Probabilidad Y Teoría Estadística

2020

Sobre los Estimadores de Bayes, el Análisis de Grupos y las Mixturas Gaussianas

Un Análisis Teórico General del Paquete ***densityMclust*** del programa estadístico **R**

Isadore Nabi

ÍNDICE

I. PRÓLOGO

La presente investigación tiene como objetivo analizar el marco teórico que es prerrequisito para estudiar y comprender plenamente las técnicas estadísticas utilizadas en (Villegas Barahona, 2018). La investigación doctoral referida tuvo como objetivo general "Estudiar la relación entre indicadores de rendimiento académico y un conjunto de dimensiones latentes y variables observadas directamente de los estudiantes para conformar un modelo estadístico que apoye la gestión académica, administrativa y la mejora continua del rendimiento académico, para lo cual se utilizará la Descomposición de la matriz CUR para la identificación y selección de las variables relevantes" (Villegas Barahona, 2018, pág. 42) en cuyo proceso aparece un aporte que radica en una mejora de la técnica CUR (en el contexto de la pedagogía) mediante lo que se denominó *Dinamic CUR* o lo aquí se considera justo denominar *CUR de Villegas*. La relevancia de esta investigación radica fundamentalmente en que permite mejoras significativas (y no sólo estadísticamente) en el proceso enseñanza-aprendizaje como resultado de permitir al investigador conocer tres aspectos relevantes del proceso educativo:

1) La relevancia estadística que tienen las variables que resultan significativas en su correlación con el rendimiento académico encontradas en la revisión sistemática de literatura, análisis bibliométrico y metaanálisis.
2) Realizar análisis mediante la técnica multivariante descomposición de la matriz CUR, para determinar las variables altamente relacionadas con el rendimiento académico y así establecer la plataforma de datos para la conformación de un modelo estadístico pedagógico con impacto en el mejoramiento continuo del rendimiento de los estudiantes universitarios que incorpore esas variables estadísticamente significativas.

3) Diseñar un modelo estadístico de clasificación de estudiantes que permite predecir el percentil de rendimiento académico esperado de un estudiante en el futuro próximo.

Así, debido a la evidente relevancia de la investigación referida en términos del impacto que tendría en el proceso de enseñanza-aprendizaje público si se diese seguimiento a los procesos de tal índole suscitados en la esfera pública del sistema educativo como parte del control de las políticas públicas en materia de educación, se ha decidido realizar esta investigación con la finalidad de proporcionar el marco teórico adecuado para su comprensión.

II. CONCEPTOS PRELIMINARES

II.I. Función de Probabilidad Posterior

Como localiza en el reporte (Nabi, Algunas Reflexiones Sobre la Distribución Binomial Negativa II (Un Análisis Teórico y Aplicado), 2020, págs. 31-33), el concepto de probabilidad condicional obedece una lógica similar, aunque menos general, que el concepto del Aufheben hegeliano. Sin embargo, el concepto de probabilidad condicional carecería de la suficiente profundidad filosófica para que su utilidad práctica fuese evidente hasta que el reverendo Thomas Bayes dijo "hágase la claridad analítica" (quizás la pidió, le fue dada y/o la encontró) y esta llegó, aunque como en promedio sucede con los genios revolucionarios, necesitó del nacimiento ni más ni menos que de un equivalente a escala aún más general (Pierre-Simon Laplace) para empezar a ser apreciado en una medida que hiciese justicia a su estatura intelectual, independientemente de sus alcances en contextos más generales y de su interpretación en tales contextos. El Teorema de Bayes es, por tanto, una forma de calcular e interpretar las probabilidades condicionales.

Así, la interpretación de las probabilidades condicionales se hará directamente en su forma más fundamental, desde el teorema de Bayes localizado en (DeGroot & Schervish, 2012, pág. 77):

$$\Pr(B_i|A) = \frac{\Pr(B_i)\Pr(A|B_i)}{\sum_{j=1}^{k}\Pr(B_j)\Pr\left(A\middle|B_j\right)} \qquad (17)$$

Es necesario comenzar por permitir que sea el mismo Bayes quien nos diga cómo se interpreta su teorema:

"DEFINITION (…) 5. The *probability of any event* is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon it's happening." (Bayes, 1763, pág. 376).

La interpretación que del Teorema de Bayes se hará en esta investigación estará orientada a la experimentación científica, por lo que se planteará desde el contexto de prueba de hipótesis[1]. Sin embargo, es fundamental conocer la interpretación objetiva, que es la interpretación óptima por cuanto lo objetivo contiene en última instancia a lo subjetivo (i.e., lo subjetivo es criatura siempre de lo objetivo[2]) del teorema en cuestión. Esta interpretación se establece en los siguientes términos: "Many applications of probability invoke a notion of probability that is objective in a logical sense: there is a fact of the matter as to what the probabilitie sare; if two agents disagree about a probability, at least one of them must be wrong. (Logical objectivity contrasts with the ontological sense of objectivity: probabilities are ontologically objective if they exist as entities or are reducible to existing entities, and are ontologically independent of mental or epistemological considerations.) For example, the probability that a patient's breast cancer will recur after treatment apparently depends on features of the cancer, of the treatment, and of the patient. It is not simply a matter of personal opinion: if two prognostic probabilities differ, at least one of them must be wrong. A philosophical interpretation of probability should, if possible, yield a notion of probability that is suitably objective in this logical sense—otherwise, it is revising rather than faithfully interpreting probabilistic statements as they occur in these applications." (Williamson, 2010, pág. 11).

Por otro lado, la interpretación del Teorema de Bayes en el contexto de las pruebas de hipótesis puede hacerse siguiendo la lógica de (Russell, 2014). Ahí, el teorema mencionado toma la forma:

---

[1] En los anexos de esta investigación se tratarán las probabilidades condicionales y la probabilidad total.

[2] Así, la definición de probabilidades aquí esbozada es compatible con los métodos utilizados por los subjetivistas (en cualquiera de sus niveles de radicalización), puesto que tales herramientas obedecen a un conjunto de axiomas (elaborados por Kolmogórov, como antes explicó), más no con la visión filosófica que orquesta sus espíritus científicos.

$$p(h|d) = \frac{p(d|h)\, p(h)}{p(d)} \qquad (18)$$

Como se menciona en la fuente citada, $p(h|d)$ puede interpretarse como qué tan verosímil es que nuestra hipótesis sea verdadera dada la evidencia científica disponible. Con ello, ahora resulta más intuitivo dar una explicación sobre el Teorema de Bayes, re-expresando las identidades (17) y (18) en una expresión diferente y de diferente ordenamiento de sus componentes:

$$\Pr(B_i|A) = \frac{\Pr(A|B_i)}{\Pr\,(A)} \Pr(B_i) \qquad (19)$$

En la expresión anterior, $\Pr(B_i|A)$ denota la probabilidad posterior, $\Pr(B_i)$ es la probabilidad a priori (conocida en teoría Bayesiana usualmente como *prior*, la cual representa el estado de conocimientos del investigador previo al encuentro de nueva evidencia relevante -la relevancia establecida por criterios que pueden variar según cada caso-) y $\frac{\Pr(A|B_i)}{\Pr\,(A)}$ es la razón (que en el numerador contiene al cociente de verosimilitud y en el denominador la probabilidad marginal asociada a una nueva y relevante evidencia encontrada -la probabilidad de cada "pieza de evidencia"-) e índice que mide la verosimilitud o confiabilidad asociada a esa nueva evidencia encontrada. Esto se entiende desde la lógica dialéctica-materialista del Teorema de Bayes y de las probabilidades condicionales en su verdad natural.

En el reporte citado, específicamente en los anexos del mismo, se puede encontrar un debate filosófico sobre la interpretación del teorema de Bayes, sin embargo, la misma naturaleza de la presente investigación hace que no sea necesario tal nivel de profundidad para ese aspecto, aunque el lector debe contar con que será necesario más adelante, por el momento diremos lo que dijo Albert Einstein a Niels Bohr en el debate sobre la interpretación del colapso de onda en la Mecánica Cuántica: la luna está ahí aunque nadie la esté mirando, *i,e.*, las probabilidades sólo son un recurso epistemológico y la verdad es siempre objetiva e independiente de

la voluntad de los hombres y mujeres, obedece a leyes fundamentales de carácter rigurosamente determinista, dinámico y complejo, en una palabra, dialécticos.

II.II. Función de Pérdida

Una función de pérdida es una regla de decisión/asignación (lo que en Álgebra Abstracta se conoce como "operación binaria") que toma un elemento de un conjunto de entrada (que le proporciona los "insumos") y, tras aplicar sobre él una operación matemática (cuyo carácter es también binario, como por ejemplo la adición -puede ser cualquier otra que el lector conozca o pueda crear como reglas de espacios vectoriales, que en los Espacios Euclidianos son los ampliamente conocidos Axiomas de Campo-) arroja un elemento de un conjunto de salida (que representa el "producto" o "resultado final" -que en inglés es el "output"- de aplicarle al "insumo" la operación). En esta función, el conjunto de entrada está dado por el valor de una o más variables aleatorias y el del conjunto de salida con su "costo asociado". Como se señala en (Wikipedia, 2020), los problemas de optimización buscan minimizar la función de pérdida. La función objetivo del problema de optimización (nótese que esta es una forma de ver la función de pérdida -como una función objetivo de un problema de optimización-) puede ser una función de pérdida o su negativo (en ramas de las ciencias en concreto, denominadas como función de recompensa, una función de ganancia, una función de utilidad, una función de aptitud, etc.), en cuyo caso debe maximizarse (la lógica del problema de optimización subordinará eso), todo refleja relaciones de poder.

Nótese que la anterior lógica, aunque útil en múltiples contextos, es profundamente limitada si se quiere expandir su comprensión analítica en distintas y diferentes aplicaciones concretas, es decir, su aplicabilidad en distintas existencias localizadas del plano de la realidad. En su lugar, conviene mejor plantear que la función de pérdida es la estructura matemática que relaciona un evento con alguna implicación (de cualquier naturaleza) a la ocurrencia de tal evento y que puede verse como un costo asociado, una implicación de recursos, un

complemento, una finalidad subyacente, etcétera. Así, puede entenderse como una función de implicación, por lo que de ello se desprende que su estructura matemática es orquestada por una naturaleza biyectiva. En el fondo, este objeto matemático captura la idea intuitiva de medir los errores en las mediciones realizadas por el investigador.

II.III. Función de Pérdida Bayesiana

Como se señala en la última fuente referida, además de las características definidas para la función de pérdida en general, esta versión concreta de la función de pérdida es la función de pérdida esperada considera dentro de su estructura matemática la incorporación de la distribución de probabilidad posterior (que para este caso se representa mediante π) del parámetro θ:

$$\rho(\pi^*, a) = \int_{\Theta} L(\theta, a) d\pi^*(\theta)$$

Luego, se debe escoger la acción a^* que minimiza o maximiza para cada caso concreto la implicación asociada (sea pérdida o ganancia, como ya se definió previamente). La fuente de señalada plantea que "(…) aunque esto resultará en elegir la misma acción que se elegiría usando el riesgo frecuentista, el énfasis del enfoque bayesiano es que uno solo está interesado en elegir la acción óptima bajo los datos observados reales, mientras que elegir la regla de decisión óptima frecuentista real, que es una función de todas las posibles observaciones, es un problema mucho más difícil", lo cual es cierto, pero no por ello revela la verdad, ya que no por ser más difícil no es una descripción filosófica y teóricamente (no estadísticamente, sino desde el marco científico del que se encuentre el investigador, que es en última instancia el que debe regir cualquier interpretación) más confiable y, por tanto, más deseable, pero eso se discutirá en otra investigación. Finalmente, es de importancia no trivial destacar que en términos más abstractos la función de pérdida (o ganancia) expresa matemáticamente una regla de decisión, como se señala en (Wikipedia, 2020), "Para evaluar la utilidad de

una regla de decisión, es necesario tener una función de pérdida que detalle el resultado de cada acción en diferentes estados(…)".

Aquí se ha considerado que no es necesario para la finalidad de la investigación que se exponga la estructura matemática de una regla de decisión, puesto que está lo suficientemente poco conectado como con la Estadística Matemática como para ser considerado analíticamente, pero sí se consideró importante resaltar la intuición esencial. En ese sentido, una regla de decisión admisible es una regla de decisión que cumple con el criterio de eficiencia en el sentido de Pareto, es decir, que obedece a la lógica de ser la selección frente a la cual no existe una alternativa superior, y sobre ello es conveniente mencionar una cuestión cuando menos curiosa.

Este criterio de eficiencia fue planteado por Pareto en el contexto de sus estudios de Economía Política (curiosamente también, él fue el último economista neo-marginalista[3] que le llamó Economía Política a lo que se conoce en la ortodoxia desde Alfred Marshall como "Economía" a secas), sin embargo, de ello se puede evidenciar cómo en las Matemáticas, que por antonomasia despoja de su esencia concreta a los objetos de la realidad, aunque a veces resulte de ello un problema (como lo que ocurre con los conjuntos autocontenidos en la paradoja de Russell, por ejemplo[4]), a veces sirve para incorporar a su marco teórico elementos que originalmente eran planteamientos de rigurosidad científica sumamente cuestionable. Cuando Pareto plantea que la distribución inicial óptima del ingreso

[3] Me refiero aquí a la nimiedad intelectual que contiene a la economía vulgar en su expresión más soez, a la conocida popularmente (de forma equivocada a falta de fundamento histórico) como "teoría neoclásica" y que por supuesto que también a sus escuelas asociadas, como propiamente los marginalistas y también a la Escuela Austríaca, que ni siquiera en su propio círculo "intelectual" (el neo-marginalista) es tomada académicamente en serio (sin entrar en detalles esenciales -propiamente de Economía Política-, hay detalles técnicos que permiten conocerlos, por ejemplo su rechazo a la utilización de instrumental matemático en el análisis económico).

[4] Se plantea esto porque si el análisis de Russell versara sobre cuerpos con alma, *i.e.*, una colección de muñecas Matrioshka, una zapatera o la históricamente característica "bolsa de bolsas" en la generalidad de hogares latinoamericanos no habría tal paradoja porque las nociones concretas implicadas no lo permitirían (por ejemplo, el aspecto espacio-temporal), ya que proporcionarían suficiente información para poderlas distinguir como diferentes objetos matemáticos.

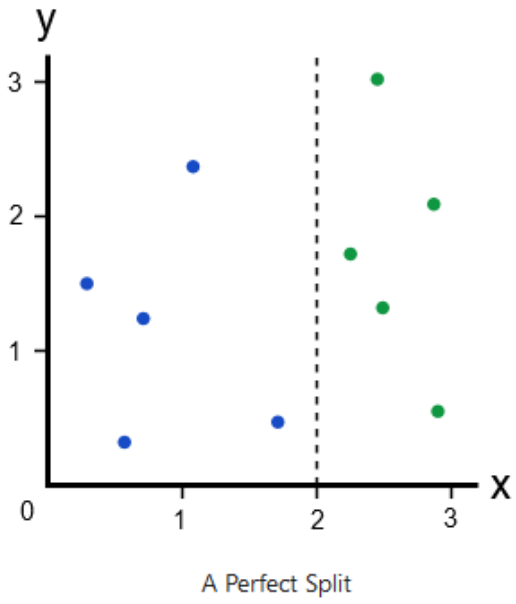
es aquella en que no se pueda mejorar el ingreso de un agente económico sin desmejorar el de otro se está haciendo, consciente o inconscientemente, una terrible omisión histórica (como es característico en la Escuela Neo-Marginalista) y es el proceso que condujo a la determinación de las condiciones iniciales bajo las cuales el análisis dinámico del fenómeno se realizará, puesto que es ampliamente conocido en *teoría del caos* que las condiciones iniciales son las que determinan de forma dinámica y no lineal (*i.e.*, compleja) el resultado de los estudios concretos del comportamiento general de los sistemas a largo plazo por lo que la rigurosidad científica con que se validen las condiciones iniciales son la esencia de la validación científica de cualquier modelo teórico dinámico y complejo, *i.e.*, de cualquier modelo que se corresponda con la complejidad de la realidad misma.

Para el caso concreto de la Economía Política, ¿cualesquiera condiciones iniciales son válidas?, y puesto que la Economía Política (como el mismo Pareto la llamaba) es una ciencia que estudia la vida económica del ser humano en sociedad (en sociedad de clases, específicamente) es, como todo es ampliamente conocido, una ciencia social, por lo que al ser la sociedad misma un fenómeno histórico, tanto como cualquier otro de la realidad, ¿cuáles son los criterios históricos, validados científicamente desde la Historia, que a su vez validan esas condiciones iniciales en la distribución del ingreso entre los agentes económicos? Pareto no cumple con absolutamente ninguno de esos criterios, sin embargo, tuvo la suerte de jugar un papel relevante en las ciencias formales a pesar de ser tan deficiente filosóficamente y en las ciencias sociales como resultado de que las Matemáticas le hicieron el favor de borrar de su planteamiento teórico el núcleo del mismo, su esencia teórica. Es parecido a lo que Marx describe sobre Napoleón III en el *XVIII Brumario de Luis Bonaparte*, a veces las condiciones históricas permiten que tipos de poca monta jueguen papeles históricamente relevantes, aun cuando la intencionalidad de sus actos era diferente e incluso distinta al desenlace histórico posterior.

II.IV. Árbol de Decisión Iterativo (Regla de Decisión Iterativa)

Como se expone en (Zhou, 2020), es aquel algoritmo que iterativamente parte los datos en dos subconjuntos disjuntos (sin elementos en común) en donde o bien cada uno de estos subconjuntos posee elementos de naturaleza equivalente o similar (según sea definido por el investigador) o bien puede construirse de forma aleatoria, como por ejemplo con la metodología conocida como "Random Forest".

Figura 2



A Perfect Split

This is a **perfect** split! It breaks our dataset perfectly into two branches:

- Left branch, with 5 blues. ● ● ● ● ●
- Right branch, with 5 greens. ● ● ● ● ●

Fuente: (Zhou, 2020).

La figura anterior es lo que se conoce como partición perfecta. Como se señala en (Yanchick, 2020), cuando el árbol de decisión realiza particiones perfectas se convierte en un *algoritmo codicioso* (en inglés "Greedy Algorithm"), es decir, que toma la decisión óptima a cada paso.

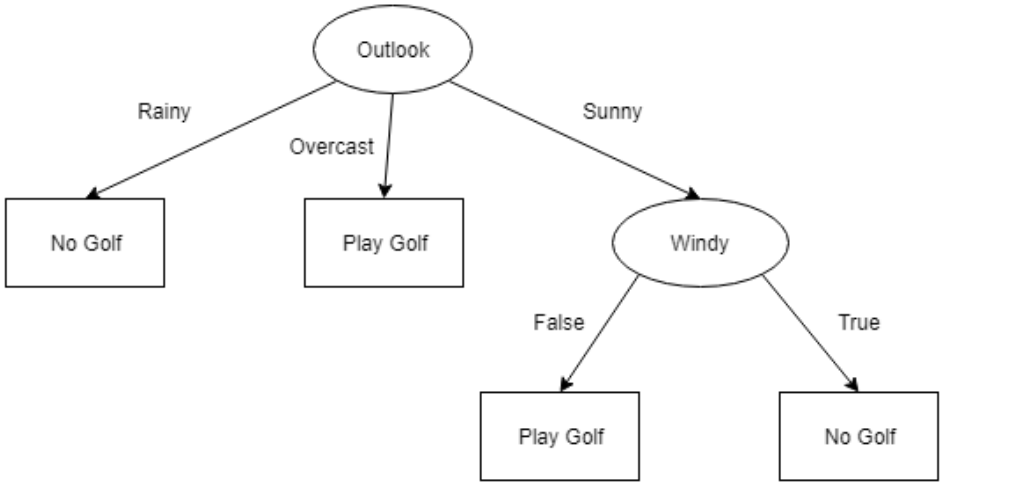
Para que clarificar las definiciones anteriores, se expondrá un ejemplo sobre el pronóstico del clima y cómo sobre tal pronóstico se tomará la decisión si jugar golf o no, según lo planteado en (Ganegedara, 2020).

Figura 2

Outlook	**Temperature**	**Humidity**	**Windy**	**Play Golf**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes

Fuente: (Ganegedara, 2020).

Figura 3



An example decision tree. Round nodes denote decision nodes, where square nodes denote leaf nodes

Fuente: (Ganegedara, 2020).

II.V. Error Cuadrático Medio

Como se señala en (Wikipedia, 2020), en el contexto bayesiano previamente descrito, la función de pérdida (o ganancia) es equivalente al error cuadrático medio, es decir, a la esperanza matemática de la distancia elevada al cuadrado entre la estimación del parámetro poblacional y su valor observable:

$$MSE(\hat{\theta}) = E_{\theta}\left[(\hat{\theta} - \theta)^2\right]$$

La expresión anterior puede presentarse de forma más familiar, tal como aparece en (Ranganathan, Nakain, Schönback, & Gribskov, 2019, pág. 408):

$$\frac{\sum_{i=1}^{d} |y_i - y_i'|}{d}$$

En donde d en el contexto de la Bioinformática y la Biología Computacional es el número total de pruebas, que en el contexto de la Teoría Estadística puede interpretarse como el número total de observaciones.

II.VI. Media Posterior

Usando el error cuadrático medio como riesgo de la estimación (*i.e.*, como función de pérdida), la estimación de Bayes del parámetro desconocido es simplemente la media de la distribución posterior, como se presenta en (Jaynes, 2003, pág. 172).

II.VII. Valores Ajustados de una Variable Aleatoria Dependiente

Como se señala en (Frost, 2020), los valores ajustados de una variable aleatoria son las predicciones de un modelo estadístico del valor de respuesta medio cuando se ingresan los valores de los predictores o variables aleatorias independientes en el modelo estadístico en cuestión.

II.VIII. Error Cuadrático Medio Mínimo

El error cuadrático medio se minimiza en este escenario con un método conocido como error cuadrático medio mínimo, que es una medida ampliamente utilizada de la calidad de un estimador de los valores ajustados de la variable aleatoria dependiente. En el contexto bayesiano, como se señala en (Wikipedia, 2020), el término *MMSE* (por su nombre en inglés) hace alusión a la función de pérdida cuadrática y en tal caso, el estimador es obtenido mediante el promedio de la distribución posterior del parámetro que se desea estimar en el marco de la metodología estadística MMSE. Así, la forma matemática del MMSE es la siguiente:

$$MSE = E[\left(\hat{\theta}(x) - \theta\right)^2]$$

En donde la esperanza matemática se estima con base en la distribución de probabilidad conjunta de θ y x.

Finalmente, según (Jaynes, 2003, pág. 172), adaptando la notación contenida en la fuente referida a la utilizada en la presente investigación[5], la media posterior se estima de la siguiente manera:

$$\hat{\theta}(x) = E[(\theta|x)] = \int \theta \, p(\theta|x) \, d\theta$$

La media posterior o media de Bayes no es otra cosa que la "decisión" a tomar, viéndola desde el punto de vista de la Teoría de la Decisión.

II. IX. Estimador de Bayes y Riesgo de Bayes

Como se señala en (Wikipedia, 2020), es el estimador que minimiza la esperanza matemática de la función de pérdida posterior (función de pérdida bayesiana). Su definición formal puede comprenderse con facilidad siempre que se recuerde con antelación, como se señala en (Dey & Rao, 2005, pág. 318), que los estimadores bayesianos son simplemente medias de distribuciones de probabilidad posteriores en muestras grandes bajo funciones de pérdida cuadráticas, en palabras justas de los autores "(…) namely, posterior means under squared los function (…)", que fue a su vez lo que se planteó en el acápite sobre el MMSE.

Suponga el lector que un parámetro desconocido θ es conocido por tener una distribución a priori o prior π. Entonces, sea $\hat{\theta} = \hat{\theta}(x)$ un estimador de θ (basado en algunas medidas de x) y sea $L(\theta, \hat{\theta})$ una función de pérdida, por ejemplo, el error cuadrático medio. Bajo tales condiciones, el riesgo de Bayes del estimador del parámetro poblacional θ, es decir, $\hat{\theta}$, está definido matemáticamente como

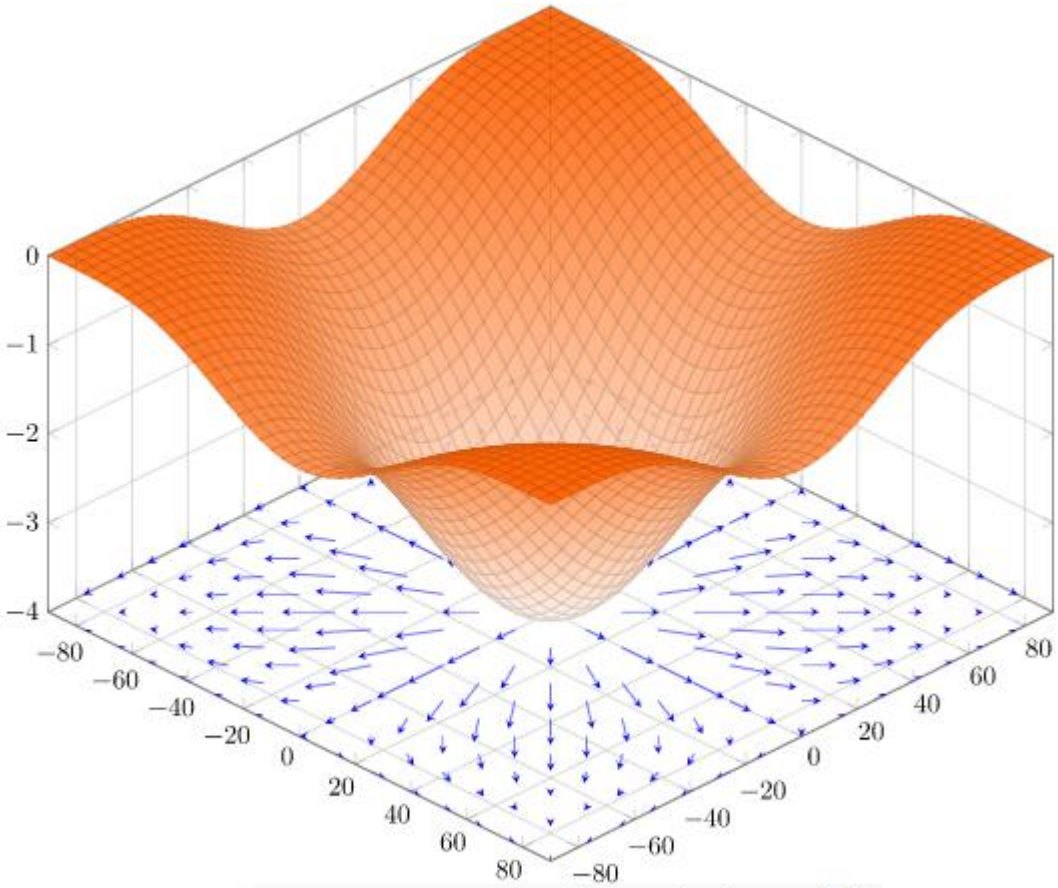
[5] Jaynes, en la localización referida, utiliza la notación $\alpha^* = \langle\alpha\rangle = \int d\alpha \, p(\alpha|DI)$.

$E_\pi[L(\theta, \hat{\theta})]$. Un estimador se dice que es un estimador de Bayes si este minimiza el riesgo bayesiano entre todos los estimadores, lo que en términos simples significa que es la regla de decisión que permite construir la función distancia o métrica del Campo Probabilístico de Kolmogórov que permite minimizar los errores de predicción de los valores de los parámetros estadísticos estimados (que por antonomasia son siempre poblacionales) dado un conjunto de datos observados (los datos muestrales, como se verifica en (Wikipedia, 2020)). De manera equivalente, el estimador que minimiza la media de la función de distribución de pérdida esperada $E_\pi[(L(\theta, \hat{\theta})|x)]$ para cada x también minimiza el riesgo de Bayes y, por tanto, es un estimador de Bayes.

II.X. Gradiente

Como se señala en (Wikipedia, 2020), un gradiente es la colección de las derivadas parciales de una función multivariable, por lo que a su vez es una generalización del concepto de derivada parcial y puede interpretarse intuitivamente como la dirección y la tasa de crecimiento más rápida, es decir, es la región del espacio (sea un campo probabilístico o de otro tipo) en donde las derivadas de la función alcanzan su máximo. Así, la dirección del gradiente es la dirección en que la función crece más rápido partiendo de algún punto de referencia p, mientras que la magnitud del gradiente es la tasa de incremento en esa dirección.

Figura 4



The gradient of the function $f(x,y) = -(\cos^2 x + \cos^2 y)^2$ depicted as a projected vector field on the bottom plane.

Fuente: (Wikipedia, 2020).

II.XI. Propiedades de los Estimadores de Bayes

Las propiedades de los estimadores de Bayes son fundamentalmente:

1) Que su estimación es resultado de una regla de decisión admisible, es decir, no existe otra regla de decisión que sea siempre "mejor" que ella (o al menos algunas veces mejor y nunca peor), en el sentido preciso en que "mejor se define a continuación: una regla de decisión es admisible si y solo si ninguna otra regla la domina. Aquí debe entenderse la dominancia como se entiende en la Teoría de Juegos, es decir, "La estrategia s_i' está **estrictamente dominada** por la estrategia s_i'' si para cada combinación posible de estrategias de los restantes jugadores la ganancia de i por utilizar s_i' es

estrictamente menor que la ganancia de i por utilizar s_i'' (…)" (Gibbons, 1992, pág. 5). Nótese que "estrictamente dominada" hace referencia a que el operador utilizado en la relación es $<$, mientras que en simplemente "dominada" se expresa usualmente mediante el operador $\leq$.

2) Son eficientes asintóticamente. Esto implica que en muestras grandes la densidad posterior se vuelve normal. En otras palabras, para un tamaño de muestra lo suficientemente grande el efecto de la probabilidad anterior sobre la posterior es insignificante. Esta propiedad, cuando el estimador de Bayes tiene como función de pérdida esperada al error cuadrático medio, se puede expresar como que el estimador es asintóticamente insesgado. Recordando que un estimador insesgado en el infinito (en muestras grandes) es una sucesión de estimadores que converge en probabilidad a la cantidad que está siendo estimada a medida el índice (a menudo el tamaño de la muestra) crece sin restricción, es decir, a medida incrementa el tamaño de la muestra también incrementa la probabilidad de que el estimador se acerque al parámetro poblacional. Matemáticamente hablando, una sucesión de estimadores $\{t_n; n \geq 0\}$ es un estimador consistente para el parámetro θ si para todo número positivo arbitrariamente pequeño (no importa cuán pequeño) $\varepsilon > 0$ se tiene que $\lim_{n\to\infty} \Pr(|t_n - \theta| < \varepsilon) = 1$ y tal límite al infinito converge en distribución a la Normal de la forma $\sqrt{n}(\delta_n - \theta) \to N(0, \frac{1}{I(\theta)})$. En la expresión anterior, $I(\theta)$ representa la *Información de Fisher*, la cual es una forma de medir la cantidad de información que una variable aleatoria x contiene sobre un parámetro desconocido θ de una distribución que modela x. Formalmente hablando, es la varianza del puntaje estadístico (que a su vez se modelan matemáticamente mediante gradientes) o bien, también puede ser la esperanza matemática de la información observada. Cuando se modela matemáticamente la Información de Fisher como la esperanza matemática,

entonces su proceso de cálculo se realiza mediante el negativo de la segunda derivada de la función de log-verosimilitud, es decir, el negativo del determinante de la matriz Hessiana obtenida del logaritmo de la función de verosimilitud. En este sentido, la matriz de información de Fisher es usada para calcular las matrices de covarianza asociadas con las estimaciones de máxima verosimilitud.

II.XII. Definición Formal de Límite. La Definición Épsilon-Delta

Es necesario comenzar por familiarizarse un poco con la definición formal de un límite, que no es más que la formalización de la noción intuitiva de aproximación hacia un punto concreto de una sucesión o una función, a medida que los parámetros[6] de esa sucesión o función se acercan a un determinado valor. Se dice que $f(x)$ se acerca a un límite cuando x se acerca a un valor a. Ahora bien, al hablar de un acercamiento o aproximación, se está hablando implícitamente de la distancia entre dos valores. ¿Cómo es posible representar la noción de cercanía?, restando un valor de otro para conocer dicha distancia. En el caso de la definición formal de un límite, al decir que $f(x)$ se acerca hacia un valor límite L se tiene que sustraer L de $f(x)$, es decir, $f(x) - L$ y al decir que x se acerca hacia un valor a se tiene que hacer lo mismo, es decir, $x - a$. Sin embargo, las distancias no puede ser valores negativos, por lo que se recurre a la utilización del valor absoluto para garantizar valores positivos en la resta, por tanto, se tendría $|f(x) - L|$ y $|x - a|$. Esto a su vez implicaría que $|f(x) - L|$ es un valor muy pequeño cuando $|x - a|$ es un valor muy pequeño.

Ahora bien, para formalizar aún más la noción intuitiva de límite, se tendrá que elegir un determinado valor para $|f(x) - L|$ y otro determinado valor para $|x - a|$, debido a que es necesario acotar o "encerrar" cada uno de estos valores absolutos, ¿cómo es que se logra acotar cada uno de estos valores absolutos?, la respuesta se encuentra en la definición misma del valor absoluto. Un valor absoluto no es más

[6] Constantes que pueden ser variables.

que el valor numérico de un número real sin tener en cuenta su signo. Entonces cuando se "encierra" cada una de las restas automáticamente se está acotando o "cercando" alrededor de determinado valor. Por ejemplo, el $|3|$ significa que $-3 < |3| < 3$. Quien planteó esta definición formal de un límite fue el matemático francés Augustin Louis Cauchy, diciendo que habría un error en la aproximación de $f(x)$ hacia L, lo que denotó por la letra ε y que habría a su vez una distancia que representaría el cambio en las abscisas de x hacia a, lo que denotó con la letra δ. Lo anterior significa que Épsilon (ε) es el error de aproximación de $f(x)$ hacia L y Delta (δ) es la distancia recorrida o la variación en las abscisas al pasar de un valor x hacia un valor a.

Por supuesto, realizar el acotamiento de forma adecuada en términos matemáticos es necesario que tanto Épsilon como Delta sean positivos, es decir, mayores que cero, pues de lo contrario, el acotamiento no sería posible. Finalmente, se requiere también que el valor absoluto de x menos a sea también mayor que cero, ¿por qué?, pues al plantear la Definición Épsilon-Delta se está planteando a su vez que x tiende hacia a, pero es una tendencia de x en que esta variable tomará valores cercanos en relación con a, sean estos mayores o menores que a, pero nunca iguales que a. Al introducir el valor absoluto y garantizar que la resta sea positiva, se está introduciendo también que $0 < |x - a|$, pues cero es siempre menor que cualquier número positivo. Por supuesto, lo anteriormente expuesto debe complementarse con el hecho que para que un límite L exista, el valor del límite L cuando x tiende hacia a tanto por la izquierda como por la derecha, debe ser el mismo. ¿Cómo es posible esto?, la respuesta se encuentra en la misma definición formal de un límite y en el Teorema del Encaje visto anteriormente. Al plantear que $|f(x) - L| < \varepsilon$ se está expresando a su vez que $L - \varepsilon < f(x) < L + \varepsilon$, es decir, que el valor de la función evaluada en x se encontrará en un intervalo equivalente al límite menos Épsilon y el límite más Épsilon, lo que matemáticamente significa que el límite será idéntico en ambos extremos del intervalo; análogamente, al plantear que $|x - a| < \delta$ se está

expresando a su vez $a - \delta < x < a + \delta$, es decir, que el de x se encontrará en un intervalo equivalente a la tendencia menos Delta y la tendencia más Delta, lo que matemáticamente significa que la tendencia será idéntica en ambos extremos del intervalo. Siendo esto así, no importa si tomemos un valor por la izquierda o por la derecha de la tendencia de x, la tendencia en sí misma será igual y el valor de la función evaluada en x tendrá también el mismo límite tanto por la izquierda como por la derecha.

En otras palabras, esta definición lo que dice es que se busca un intervalo alrededor del límite L muy pequeño y que L siempre será un valor que, tendencialmente, le corresponderá a y.

ε (Épsilon): Es un número infinitamente pequeño que se le sumará y restará al valor del límite para poder delimitar la tendencia alrededor de ese intervalo en y.

δ (Delta): Es un número que dependerá del valor de ε (Épsilon) y servirá para delimitar la tendencia alrededor de ese intervalo en x.

Formalización de ε (Épsilon)

$$-\varepsilon < f(x) - L < \varepsilon$$

$$L - \varepsilon < f(x) < L + \varepsilon$$[7]

$$\lim_{x \to a} f(x) = L$$[8]

[7] Al restarle y sumarle al límite un valor Épsilon a la izquierda y derecha de la función, respectivamente, lo que se establece es un acotamiento de dicha función en un intervalo. Lo anterior implica que el acotamiento por la izquierda representa un valor menor a la función evaluada en x y la acotación por la derecha representa un valor mayor a la función evaluada en x.

[8] Esto es, ni más ni menos, que la definición formal de un límite para una función $f(x)$ cuando x tiende hacia un valor a.

Formalización de δ (Delta)[9]

$$0 < |x - a| < \delta$$

$$|x - a| < \delta$$

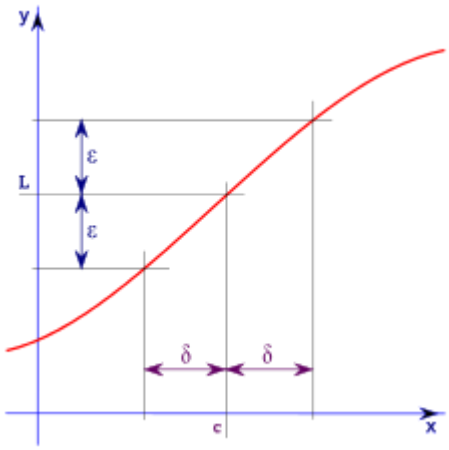
$$-\delta < x - a < \delta$$

$$-\delta + a < x < \delta + a$$

Obsérvese que al contrario de lo que en Matemáticas se acostumbra a realizar (establecer que y dependa de x), aquí Delta depende de Épsilon, porque se parte de Épsilon para acotar el límite, esto debido a que el valor del límite será siempre un valor en y. Posteriormente se procede a definir un valor Delta en x. A su vez, se acota primero el valor del límite y no x porque se está probando que el resultado del límite en la función evaluada en x es L y porque para todo $\varepsilon > 0, \exists\ \delta$. En otras palabras, se acota primero el valor en y por la conveniencia que esto representa, es decir, porque se requiere acotar el límite y este siempre será un valor en y, independientemente que sea un valor resultante de evaluar la función en x, dado que el acotamiento previo no está relacionado directamente en términos matemáticos con la evaluación de la función en x.

[9] Nótese que a es equivalente al c de la gráfica de la Definición ε- δ, pues ambas representan una constante cuya notación se escoge arbitrariamente, hacia la cual tiende la variable independiente x.

Figura 5

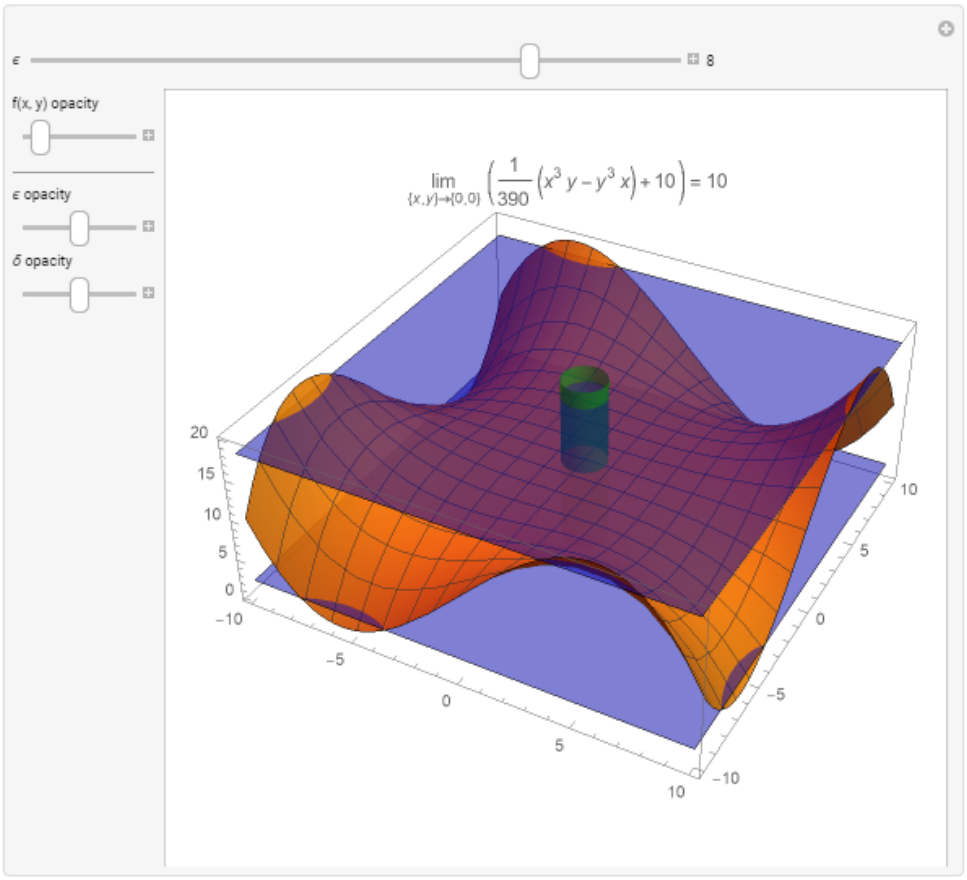


Fuente: (Wikipedia, 2020).

Es necesario mencionar que para el caso de dos o más variables la idea intuitiva que orquestó el caso univariable se conserva, independientemente de que el desarrollo matemático incremente su complejidad y tome formas diferentes. Esto puede verificarse mediante las dos figuras presentadas a continuación.

Figura 6

Multivariable Epsilon-Delta Limit Definitions



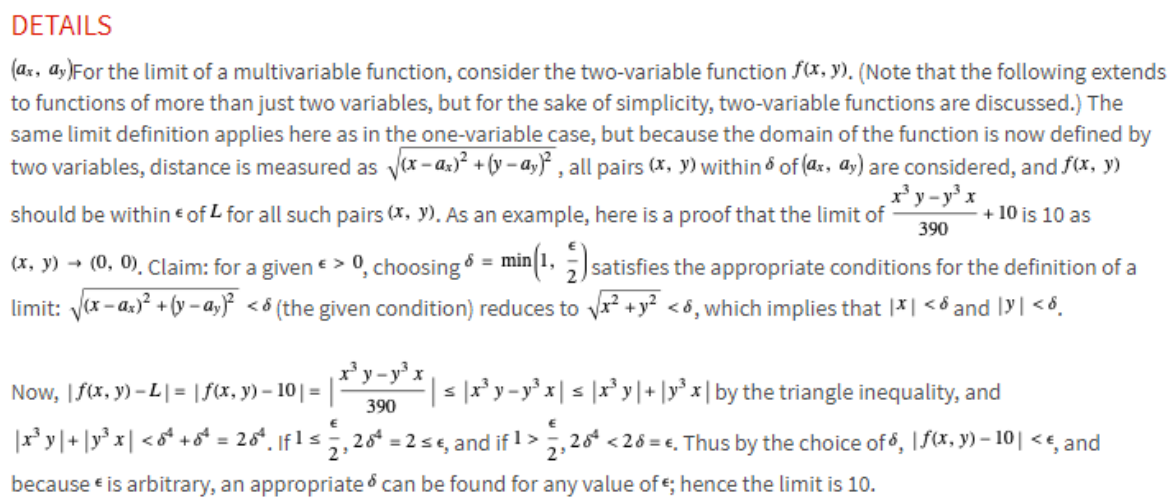
The definition of a limit: The expression $\lim_{x \to a} f(x) = L$ is an abbreviation for: the value of the single-variable function $f(x)$ approaches L as x approaches the value a. More formally, this means that $f(x)$ can be made arbitrarily close to L by making x sufficiently close to a, or in precise mathematical terms, for each real $\epsilon > 0$, there exists a $\delta > 0$ such that $0 < |x - a| < \delta \rightarrow |f(x) - L| < \epsilon$. In other words, the inequalities state that for all x except a within δ of a, $f(x)$ is within ϵ of L.

This definition extends to multivariable functions as distances are measured with the Euclidean metric.

In the figure, the horizontal planes $10 \pm \epsilon$ represent the bounds on $f(x, y)$ and the cylinder is $|x - a| = \delta$. No matter what ϵ is given, a δ is found (represented by the changing radius of the cylinder) so that all points on the surface $z = f(x, y)$ inside the cylinder are between the two planes.

Fuente: (Liang, 2011).

Figura 7



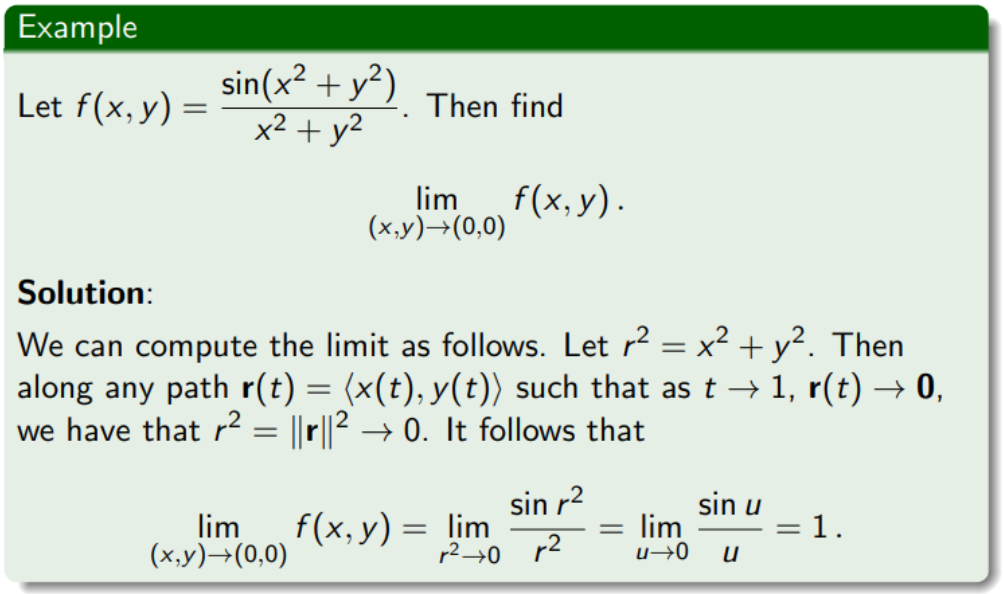
DETAILS

(a_x, a_y)For the limit of a multivariable function, consider the two-variable function $f(x, y)$. (Note that the following extends to functions of more than just two variables, but for the sake of simplicity, two-variable functions are discussed.) The same limit definition applies here as in the one-variable case, but because the domain of the function is now defined by two variables, distance is measured as $\sqrt{(x-a_x)^2+(y-a_y)^2}$, all pairs (x, y) within δ of (a_x, a_y) are considered, and $f(x, y)$ should be within ϵ of L for all such pairs (x, y). As an example, here is a proof that the limit of $\frac{x^3 y - y^3 x}{390} + 10$ is 10 as $(x, y) \to (0, 0)$. Claim: for a given $\epsilon > 0$, choosing $\delta = \min\left(1, \frac{\epsilon}{2}\right)$ satisfies the appropriate conditions for the definition of a limit: $\sqrt{(x-a_x)^2+(y-a_y)^2} < \delta$ (the given condition) reduces to $\sqrt{x^2+y^2} < \delta$, which implies that $|x| < \delta$ and $|y| < \delta$.

Now, $|f(x, y) - L| = |f(x, y) - 10| = \left|\frac{x^3 y - y^3 x}{390}\right| \le |x^3 y - y^3 x| \le |x^3 y| + |y^3 x|$ by the triangle inequality, and $|x^3 y| + |y^3 x| < \delta^4 + \delta^4 = 2\delta^4$. If $1 \le \frac{\epsilon}{2}$, $2\delta^4 = 2 \le \epsilon$, and if $1 > \frac{\epsilon}{2}$, $2\delta^4 < 2\delta = \epsilon$. Thus by the choice of δ, $|f(x, y) - 10| < \epsilon$, and because ϵ is arbitrary, an appropriate δ can be found for any value of ϵ; hence the limit is 10.

Fuente: (Liang, 2011).

Figura 8



Example

Let $f(x, y) = \dfrac{\sin(x^2 + y^2)}{x^2 + y^2}$. Then find

$$\lim_{(x,y)\to(0,0)} f(x, y)\,.$$

Solution:

We can compute the limit as follows. Let $r^2 = x^2 + y^2$. Then along any path $\mathbf{r}(t) = \langle x(t), y(t)\rangle$ such that as $t \to 1$, $\mathbf{r}(t) \to \mathbf{0}$, we have that $r^2 = \|\mathbf{r}\|^2 \to 0$. It follows that

$$\lim_{(x,y)\to(0,0)} f(x, y) = \lim_{r^2\to 0} \frac{\sin r^2}{r^2} = \lim_{u\to 0} \frac{\sin u}{u} = 1\,.$$

Fuente: (Havens, 2019, pág. 7).

Figura 9

The previous example has a geometric solution as well: the graph for $z = f(x, y) = \sin(r^2)/r^2$ is a surface of revolution.

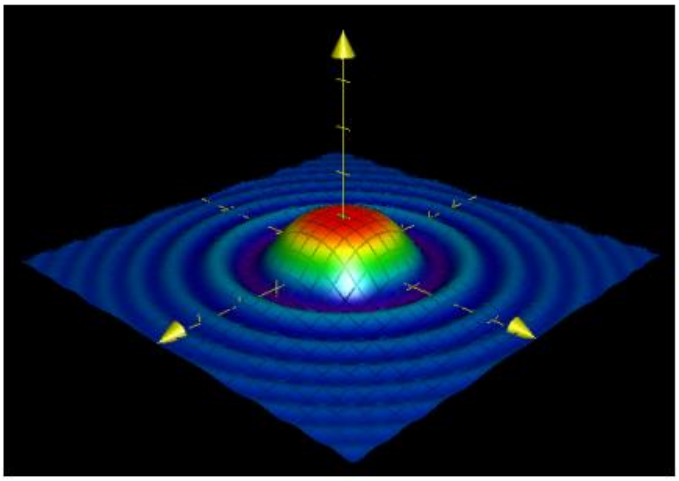


Figure: The graph of $z = \sin(r^2)/r^2$.

Fuente: (Havens, 2019, pág. 8).

II.XIII. Continuidad Absoluta

Como se señala en (Wikipedia, 2020), "In calculus, absolute continuity is a smoothness property of functions that is stronger than continuity and uniform continuity. The notion of absolute continuity allows one to obtain generalizations of the relationship between the two central operations of calculus—differentiation and integration. This relationship is commonly characterized (by the fundamental theorem of calculus) in the framework of Riemann integration, but with absolute continuity it may be formulated in terms of Lebesgue integration. For real-valued functions on the real line, two interrelated notions appear: absolute continuity of functions and absolute continuity of measures. These two notions are generalized in different directions. The usual derivative of a function is related to the Radon–Nikodym derivative, or density, of a measure."

Sea I un intervalo de la recta real $\mathbb{R}$. Una función $f: I \to \mathbb{R}$ es absolutamente continua en I si para todo $\epsilon > 0$ existe un $\delta > 0$ tal que cada vez que una sucesión infinita de subintervalos disjuntos por parres (x_k, y_k) de I satisfacen que:

$$\sum_k (y_k - x_k) < \delta$$

Por lo que entonces,

$$\sum_k |f(y_k) - f(x_k)| < \varepsilon$$

La colección de todas las funciones absolutamente continuas en I se denota $AC(I)$.

II.XIV. Espacio Paramétrico

Como señala (Lehmann, 1959, pág. 1), "The raw material of a statistical investigation is a set of observations; these are the values taken on by random variables X whose distribution P_θ is at least partly unknown. Of the parameter θ, which labels the distribution, it is assumed known only that it lies in a certain set Ω, the *parameter space*." Históricamente, el concepto de espacio paramétrico proviene de la Física, es el análogo estadístico de los espacios de fase estudiados en el marco de los sistemas dinámicos, los cuales fueron desarrollados por Ludwig Boltzmann (padre de la Mecánica Estadística y de la explicación estadística a la segunda ley de la termodinámica), Henri Poincaré (padre de la Topología y el último gran polímata de las ciencias formales) y Josiah Willard Gibbs (padre de la Mecánica Estadística -él acuñó tal termino-, el Cálculo Vectorial y la Física Química). Sin embargo, la definición de Lehmann permite plantear los espacios paramétricos en términos de Teoría de Conjuntos, lo cual es congruente con la forma en que Kolmogórov axiomatizó la Teoría de las Probabilidades y permite ver a los elementos involucrados como datos y no como objetos o puntos materiales (masa puntual o partícula) y con la finalidad de utilizar una terminología que le resulte más familiar al lector es que se presenta la definición de Lehmann.

II.XV. Criterio Bayesiano de Información (BIC)

Como se señala en (Wikipedia, 2020), "In statistics, the Bayesian information criterion (BIC) or Schwarz information criterion (also SIC, SBC, SBIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

The BIC was developed by Gideon E. Schwarz and published in a 1978 paper (…) where he gave a Bayesian argument for adopting it."

Vale la pena hacer una pausa para hablar de la información. ¿Qué es la información?, según (Real Academia Española, 2020) es enterar o dar noticia de algo, así como también puede ser fundamentar algo, e incluso señala que en su sentido filosófico es dar forma sustancial a algo. La información es, más allá de la Estadística, lo que se obtiene de los datos recolectados (sobre un determinado fenómeno o conjunto de fenómenos naturales o sociales) tras aplicarlo sobre estos los métodos y las técnicas estadísticas (de la simpleza o complejidad que sea) y analizar tales resultados a la luz del marco teórico científico respectivo. Sin embargo, en el contexto puramente de la Estadística, la definición de información es mucho más limitada y es equivalente a los datos disponibles.

Ahora bien, ¿para qué se necesita un criterio de información de cualquier índole? A esta pregunta responde (Schwarz, 1978, pág. 461). En la investigación referida nace el ahora conocido como *Criterio Bayesiano de Información*.

Al respecto, el autor señala que "Statisticians are often faced with the problem of choosing the appropriate dimensionality of a model that will fit a given set of

observations. Typical examples of this problem are the choice of degree for a polynomial regression and the choice of order for a multi-step Markov chain[10]. In such cases the maximum likelihood principle invariably leads to choosing the highest possible dimension. Therefore, it cannot be the right formalization of the intuitive notion of choosing the "right" dimension."

¿En qué se diferencia entonces del criterio de información cuya autoría responde a Hirotugu Akaike y que le precede históricamente?, pues la respuesta a la pregunta anterior, junto con la explicación concisa del criterio de Akaike, es el punto de partida de la investigación Schwarz, a lo que este plantea que "An extension of the maximum likelihood principle is suggested by Akaike (…) for the slightly more general problem of choosing among different models with different number of parameters. His suggestion amounts to maximazing the likelihood function separately for each model j, obtaining, say, $M_j(X_1, \dots, X_n)$, and then choosing the model for which $\log\left[M_j(X_1, \dots, X_n)\right] - k_j$ is largest, where k_j is the dimension of the model. We present an alternative approach to the problem (…)"

La dimensionalidad de un conjunto de datos es la cantidad de atributos correspondiente a tal conjunto de datos. En la investigación de Schwarz se plantea una alternativa a Akaike que consiste en que, en un modelo de dimensionalidad dada, los estimadores de máxima verosimilitud (EMV) pueden obtenerse como los límites en muestra grande de los estimadores de Bayes, para el caso de cierta clase especial de distribuciones a priori.

Como señala Schwarz, esta clase de distribuciones no son absolutamente continuas, puesto que asignan probabilidad positiva en algunos subespacios (del espacio paramétrico) de menor dimensión, cuando por la misma definición de continuidad absoluta la probabilidad asignada a tales subespacios debería ser nula.

[10] En referencia a Markov Jr.

Finalmente, el criterio de información Bayesiano consiste en escoger el modelo estadístico para el cual la siguiente expresión encuentra su valor máximo:

$$\log\left[M_j(X_1, \ldots, X_n)\right] - \frac{1}{2} k_j \log(n)$$

II.XVI. Minería de Datos

Como se señala en (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2005, pág. 3), "El primer pensamiento de muchos al oír por primera vez el término "minería de datos" fue la reflexión "nada nuevo bajo el sol". En efecto, la "minería de datos" no aparece por el desarrollo de tecnologías esencialmente diferentes a las anteriores, sino que se crea, en realidad, por la aparición de nuevas necesidades y, especialmente, por el reconocimiento de un nuevo potencial: el valor, hasta ahora generalmente infrautilizado, de la gran cantidad de datos almacenados informáticamente en los sistemas de información de instituciones, empresas, gobiernos y particulares. Los datos pasan de ser un "producto" (el resultado histórico de los sistemas de información) a ser una "materia prima" que hay que explotar para obtener el verdadero "producto elaborado", el conocimiento; un conocimiento que ha de ser especialmente valioso para la ayuda en la toma de decisiones sobre el ámbito en el que se han recopilado o extraído los datos. Es bien cierto que la estadística es la primera ciencia que considera los datos como su materia prima, pero las nuevas necesidades y, en particular, las nuevas características de los datos (en volumen y tipología) hacen que las disciplinas que integran lo que se conoce como "minería de datos" sean numerosas y heterogéneas."

Por otro lado, "Los juegos de datos encierran estructuras, patrones y reglas de los que es posible extraer conocimiento sobre los eventos que los han generado. La física teórica más avanzada, cada vez más, habla de un universo de eventos y no de partículas, de modo que tratar de entender o incluso predecir estos eventos se ha convertido en un reto para muchas disciplinas y la razón de ser para la minería

de datos, que ha pasado de tratar de entender los datos a tratar de comprender los eventos que hay detrás. Las cosas no son como aparentan ser, por esta razón técnicas como la visualización de datos, aunque necesarias, son del todo insuficientes para llegar hasta el conocimiento que se esconde detrás de estructuras y relaciones no triviales en los juegos de datos. Esta nueva visión convierte necesariamente los equipos de analistas de datos en equipos interdisciplinares donde se requieren habilidades matemáticas e informáticas, pero también conocedoras del negocio y de la organización empresarial. Solo así se podrá cubrir el proceso de extracción de conocimiento de principio a fin. Por medio de una metodología adecuada y de entender tanto qué tipo de problemas trata de resolver, como con qué técnicas hace frente a estos retos, trataremos de ofrecer una visión lo más completa posible de lo que es hoy en día la minería de datos." (Gironés Roig, Casas Roma, Minguillón Alfonso, & Caihuelas Quiles, 2017, pág. 25).

II.XVII. Pre-Procesamiento de Datos

Como se señala en (PowerData, 2016), "El preprocesamiento de datos es un paso preliminar durante el proceso de minería de datos. **Se trata de cualquier tipo de procesamiento que se realiza con los datos brutos para transformarlos en datos que tengan formatos que sean más fáciles de utilizar** (…) En el mundo real, los datos frecuentemente no están limpios, faltan valores clave, contienen inconsistencias y suelen mostrar ruido, conteniendo errores y valores atípicos. Sin un preprocesamiento de datos, estos errores en los datos sobrevivirían y disminuirían la calidad de la minería de datos (…) La falta de limpieza adecuada en los datos es el problema número uno en data warehousing. Algunos de las tareas de preprocesamiento de datos son las siguientes (…) Rellenar valores faltantes (…) Identificar y eliminar datos que se pueden considerar un ruido (…) Resolver redundancia (…) Corregir inconsistencias (…) Los datos están disponibles en varios formatos, tales como formas estáticas, categóricas, numéricas y dinámicas (…) Algunos ejemplos incluyen metadatos, webdata, texto, vídeo, audio e imágenes. Estas formas de datos tan variadas contribuyen a que el procesamiento

de datos continuamente se encuentre con nuevos desafíos (…) Además de manejar datos faltantes, es esencial identificar las causas de la falta de datos para evitar que esos problemas evitables con los datos no vuelvan a ocurrir. Las soluciones para datos faltantes incluyen rellenar manualmente los valores perdidos y rellenar automáticamente con la palabra "desconocido" (…) La duplicación de datos puede ser un problema importante en minería de datos, ya que a menudo hace que se pierdan negocios, se pierda el tiempo y sea difícil de tratar. Un ejemplo común de un problema de duplicación de datos típico incluye varias llamadas de ventas al mismo contacto. Las posibles soluciones implican actualizaciones de software o cambiar la forma en que tu negocio controla la gestión de relaciones con clientes. Sin un plan específico y el software adecuado, es difícil eliminar la duplicación de datos (…) Otra fuente común de duplicación de datos es cuando una empresa tiene un número excesivo de bases de datos. Como parte de su preprocesamiento de datos debe revisar periódicamente oportunidades para reducir y eliminar algunas de esas bases de datos. Si no se hace, la duplicación de datos es probable que sea un problema recurrente con el que vas a tener que lidiar una y otra vez (…) Alcanzar la calidad de datos en minería de datos (…) La mayoría de las empresas quieren hacer un mejor uso de sus extensos datos, pero no están seguros acerca de por dónde empezar. La limpieza de datos es un primer paso prudente de un largo camino hacia la mejora de la calidad de los datos. La calidad de los datos puede ser un objetivo difícil de alcanzar sin una metodología eficaz que acelere la limpieza de datos: 1. Reconocer el problema e identificar las causas fundamentales (…) 2. Creación de una estrategia y visión de calidad de datos (…) 3. Priorizar la importancia de los datos (…) 4. Realización de evaluaciones de datos (…) 5. Estimación del ROI para mejorar la calidad de los datos frente al coste de no hacer nada (…) Establecer la responsabilidad de la calidad de los datos."

II.XVIII. Inteligencia Artificial

Como se señala en La inteligencia artificial (IA) es la inteligencia llevada a cabo por máquinas. En ciencias de la computación, una máquina «inteligente» ideal es un agente flexible que percibe su entorno y lleva a cabo acciones que maximicen sus posibilidades de éxito en algún objetivo o tarea (…) Coloquialmente, el término inteligencia artificial se aplica cuando una máquina imita las funciones «cognitivas» que los humanos asocian con otras mentes humanas, como por ejemplo: «percibir», «razonar», «aprender» y «resolver problemas» (…) Andreas Kaplan y Michael Haenlein definen la inteligencia artificial como «la capacidad de un sistema para interpretar correctamente datos externos, para aprender de dichos datos y emplear esos conocimientos para lograr tareas y metas concretas a través de la adaptación flexible» (…) A medida que las máquinas se vuelven cada vez más capaces, tecnología que alguna vez se pensó que requería de inteligencia se elimina de la definición. Por ejemplo, el reconocimiento óptico de caracteres ya no se percibe como un ejemplo de la «inteligencia artificial» habiéndose convertido en una tecnología común (…) Avances tecnológicos todavía clasificados como inteligencia artificial son los sistemas de conducción autónomos o los capaces de jugar al ajedrez o al Go (…) Según Takeyas (2007) la IA es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos con base en dos de sus características primordiales: el razonamiento y la conducta (…) En 1956, John McCarthy acuñó la expresión «inteligencia artificial», y la definió como «la ciencia e ingenio de hacer máquinas inteligentes, especialmente programas de cómputo inteligentes» (…) También existen distintos tipos de percepciones y acciones, que pueden ser obtenidas y producidas, respectivamente, por sensores físicos y sensores mecánicos en máquinas, pulsos eléctricos u ópticos en computadoras, tanto como por entradas y salidas de bits de un software y su entorno software (…) Varios ejemplos se encuentran en el área de control de sistemas, planificación automática,

la habilidad de responder a diagnósticos y a consultas de los consumidores, reconocimiento de escritura, reconocimiento del habla y reconocimiento de patrones. Los sistemas de IA actualmente son parte de la rutina en campos como economía, medicina, ingeniería, el transporte, las comunicaciones y la milicia, y se ha usado en gran variedad de aplicaciones de software, juegos de estrategia, como ajedrez de computador, y otros videojuegos.

II.XIX. Aprendizaje Automático

Popularmente conocido como "Machine Learning" dados los procesos de transculturización aparejados con la globalización aparejada a la difusión internacional del internet y las distintas tecnologías de la información y las comunicaciones, es una disciplina científica ligada a las Ciencias de la Computación y la Inteligencia Artificial. Según se señala en (Wikipedia, 2020), "El aprendizaje automático o aprendizaje automatizado o aprendizaje de máquinas (del inglés, machine learning) es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan. Se dice que un agente aprende cuando su desempeño mejora con la experiencia; es decir, cuando la habilidad no estaba presente en su genotipo o rasgos de nacimiento (…) De forma más concreta, los investigadores del aprendizaje de máquinas buscan algoritmos y heurísticas para convertir muestras de datos en programas de computadora, sin tener que escribir los últimos explícitamente. Los modelos o programas resultantes deben ser capaces de generalizar comportamientos e inferencias para un conjunto más amplio (potencialmente infinito) de datos (…) En muchas ocasiones el campo de actuación del aprendizaje automático se solapa con el de la estadística inferencial, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el aprendizaje automático incorpora las preocupaciones de la complejidad computacional de los problemas2. Muchos problemas son de clase NP-hard, por lo que gran parte de la investigación realizada en aprendizaje automático está enfocada al diseño de soluciones factibles a esos problemas. El aprendizaje automático también está

estrechamente relacionado con el reconocimiento de patrones. El aprendizaje automático puede ser visto como un intento de automatizar algunas partes del método científico mediante métodos matemáticos. Por lo tanto, es un proceso de inducción del conocimiento (…) El aprendizaje automático tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica."

II.XX. Entrenamiento de un Conjunto de Datos

Como se señala en (Amazon Web Services, 2020), "El proceso de entrenamiento de un modelo de ML consiste en proporcionar datos de entrenamiento de los cuales aprender a un algoritmo de ML (es decir, el algoritmo de aprendizaje). El término modelo de ML se refiere al artefacto de modelo que se crea en el proceso de entrenamiento (…) Los datos de entrenamiento deben contener la respuesta correcta, que se conoce como destino o atributo de destino. El algoritmo de aprendizaje encuentra patrones en los datos de entrenamiento que asignan los atributos de los datos de entrada al destino (la respuesta que desea predecir) y genera un modelo de ML que captura dichos patrones (…) Puede utilizar el modelo de ML para obtener predicciones sobre datos nuevos para los que no se conoce la respuesta de destino. Por ejemplo, si desea entrenar un modelo de ML para que prediga si un mensaje de correo electrónico es spam o no. Le proporcionaría datos de entrenamiento a Amazon ML que contienen correos electrónicos para los que conoce el destino (es decir, una etiqueta que indica si un mensaje es spam o no). Amazon ML entrenaría un modelo de ML mediante la utilización de estos datos, lo que se traduce en un modelo que intenta predecir si los correos electrónicos nuevos son spam o no."[11]

[11] "ML" hace alusión a "Machine Learning", *i.e.*, Aprendizaje Automático.

II.XXI. Separación de Datos. Conjunto de Entrenamiento y Conjunto de Prueba

Desde una perspectiva elemental[12], como se señala en (Google Developers, 2020), los conjuntos de datos pueden ser divididos en dos subconjuntos:

1) El conjunto de entrenamiento: Un subconjunto para entrenar un modelo.
2) El conjunto de prueba: Un subconjunto para probar el modelo entrenado.

El lector puede imaginar dividir un único conjunto de datos de la siguiente manera:

Figura 10

Fuente: (Google Developers, 2020).

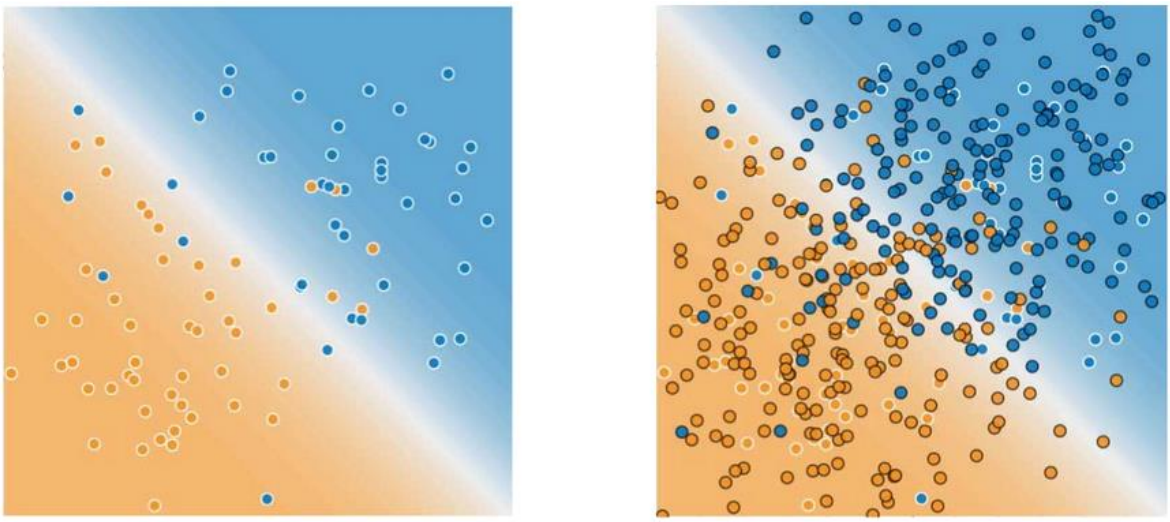
El conjunto de datos de prueba debe cumplir dos condiciones:

1) Debe ser lo suficientemente grande como para generar resultados significativos desde el punto de vista estadístico.
2) Debe ser representativo de todo el conjunto de datos. En otras palabras, no se debe elegir un conjunto de datos con características diferentes al del conjunto de entrenamiento.

[12] En realidad, omitiendo toda simplificación, faltaría mencionar un tercer grupo, que es el grupo de validación. De hecho, idealmente el grupo de prueba tiene que ser "ciego", es decir, no puede ser conocido por el investigador, quien únicamente trabaja con los dos primeros primeros grupos, *i.e.*, el de entrenamiento y el de validación. Solamente hasta que el investigador ha finalizado la construcción del modelo estadístico a utilizar, se alimenta tal modelo con los datos del conjunto de prueba, con la finalidad de evitar un sesgo en la investigación.

Si se supone que el conjunto de prueba reúne estas dos condiciones, el objetivo del investigador es crear un modelo que generalice los datos nuevos de forma correcta desde el punto de vista estadístico. El conjunto de prueba sirve como "proxy" para los nuevos datos. Por ejemplo, considérese la siguiente figura que expresa la validación del modelo entrenado con los datos de prueba:

Figura 11



Datos de entrenamiento

Datos de prueba

Fuente: (Google Developers, 2020).

Como se puede observar, el modelo aprendido para los datos de entrenamiento es muy simple. Este modelo no hace un trabajo perfecto. Algunas predicciones son incorrectas. Sin embargo, este modelo funciona de la misma manera tanto en los datos de prueba como en los de entrenamiento. En otras palabras, el modelo simple descrito anteriormente no sobreajusta[13] los datos de entrenamiento.

[13] Como se señala en (Google Developers, 2020), sobreajustar los datos tiene que ver con la creación de un modelo que coincide de tal manera con los datos de entrenamiento que no puede realizar predicciones correctas con datos nuevos. La intuición detrás de esto es que el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo. Así, en términos del Aprendizaje Automático, la función objetivo cristaliza conceptualmente la métrica que el algoritmo

II.XXII. Etiquetas de un Conjunto de Datos

Como se señala en (Bhattacharjee, 2017), "Labels are the final output. You can also consider the output classes to be the labels. When data scientists speak of labeled data, they mean groups of samples that have been tagged to one or more labels."

En la misma línea, se señala en (Python Programming, 2020) que "With supervised learning, you have features and labels. The features are the descriptive attributes, and the label is what you're attempting to predict or forecast. Another common example with regression might be to try to predict the dollar value of an insurance policy premium for someone. The company may collect your age, past driving infractions, public criminal record, and your credit score for example. The company will use past customers, taking this data, and feeding in the amount of the "ideal premium" that they think should have been given to that customer, or they will use the one they actually used if they thought it was a profitable amount (…) Thus, for training the machine learning classifier, the features are customer attributes, the label is the premium associated with those attributes (…) In our case, what are the features and what is the label? We're trying to predict the price, so is price the label? If so, what are the featuers? When it comes to forecasting out the price, our label, the thing we're hoping to predict, is actually the future price. As such, our features are actually: current price, high minus low percent, and the percent change volatility. The price that is the label shall be the price at some determined point the future."

---

intenta optimizar. Otra forma de ver el sobreajuste estadístico es como un problema relacionado a la memorización de datos. Cuando el investigador posee un modelo modelo que se ajusta perfectamente a los datos seguramente va a tener problemas para predecir un dato distinto a los de entrenamiento, porque está diseñado para predecir específicamente esos a la perfección.

II.XXIII. Verdad Fundamental (Aprendizaje Automático)

La *verdad fundamental* es un conjunto de prueba que el investigador analiza y compara con los resultados de la clasificación realizada por su modelo estadístico. En palabras simples, puede concebirse como la pauta, conocida con antelación, que sirve de criterio para "calificar" la precisión predictiva del modelo estadístico utilizado por el investigador. En suma, la verdad fundamental es conocer a priori el resultado esperado de la predicción.

La traducción que aquí se hace del término inglés "ground truth" está validada porque tal y como señala (Wikipedia, 2020), según el *Oxford English Dictionary* esta palabra tiene sus raíces históricas en la noción de "fundamental truth" contenida en un poema de Henry Ellison titulado "El Cuento del Exilio Siberiano" publicado en 1833.

Figura 12

82. *And be, and do*, for without it he's naught!
Without it Wisdom, Action, Life, is none!
Now as by Nature this Belief is wrought
Out in him, nay, as *she herself* alone
Lives *in him*, as the *Groundtruth* of *her own*
Existence it must be regarded, thro'
Him in its highest, purest Aspect shown!
And he in this full Feeling calm and true
Of the great Whole, regards but as a few Grains to
83. The Seasands added, all the Wonders by
The Pen of History recorded! for
He feels God's Presence in him evernigh,
The greatest Wonder, such as Eye ne'er saw,
Nor Thought conceived! now Wonders *'gainst* the Law
Of Nature God worked out in Pity to
Man's *Frailty*, but he claims far higher Awe
For those wrought *quietly by it*, the tru-
Est, suitablest, and which *He* most delights to do!

Fuente: (Ellison, 2020).

Parece que en algunos lugares la gente entiende que las verdades fundamentales implican en su comprensión un poco de poesía.

II.XXIV. Aprendizaje Supervisado

Como se señala en (Ranganathan, Nakain, Schönback, & Gribskov, 2019, pág. 275), el aprendizaje supervisado puede ser realizado para conjuntos de datos de entrenamiento en que las etiquetas u objetivos para cada muestra. Puede pensarse en esto como un escenario de maestro-alumno, en donde el alumno aprende sobre una temática mediante la retroalimentación que recibe con base en la precisión de sus respuestas. De manera similar, un modelo de aprendizaje supervisado aprenderá de la "verdad fundamental" dada por el conjunto de datos de entrenamiento con la finalidad de construir un modelo generalizado que pueda ser aplicado para predecir las etiquetas de los nuevos datos. Precisamente en eso radica la diferencia entre aprendizaje supervisado y no supervisado, en que en este

último no se conoce la "verdad fundamental", como se expandirá en el siguiente acápite.

II.XXV. Aprendizaje No Supervisado

Como se señala en (Salian, 2018), "In a supervised learning model, the algorithm learns on a labeled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data. An unsupervised model, in contrast, provides unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own.", añadiendo que "Unsupervised machine learning finds all kind of unknown patterns in data." (Guru99, 2020).

Según (Ranganathan, Nakain, Schönback, & Gribskov, 2019, pág. 250), el aprendizaje no supervisado es fundamentalmente "clustering", mientras que el aprendizaje supervisado radica en regresión y clasificación[14]. Es precisamente sobre el "clustering" o análisis de grupos que versa el núcleo de esta investigación.

II.XXVI. Análisis de Grupos

Como se señala en (Ranganathan, Nakain, Schönback, & Gribskov, 2019, pág. 437), el término *análisis de grupos* (en inglés "clustering", de "cluster analysis") en su significado más general hace referencia a la metodología de particionar elementos en grupos de acuerdo con algunas características comunes. El análisis de grupos fue fundado por Driver y Kroeber (1932), siendo aplicadas en la Psicología por Zubin (1938) y Tryon (1939) a causa de la necesidad de caracterizar la tipología (estudio y clasificación de tipos) de diferentes individuos y culturas. Luego, fue utilizado en 1943 por Cattell para tratar la teoría de la clasificación de la personalidad. Sin embargo, el análisis de grupos no alcanzó el estatus de rama de la Estadística y de las Ciencias de la Computación importante hasta finales de la

[14] Se asume que el lector tiene al menos nociones fundamentales de regresión, es decir, que está familiarizado al menos con la regresión lineal; así, en el contexto del Aprendizaje Automático, la regresión es el resultado numérico concreto de un proceso de naturaleza estadística-matemática. Así, la clasificación es la separación de los elementos de un conjunto de datos en subconjuntos con características comunes (denominadas ***clases***) y cuyo resultado final es una clase, no necesariamente un valor numérico concreto.

década como resultado de la investigación realizada por Sokal en 1963. Luego de ello, diferentes libros de texto fueron publicados sobre esta temática, entre los cuales destacan "Les bases de la classification automatique" (Lerman, 1970), "Mathematical Taxonomy" (Jardine y Sibson, 1971) y "Cluster analysis for applications" (Anderberg, 1973), los cuales formalizaron el problema abordado por el análisis grupos y presentaban la metodología de este abordaje.

Desde 1960 el análisis de grupos ha sido utilizado en diferentes disciplinas, incluidas Biología, Psicología, Antropología, Ciencias Políticas, Sociología, Economía y Geografía. En particular, el análisis de grupos ha sido aplicado en la Biología para el conteo de partículas de polvo y bacterias, en Geografía ha sido empleado para resolver problemas de localización de tierras en planificación urbana y en Ciencias Políticas como apoyo a tareas de campaña política, por mencionar algunos ejemplos.

En la actualidad, el análisis de grupos se ha convertido en un instrumento válido para resolver problemas complejos en Estadística y Ciencias de la Computación. En particular, es ampliamente utilizado en minería de datos y resulta efectivo en el descubrimiento de patrones de interés específico que sigue un conjunto de datos y con ello convertirse en una herramienta de investigación valiosa.

El tipo específico de tarea (de preprocesamiento) depende principalmente del tipo de análisis de grupos que se utilice. Las tareas de preprocesamiento pueden incluir:

1) Remover ruido o valores atípicos de los datos, lo que es esencial cuando el tipo de análisis de grupos utilizado es sensible a ruido o valores atípicos.
2) Normalización de datos, lo que es importante cuando se emplea un análisis de grupos basado en la distancia entre los elementos que componen el conjunto de datos.

3) Reducción de datos (muestreo o reducción de atributos), que puede ser útil cuando el tipo de análisis de grupos es costoso en términos computacionales.

La normalización consiste en escalar el conjunto de datos para que oscilen en un determinado rango. La reducción de los datos puede remover atributos irrelevantes mediante la selección de atributo (conocido también como selección de variable, selección de subconjunto, selección de característica) o el análisis de componentes principales (técnica estadística utilizada para describir un conjunto de datos en términos de nuevas variables - "componentes"- no correlacionadas[15]), el cual permite representar el conjunto de datos en un estructura matemática de menor dimensionalidad (espacio de menor dimensión) así como también remover instancias de datos mediante la utilización de algún método de muestro.

Como se señala en (Pang-Ning, Steinbach, & Kumar, 2014, pág. 490), el análisis de conglomerados agrupa los datos basándose únicamente en la información que se encuentra en los datos que describen a los objetos y sus relaciones. El objetivo es que los objetos dentro de un grupo sean similares (o relacionados) entre sí y diferentes (o no relacionados) con los objetos de otros grupos. Cuanto mayor sea la similitud (u homogeneidad) dentro de un grupo y cuanto mayor sea la diferencia entre los grupos, mejor o más distinta será la agrupación. En muchas aplicaciones, la noción de agrupación no está bien definida. Para comprender mejor la dificultad de decidir qué constituye un grupo, considere la figura que en la fuente referida aparece referenciada por el número 8.1., la cual muestra veinte puntos y tres formas diferentes de dividirlos en grupos. Las formas de los marcadores indican la pertenencia a un grupo. Las figuras 8.1 (b) y 8.1 (d) dividen los datos en dos y seis partes, respectivamente. Sin embargo, la aparente división de cada uno de los dos grupos más grandes en tres subgrupos (o subconglomerados, subclústeres o

[15] Véase (Wikipedia, 2020).

"subclusters") puede ser simplemente un artefacto del sistema visual humano[16]. Además, puede que no sea irracional decir que los puntos forman cuatro grupos, como se muestra en la Figura 8.1 (c). Esta figura ilustra que la definición de un conglomerado es imprecisa y que la mejor definición depende de la naturaleza de los datos y los resultados deseados. El análisis de conglomerados está relacionado con otras técnicas que se utilizan para dividir objetos de datos en grupos. Por ejemplo, la agrupación en clústeres puede considerarse una forma de clasificación en el sentido de que crea un etiquetado de objetos con etiquetas de clase (clúster). Sin embargo, deriva estas etiquetas solo de los datos.

Finalmente, como se señala en (Pang-Ning, Steinbach, & Kumar, 2014, págs. 491-492), existe análisis de grupos de carácter jerárquico y de carácter particional, conocidos también como conglomerados anidados y conglomerados no anidados. Un conglomerado particional es simplemente una división del conjunto de datos en conjuntos que no se traslapan (clústeres) tal que cada observación es exactamente un conjunto. Tomadas individualmente, cada colección de clústeres como las figuras b y d (contenidas en la figura 8.1 de la fuente referida) es un clúster particional (es decir, que tampoco se traslapan -conjuntos disjuntos-). Si se permite que los clústeres tengan subclústeres, entonces se obtiene un análisis jerárquico de conglomerados, el cual consiste en un conjunto de clústeres anidados (jerarquizados) que son organizados como un árbol de decisión. En tal árbol[17], cada nodo (grupo) en el árbol (excepto los nodos hoja) es la unión de sus hijos (subgrupos), y la raíz del árbol es el grupo que contiene todos los objetos. A menudo, pero no siempre, las hojas del árbol son grupos únicos de objetos de

[16] O es una metáfora del autor en la fuente citada o implica una concepción idealista o quizás metafísica de la Filosofía de las Ciencias.

[17] Como se señala en (Adamchik, 2020), un árbol binario está formado por nodos, donde cada nodo contiene una referencia "izquierda", una referencia "derecha" y un elemento de un conjunto de datos. El nodo superior del árbol se llama raíz. Cada nodo (excluyendo una raíz) en un árbol está conectado por un borde dirigido desde exactamente otro nodo. Este nodo se llama padre. Por otro lado, cada nodo puede conectarse a un número arbitrario de nodos, llamados hijos. Los nodos sin hijos se llaman nodos hoja (u hojas) o nodos externos. Los nodos que no son hojas se denominan nodos internos. Los nodos con el mismo padre se denominan hermanos.

datos individuales. En el escenario de conglomerados anidados, una interpretación de la Figura 8.1 (a) es que tiene dos subgrupos (Figura 8.1 (b)), cada uno de los cuales, a su vez, tiene tres subgrupos (Figura 8.1 (d)). Los conglomerados que se muestran en las Figuras 8.1 (a – d), cuando se toman en ese orden, también forman un conglomerado jerárquico (anidado) con, respectivamente, 1, 2, 4 y 6 conglomerados en cada nivel. Por último, tenga en cuenta que un agrupamiento jerárquico puede verse como una secuencia de agrupamientos particionales y un agrupamiento particional puede obtenerse tomando cualquier miembro de esa secuencia; es decir, cortando el árbol jerárquico en un nivel particular.

Figura 13

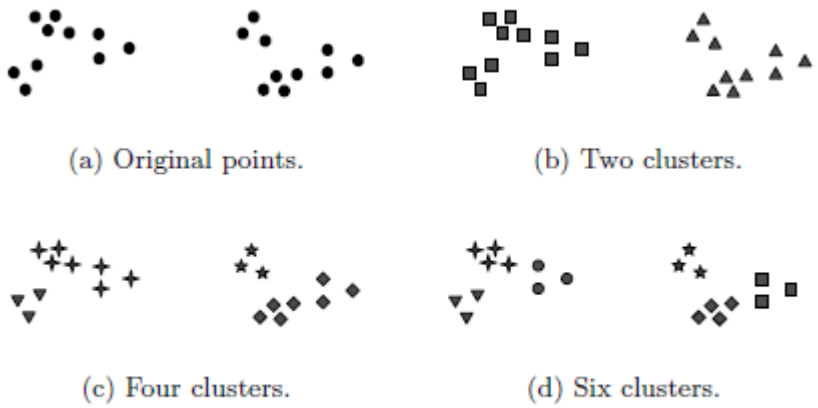


Figure 8.1. Different ways of clustering the same set of points.

Fuente: (Pang-Ning, Steinbach, & Kumar, 2014, pág. 491).

Por el contrario, la clasificación realizada mediante árboles de decisión y procesos modelos de evaluación se considera clasificación supervisada (estudiada en el capítulo 4 de la última fuente referenciada); es decir, a los objetos nuevos no etiquetados se les asigna una etiqueta de clase utilizando un modelo desarrollado a

partir de objetos con etiquetas de clase conocidas. Por esta razón, el análisis de conglomerados a veces se denomina clasificación no supervisada. Cuando el término clasificación se usa sin ninguna calificación dentro de la minería de datos, generalmente se refiere a la clasificación supervisada. Además, aunque los términos segmentación y particionamiento a veces se usan como sinónimos de agrupamiento, estos términos se usan con frecuencia para enfoques fuera de los límites tradicionales del análisis de conglomerados. Por ejemplo, el término partición se utiliza a menudo en relación con técnicas que dividen gráficos en subgrafías y que no están estrechamente relacionadas con la agrupación. La segmentación a menudo se refiere a la división de datos en grupos utilizando técnicas simples; Por ejemplo, una imagen se puede dividir en segmentos según la intensidad y el color de los píxeles, o las personas se pueden dividir en grupos según sus ingresos. No obstante, algunos trabajos en la partición de gráficos y en la segmentación de imágenes y mercados están relacionados con el análisis de clústeres.

II.XXVII. Mixturas

Según lo expuesto en (Nabi, Algunas Reflexiones Sobre la Distribución Binomial Negativa II (Un Análisis Teórico y Aplicado), 2020, pág. 34), las mixturas de probabilidad son funciones de distribución de probabilidad en que uno o más parámetros de la distribución de probabilidad (que puede ser univariante o multivariante, pero siempre implica probabilidad condicional y una modelación teórica del proceso de estimación de las distribuciones de probabilidad que responde a un diseño por etapas jerarquizadas) se modelan con base al comportamiento de alguna otra variable aleatoria, que condiciona implícitamente el comportamiento de la variable aleatoria estudiada. Estos modelos buscan capturar la noción de variable latente, variable implícita, variable oculta.

II.XXVIII. Estimación de Densidad

El concepto de "estimación" tiene también sus raíces en la Física. Según (Wikipedia, 2020), es el proceso de encontrar una estimación o aproximación, que es un valor que se puede utilizar para algún propósito, incluso si los datos de entrada pueden ser incompletos, inciertos o inestables. No obstante, el valor es utilizable porque se deriva de la mejor información disponible. Normalmente, la estimación implica utilizar el valor de una estadística derivada de una muestra para estimar el valor de un parámetro poblacional correspondiente. La muestra proporciona información que puede proyectarse, a través de varios procesos formales o informales, para determinar un rango que probablemente describa la información faltante. Una estimación que resulta ser incorrecta será una sobreestimación si la estimación excedió el resultado real, y una subestimación si la estimación no alcanzó el resultado real (contrasta con la predicción).

Así, como se señala en (Wikipedia, 2020), la estimación de densidad es una construcción de una estimación estadística, con base en los datos observados, de una función de densidad de probabilidad subyacente no observable[18]. La función de densidad no observable se considera la densidad según la cual se distribuye una gran población; generalmente, los datos de los que dispone el investigador son considerados como una muestra aleatoria de esa gran población mencionada anteriormente. Para estimar las densidades de un conjunto de datos se utiliza una gran variedad de enfoques, el que en esta investigación se estudia es la técnica mediante análisis de grupos utilizando mixturas gaussianas finitas, las cuales se expondrán más adelante.

La esencia lógica de esta técnica radica en la intuición geométrica subyacente en la Estadística y prueba de ello dan las mismas raíces históricas de la estimación de densidad. Como se expone en (Scott, 2015, pág. 4), estas raíces se localizan en el descubrimiento de la correlación hecho por Galton en 1886. Como señala el autor,

[18] Aquí aparece la relevancia de los modelos de variables latentes y las mixturas.

Karl Pearson, padre de la Estadística Matemática, era un exponente representativo de las representaciones geométricas de la Estadística, lo que se verifica por sus propias palabras, pronunciadas en una lectura realizada un 11 de noviembre de 1891 citadas en la localización referida: "Erroneous opinion that Geometry is only a means of popular representation: *it is afundamental method of investigating and analysing statistical material*." Las cursivas son autoría de Karl Pearson. Así, de la lógica esgrimida por Karl Pearson aparecen los histogramas, diagramas, cartogramas, estereogramas, topogramas, estigmogramas, radiogramas, epipedogramas, hormogramas, eutigramas, entre otros.

Como se señala en (Silverman, 1998, pág. 1), una forma de realizar estimación de densidad es mediante métodos paramétricos. Esto consiste en asumir que el conjunto de datos proviene de una familia de distribuciones de probabilidad paramétricas conocidas, por ejemplo, la distribución Normal cuyos parámetros son μ y σ^2. La densidad f subyacente al conjunto de datos puede ser estimada mediante el siguiente proceso: se realizan las estimaciones de μ y σ^2 y se sustituyen tales estimaciones en la fórmula de la densidad Normal que se ha puesto como ejemplo. Sin embargo, existe otro enfoque menos rígido, el que abandera a la familia de métodos no paramétricos.

En la metodología no paramétrica de estimación de densidad se asume únicamente que el conjunto de datos tiene una densidad de probabilidad f, sin restringir el estudio del conjunto de datos a valores específicos de los parámetros. El análisis de grupos es, en general[19], una metodología estadística que pertenece a los métodos

[19] Existen metodologías estadísticas relativamente recientes que plantean algo sensiblemente diferente. "Clustering analysis is widely used in many fields. Traditionally clustering is regarded as unsupervised learning for its lack of a class label or a quantitative response variable, which in contrast is present in supervised learning such as classification and regression. Here we formulate clustering as penalized regression with grouping pursuit. In addition to the novel use of a non-convex group penalty and its associated unique operating characteristics in the proposed clustering method, a main advantage of this formulation is its allowing borrowing some well-established results in classification and regression, such as model selection criteria to select the number of clusters, a difficult problem in clustering analysis. In particular, we propose using the generalized cross-validation (GCV) based on generalized degrees of freedom (GDF) to select the number of

del Aprendizaje Automático No Supervisado. Por otro lado, el análisis de grupos puede ser paramétrico o no paramétrico, tal y como se señala en (Roberts, 1997, pág. 261). Como se dijo anteriormente, la clasificación es dominio del Aprendizaje Automático Supervisado. El análogo de la clasificación en el Aprendizaje Automático No Supervisado es el análisis de grupos, como se señala en (Wikipedia, 2020), lo cual constituye la metodología estudiada en esta investigación. Además, en esta investigación se realiza un abordaje paramétrico del problema.

A partir de aquí el lector requerirá caminar guiado de su intuición (que deberá estar fundamentada en un sentido común científico adecuadamente fundamentado y la lógica estadística-matemática) y por ningún motivo soltarse de ella en el recorrido restante.

II.XXIX. Señal

Una señal es una función que provee al investigador información sobre un fenómeno natural o social. Sin embargo, esa es una definición muy limitada. En su sentido pleno, una señal es una construcción matemática que cristaliza la noción dialéctica-materialista de *reflejo*. A continuación, se presentarán las definiciones formales de una señal, fundamentalmente las concernientes a los campos de la Ingeniería Eléctrica (específicamente en un subcampo de sí conocido como *Procesamiento de Señales*) y la Estadística, para luego pasar a definirlas desde el Materialismo Dialéctico.

(Wikipedia, 2020) señala que "Signal processing is an electrical engineering subfield that focuses on analysing, modifying, and synthesizing signals such as sound, images, and scientific measurements (…) Signal processing techniques can be used to improve transmission, storage efficiency and subjective quality and to

---

clusters. We use a few simple numerical examples to compare our proposed method with some existing approaches, demonstrating our method's promising performance." (Pan, Shen, & Liu, 2013, pág. 1865).

also emphasize or detect components of interest in a measured signal (…)" Con fines de construir una definición general para señal (sin entrar en consideración de los tipos que existen sobre la misma) el lector debe leer la definición anterior como "Procesamiento de señales es el proceso que se encarga de analizar, modificar y sintetizar (entiendo por esta palabra en el sentido de superación dialéctica -el equivalente en el terreno de lo heterogéneo al *Aufheben*-) señales (definidas estas como mediciones científicas de cualquier índole).

El procesamiento de señales puede ser utilizado para mejorar la transmisión de señales, la eficiencia del almacenamiento de señales y facilitar que el investigador realice valoraciones sobre la calidad de las señales de acuerdo a su criterio experto (entendido este como su opinión profesional basada su experiencia como investigador en el campo, su lógica científica y su intuición, todo formado con base a sus condiciones genéticas, biológicas y a un proceso de socialización del que fue criatura y no demiurgo, aunque algunos con profunda ingenuidad se crean por encima de su proceso de socialización[20]).

Ahora el lector inexorablemente deberá estar preguntándose ¿y qué son entonces las señales?, y es que estas, más allá de ser mediciones científicas de cualquier índole, lo cual es una definición que únicamente de forma provisional resulta completamente satisfactoria debido a su nivel de abstracción y generalidad

[20] Si se quiere ampliar sobre la posición del investigador en Filosofía de la Estadística y en Filosofía de la Ciencias en general, pueden consultarse los anexos de (Nabi, Algunas Reflexiones Sobre la Distribución Binomial Negativa II (Un Análisis Teórico y Aplicado), 2020).

excesivos, son "(…) a function representing a physical quantity or variable, and typically it contains information about the behavior or nature of the phenomenon. For instance, in an RC[21] circuit the signal may represent the voltage across the capacitor or the current flowing in the resistor. Mathematically, a signal is represented as a function of an independent variable t. Usually t represents time. Thus, a signal is denoted by x(t)." (Hsu, 2011, pág. iii). Hasta aquí sólo se ha avanzado en su clarificación lógico-formal y aplicado con un grado de generalidad no despreciable (debido a todos los modelos que dependen del tiempo en diferentes disciplinas científicas), sin embargo, falta más generalidad y menos abstracción.

Como se mencionó anteriormente, es posible realizar una interpretación del significado de las señales en términos estadísticos. En ese sentido, "*A signal is a set of data or information*. Examples include a telephone or a television signal, monthly sales of a corporation, or daily closing prices of a stock market (e.g., the Dow Jones averages). In all these examples, *the signals are functions of the independent variable time*. *This is not always the case, however*. When an electrical charge is distributed over a body, for instance, the signal is the charge density, a function of space rather than time. In this book we deal almost exclusively with signals that are functions of time. *The discussion, however, applies equally well to other independent variables*." (Lathi & Green, 2018, pág. 64)[22].

Ahora no deben existir dudas sobre la definición de una señal en su sentido aplicado y su sentido lógico-formal, sin embargo, falta ahora ligar todas las definiciones anteriores en su interconexión dialéctica, para lo que es necesario

[21] "Resistor-capacitor", que se escribe de igual forma en el idioma inglés.
[22] Cursivas añadidas por el autor de la presente investigación con la finalidad de destacar los componentes esenciales de la definición y sus aspectos teóricos complementarios.

construir un envoltorio teórico más robusto y flexible, más general y abstracto, sin perder el nivel de concreción alcanzado con antelación.

Según (Rosental & Iudin, 1971, pág. 393), un reflejo "Es uno de los conceptos fundamentales de la gnoseología materialista. El materialismo dialéctico diferencia el reflejo psíquico como propiedad de la materia altamente desarrollada y la propiedad general del reflejo, inherente a la materia toda. El reflejo psíquico surge como resultado de la incidencia de los objetos sobre el aparato reflectante de los animales y del hombre, por la reelaboración analítico-sintética de las huellas de tales incidencias, así como por el empleo de los productos reelaborados en calidad de sustitutos, representaciones o modelos de los objetos. Gracias a los modelos de las cosas y de sus propiedades, el sujeto se orienta en el medio que le rodea. El reflejo psíquico tiene dos aspectos: 1) el contenido del reflejo o imagen[23], y 2) el

[23] En este caso "imagen" representaría lo inverso que representa matemáticamente, es decir, representaría el contenido del reflejo, que encuentra su análogo matemático en el dominio o conjunto de entrada del mapeo relacional (función), mientras que en Matemáticas se entiende por "imagen" el conjunto de salida o algún subconjunto de este último.

modo de su existencia material, es decir, los procesos para reelaborar la acción de los objetos sobre el aparato reflector[24].

El contenido del reflejo psíquico está caracterizado por dos rasgos fundamentales: 1) la relación de isomorfismo[25] existente entre la huella en el aparato reflectante y

---

[24] Esto es equivalente al procesamiento de señales. La síntesis del ser humano (Ser Para Sí, *Lo Particular*) con la Naturaleza (Ser En Sí, *Universal Abstracto*) sólo es alcanzada cuando el ser humano aprehende a la Naturaleza a través de la razón (entendiendo lo anterior en su proceso, como *Universal Concreto*, *Ser En Sí-Para Sí*), pero en tal proceso el ser humano recibe influencia sobre su aparato receptor (el cerebro, lo cual hace que "su instrumento de medición sea sesgado -no necesariamente este sesgo es favorable o desfavorable, dependerá ello del contexto, de la magnitud y de su combinación dialéctica con los factores intrínsecos a la naturaleza del caso estudiado- y con ello la estructura analítica bajo la cual estudia al fenómeno natural o social en cuestión sea alterada y así se alteren -favorable o desfavorablemente- los resultados del estudio. Usualmente y de forma sumamente ingenua se huye de los sesgos, cuando más allá de su inevitabilidad absoluta, los instrumentos de medición no son perfectos (*i.e.*, los mismos instrumentos de medición están sesgados por su imperfección) los sesgos personales pueden guiar al investigador por caminos que de otra forma -si fuese un autómata sin sesgos- no habría transitado). Sin embargo, es siempre deseable que el investigador estudie los fenómenos dentro de un marco analítico y experimental estandarizado (aquí entra en juego el Método Científico, que ha evolucionado con el paso del tiempo -igual que las ciencias mismas-) que permita una recalibración de sus sesgos personales, es decir, que el investigador debe realizar sus estudios en el marco de estructuras analíticas y de experimentación que corrijan y faciliten corregir la generalidad de sesgos indeseables (que sean indeseables por su naturaleza y/o por su magnitud). Esta recalibración del aparato receptor, es decir, esta recalibración de los sesgos personales del investigador (contenidos en su psique como resultado de un proceso específico de socialización) por un conjunto de reglas que conocemos como Método Científico (y las reglas inherentes a la disciplina científica en que radique el investigador, así como otros mecanismos que para casos particulares pudiesen sumar en ese sentido), es el procesamiento de señales del que hablan los filósofos soviéticos citados.

Esto sería equivalente a la forma que toman los fenómenos naturales o sociales que estudia el investigador. Esto a su vez encuentra su análogo matemático en la misma estructura matemática de la función, que es equivalente a su vez a la variable dependiente.

[25] Un isomorfismo es, matemáticamente hablando, una función biyectiva que posee una inversa. Desde un aspecto más lógico-formal, un isomorfismo es una relación uno a uno entre los elementos de dos conjuntos de observaciones que representan cada una a una variable aleatoria y en el que para la función matemática que modela dicha relación existe una función inversa. Sea $y = f(x)$, la función inversa $x = f^{-1}(y)$ tiene la característica de que cuando se evalúa el valor de $f(x)$ en $x = f^{-1}(y)$ el resultado es x y cuando se evalúa $f^{-1}(y)$ en $f(x)$ el resultado también es x; esto captura la idea intuitiva de dos fenómenos con la misma estructura interna, lo cual se refleja en su estrecha interrelación . Así, lo que la definición de isomorfismo busca capturar es la idea intuitiva de dos fenómenos (naturales o sociales) que poseen la misma estructura interna (por ello posee inversa) y que como consecuencia de ello, existe entre tales fenómenos una estrecha interrelación que se manifiesta en la correspondencia uno a uno de sus elementos sin importar si se establece la relación partiendo de un conjunto de elementos hacia el otro o viceversa, idea que se cristaliza en el hecho que tanto la función $y = f(x)$ como su inversa $x = f^{-1}(y)$ son funciones biyectivas y que por ello al evaluar una en otra y viceversa se obtiene siempre x, que puede ver vista como el aspecto cuantitativo que toma la esencia del fenómeno dado. En este sentido, un isomorfismo es una "(…)

un aspecto determinado del objeto incidente; en los casos particulares, el isomorfismo presenta diferentes tipos y niveles de semejanza; 2) la propiedad de poseer características de objeto. Esto último significa que, en el contenido del reflejo, no se da al sujeto el estado de sus receptores, nervios y cerebro, como creían los representantes del *idealismo fisiológico*, sino el contenido de los objetos de los objetos del mundo exterior[26]. El contenido objetual aparece directamente para el

---

Relación entre objetos que tienen una estructura igual, idéntica. Dos estructuras (sistemas o conjuntos) son isomorfas entre sí cuando a cada elemento de la primera estructura corresponde sólo un elemento de la segunda, y a cada operación (nexo) de una estructura corresponde una única operación (nexo) en la otra, y recíprocamente. Por lo general, la relación isomórfica caracteriza una de las relaciones o propiedades de los objetos que se comparan. Sólo puede darse el isomorfismo completo entre dos objetos abstractos, por ejemplo, entre una figura geométrica y su expresión analítica bajo el aspecto de fórmula matemática. El concepto de "isomorfismo" se emplea mucho en matemática y también en lógica matemática, en física teórica, en cibernética y en otras esferas del saber. El concepto de "isomorfismo" se halla relacionado con los conceptos de "modelo" (*Modelación*), "señal" e "imagen" (*Reflejo*, *Ideal*)." (Rosental & Iudin, 1971, pág. 249).

[26] Aquí hay que destacar algo de importancia fundamental, y es el hecho de que al decir "(…) en el contenido del reflejo, no se da al sujeto el estado de sus receptores, nervios y cerebro (…) sino el contenido de los objetos de los objetos del mundo exterior" se está, consciente o no de ello, estableciendo la independencia relativa de la realidad con respecto del investigador (el Ser Para Sí está contenido en el Ser En Sí, pero a pesar de ello existe cierta independencia relativa mutua -como la madre que contiene a su hijo, que es parte de ella y a la vez no-) y con ello permitiendo establecer lo que diría el célebre padre francés de la Sociología, Émile Durkheim "No decimos que los hechos sociales son cosas materiales, sino que son cosas con el mismo derecho que las cosas materiales aunque de otro modo. ¿Qué es una cosa? La cosa se opone a la idea como aquello que es conocido desde fuera a aquello que se conoce desde dentro. Es cosa todo objeto de conocimiento que no es naturalmente compenetrable por la inteligencia, todo aquello de lo que no podemos hacernos una noción adecuada por un mero procedimiento de análisis mental, todo aquello que el espíritu no puede llegar a comprender más que a condición de salir de sí mismo, por medio de observaciones y experimentos, pasando progresivamente de los caracteres más exteriores e inmediatamente más accesibles a los menos visibles y más profundos. Tratar a hechos de un cierto orden de cosas no es, pues, clasificarlos en tal o cual categoría de lo real; es observar con respecto a ellos una cierta actitud mental, Es abordar el estudio de los mismos adoptando el principio de que se ignora por completo lo que son y de que tanto sus propiedades características cuanto las causas desconocidas de que dependan no pueden ser descubiertas ni siquiera por la introspección más cuidadosa." (Durkheim, 2009, págs. 37-38). Lo anterior coincide también por lo planteado por el célebre politólogo salvadoreño Dagoberto Gutiérrez: "la realidad es lo que se me resiste. ¿A qué se resiste?, a mi pensamiento", en donde él entiende por "realidad" la esencia del mundo físico, lo que Durkheim considera "lo real", aquí no es necesario entrar en ese nivel de profundidad. Pero, además, en lo planteado por los filósofos soviéticos, se está diciendo que precisamente a causa de esa independencia existe también una diferencia en los determinantes del estado de los receptores del ser humano (el investigador) y el contenido de los objetos del mundo exterior. Finalmente, cuando dicen "El contenido objetual aparece directamente para el sujeto en la forma ideal de reflejo (*Ideal*)" no debe perderse de vista que en ese aparecer se implican también las diferencias existentes entre el *fenómeno* ("Concepto que designa lo que se nos da en la experiencia y conocemos a través de los sentidos." (Rosental & Iudin, 1971, pág. 171)), el conjunto de hechos subyacentes a ese

sujeto en la forma ideal de reflejo (*Ideal*), es decir, bajo la forma de imagen del objeto. La cognición humana se diferencia cualitativamente del reflejo psíquico de los animales por su naturaleza social, que se manifiesta por la presencia de la *conciencia*, relacionada con el lenguaje, y por la transformación activa del mundo exterior. La propiedad general del reflejo[27], inherente a toda la materia, es afín a la sensación gracias a la presencia del rasgo de isomorfismo; pero no es idéntica a la sensación por carecer de la propiedad de ser objetual: las huellas isomorfas en la naturaleza orgánica son muertas, es decir, no se utilizan en la función de modelos, en calidad de instrumentos de orientación[28]. Gracias al carácter isomorfo entre las incidencias y las huellas en la naturaleza inorgánica, la propiedad general del

---

fenómeno (el conjunto general de leyes de carácter objetivo que lo explican -no hay que olvidar que así como los fenómenos evolucionan, sus leyes objetivas explicativas también lo harán en la misma medida, lo que explica la estabilidad temporal de los paradigmas alrededor del cual se constituye cada una de las ciencias normales en el sentido definido por Thomas Kuhn, es decir, "(…) 'ciencia normal' significa investigación basada firmemente en una o más realizaciones científicas pasadas, realizaciones que alguna comunidad científica en particular reconoce, durante cierto tiempo, como fundamento para su práctica posterior." (Kuhn, 2004, pág. 33)-), *i.e.*, el *contenido* subyacente al fenómeno (natural o social) y dentro de ese contenido, el conjunto fundamental de leyes que lo rigen (que es un subconjunto de ese conjunto general de leyes que llamamos contenido), es decir, la *esencia* subyacente del fenómeno natural o social estudiado.

[27] Cuando se refieren a "la propiedad general del reflejo" se están refiriendo a la huella (entendida esta como "rastro, seña, vestigio que deja algo o alguien por donde pasa", "Señal que deja una lámina o forma de imprenta en el papel u otra cosa que se estampa", Impresión profunda y duradera", "Indicio, mención, alusión", como se define en (Real Academia Española, 2020)) que la totalidad deja como "marca distintiva" en sus partes componentes, *i.e.*, las características distintivas del reflejo del todo en sus partes, "el sello que las distingue como de su propiedad". Por supuesto, esta huella o reflejo usualmente no es accesible inmediatamente a la razón (lo que señalaba Durkheim en la referencia realizada), lo que es equivalente al hecho de que existe una contradicción entre cómo los hechos naturales o sociales aparecen y su esencia (lo que Durkheim llama "lo real" y Dagoberto Gutiérrez "la realidad") y es precisamente en este sentido que toma relevancia el *Procesamiento de Señales*, que como subcampo de las ciencias formales cristaliza la idea intuitiva de depuración de las mediciones estadísticas preliminares de un hecho de carácter natural o social. Estas mediciones preliminares pueden considerarse como el fenómeno, es decir, como la fenomenología del hecho.

[28] Por un lado, está viendo a los instrumentos de medición como instrumentos de orientación (tomando la intuición que medimos en espacios de fases, lo que sirve para tener conocimiento de los estados que tomarán los objetos físicos dentro del sistema, que implica a su vez una dirección y sentido -desde la perspectiva de los espacios vectoriales-), lo que en Navegación es conocido como el rumbo y tiene que ver con la orientación del navegante. Por otro lado, es necesario destacar que esto está relacionado con los conceptos de la Geoquímica de Isótopos que es de enorme relevancia en la investigación geológica a causa del papel que en ella juega la utilización de pruebas como la del Carbono-14.

reflejo constituye la base genética del reflejo psíquico[29], con la premisa de su aparición. También es la base dada en toda la naturaleza (física) del proceso en virtud del cual el hombre entra en conocimiento de la realidad circundante, es decir, el hombre en su actividad cognoscitiva, utiliza tanto los resultados inmediatos de la interacción de las cosas como los resultados mediatos, se basa en unos y otros, descubriendo las propiedades y relaciones esenciales de las cosas."

II.XXX. Cuantización

Como se señala en (Wikipedia, 2020), cuantización es "(…) the process of mapping input values from a large set (often a continuous set) to output values in a (countable) smaller set, often with a finite number of elements (…) Quantization is involved to some degree in nearly all digital signal processing, as the process of representing a signal in digital form ordinarily involves rounding (proceso de descartar cifras de una expression decimal). Quantization also forms the core of essentially all lossy compression algorithms (clase de métodos de codificación de datos que utiliza aproximaciones inexactas y descarte parcial de datos para representar el contenido)"

[29] Con "base genética" debe entenderse la base "Perteneciente o relativa a la génesis u origen de las cosas" (Real Academia Española, 2020), es decir, como el origen de los hechos (vistos como un proceso) que condiciona tal proceso y por consiguiente el resultado final de dicho proceso.

II.XXXI. Técnica de Clasificación (Clasificador)

Como se señala en (Pang-Ning, Steinbach, & Kumar, 2014, pág. 148), una técnica de clasificación (o clasificador) es un abordaje sistemático ara construir modelos de clasificación a partir de un conjunto de datos que sirven de insumo para tales fines. Ejemplos de esto son los árboles de decisión, los clasificadores basados en reglas, las redes neuronales artificiales, los clasificadores ingenuos de Bayes y las Máquinas de Vectores de Soporte.

II.XXXII. Teoría del Aprendizaje Estadístico

La teoría del aprendizaje estadístico es el subcampo de las ciencias formales que engloba el marco conceptual y aplicado tanto del Aprendizaje Supervisado como del Aprendizaje No Supervisado. "Broadly speaking, supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs. Problems of this nature occur in fields as diverse as business, medicine, astrophysics, and public policy. With unsupervised statistical learning, there are inputs but no supervising output; nevertheless, we can learn relationships and structure from such data." (James, Witten, Hastie, & Tibshirani, 2013, pág. 1).

II.XXXIII. Maldición del Problema de Dimensionalidad

Este término acuñado por Richard Bellman en 1957 consiste, según sus propias palabras en que: "There are, however, some details to consider. In the first place, the effective analytic solution of a large number of even simple equations as, for example, linear equations, is a difficult affair. Lowering our sights even a computational solution usually has a number of difficulties of both gross and subtle nature. Consequently, the determination of this maximum is quite definitely not routine when the number of is large. All this may be subsumed under the heading "the curse of dimensionality" (Bellman, 1972, pág. ix).

Intuitivamente, como se señala en (Wikipedia, 2020), esta categoría conceptualiza el hecho de que el común denominador de estos problemas radica en que cuando aumenta la dimensionalidad del conjunto de datos, el volumen del espacio aumenta tan rápido que los datos disponibles se vuelven escasos. Esta escasez es problemática para cualquier método que requiera significación estadística. *Para obtener un resultado estadísticamente sólido y confiable, la cantidad de datos necesarios para respaldar el resultado a menudo aumenta exponencialmente con la dimensionalidad*. Además, la organización y la búsqueda de datos a menudo se basan en la detección de áreas donde los objetos forman grupos con propiedades similares; sin embargo, en datos de alta dimensión, todos los objetos parecen ser escasos y diferentes en muchos aspectos, lo que impide que las estrategias comunes de organización de datos sean eficientes.

II.XXXIV. Aridad (Lógica Matemática)

Según (Wikipedia, 2020), es el número de argumentos u operandos tomados por una función u operación en Lógica, Matemáticas y Ciencias de la Computación.

II.XXXV. Función Booleana

Según lo expuesto en (Wikipedia, 2020), una función booleana es una función cuyos argumentos, así como la función en sí misma, asume valores y exclusivamente de un conjunto de elementos cuya cardinalidad o medida es equivalente a 2, tal conjunto usualmente es el conjunto $\{0, 1\}$.

Una función booleana toma la forma $f : \{0, 1\}^k$, en donde k es la aridad de la función f.

II.XXXVI. Concepto (Aprendizaje Automático)

Como se señala en (Stover, Concept, 2020), debe entenderse por "concepto" en el contexto de la Teoría del Aprendizaje Automático y la Inteligencia Artificial lo que se expone a continuación. Un concepto c sobre un dominio X es una función booleana $c : X \to \{0, 1\}$. Una colección de conceptos es denominada *clase de conceptos*. En el contexto de aplicaciones específicas, los conceptos son usualmente pensados para asignar sea un resultado "positivo" o "negativo" (asignando 1 o 0, respectivamente) a cada elemento x del dominio X. En ese sentido, los conceptos son componentes fundamentales de la Teoría del Aprendizaje.

II.XXXVII. Conjuntos Altamente Fragmentados ("Shattered Set")

El término, según los registros históricos disponibles, tiene sus orígenes en la tesis doctoral de Michael Steele presentada en la Universidad de Stanford en 1975 titulada "Entropía Combinatoria y las Leyes Límites Uniformes", la cual se puede localizar en (Steele M. J., 2015). Con base en el espíritu que orquestó su acuñamiento (véase (Steele J. M., 2015)), la traducción más fiel sería "conjuntos destrozados" puesto que busca hacer cristalizar la idea de un nivel de escisión cuántico (a muy pequeña escala), sin embargo, he considerado que la palabra "destrozado" pierde de vista que dentro de ese nivel de escisión existe un orden (que, de hecho, es sumamente férreo, hasta rígido si se quiere -como las mismas Matemáticas-), por lo que llamarles *conjuntos altamente fragmentados* me parece una descripción conceptual más adecuada.

Según (Stover, Shattered Set, 2020), un conjunto altamente fragmentado puede definirse como se expone a continuación. Sea X un conjunto y sea S una colección de subconjuntos de X. Un subconjunto $A \subset X$ es un conjunto altamente fragmentado si cada subconjunto $B \subset A$ puede ser expresado como una intersección de A con algún subconjunto T en S. Simbólicamente, se dice entonces que A es un conjunto altamente fragmentado por S si para todo $B \subset A$ existe algún $T \subset X$ para el cual se cumple que $T \cap A = B$. Si A es altamente fragmentado por S,

se dice que S realiza una alta fragmentación de A. También podría llamársele *fragmentación cuántica* del conjunto A por el conjunto S.

En el campo de la Teoría del Aprendizaje Automático, generalmente se considera que el conjunto A es una muestra de resultados extraídos de acuerdo con una distribución D con el conjunto S representando una colección de conceptos o leyes "conocidas". En ese contexto, decir que A es fragmentado por S cristaliza la idea intuitiva de que todos los resultados (del inglés "outcomes") constituyentes en A pueden ser conocidos solamente conociendo las leyes en S.

Alguna vez en una época pasada casi seguramente le gustó al lector aquella canción que decía "la desnudez es tu mejor lencería", pues con humildad debe cantársele ese mismo fragmento de la canción a la definición de conjuntos fragmentados y para ello habrá que profundizar someramente en su estructura matemática. Como se verifica en (Shashua, 2009, pág. 2), Definición 2, adaptando la simbología de tal definición a la expuesta en esta investigación con base en Wolfram MathWorld cuya cita inaugura este acápite, la fragmentación que S realiza de A está ligada a nivel de su estructura matemática (de la función que modela ese proceso de fragmentación) por un conjunto potencia formado con los elementos del conjunto que es fragmentado, *i.e.*, el conjunto A. Esto a su vez implica que S fragmenta a A si puede agotar los elementos de A seleccionándolos de dos en dos de cualquier forma posible, en palabras textuales de la fuente citada "In other words, S is shattered by C if C realizes all possible dichotomies of S." (Shashua, 2009, pág. 2), recordando que la notación que aquí se ha utilizado lo que en Shashua es C aquí es S y lo que ahí es S aquí es A.

Como se dijo anteriormente, citando a (Stover, Shattered Set, 2020) en el Wolfram MathWorld, "En ese contexto, decir que A es fragmentado por S cristaliza la idea intuitiva de que todos los resultados (del inglés "outcomes") constituyentes en A pueden ser conocidos solamente conociendo las leyes en S.", lo cual significa que conocer las leyes en S es equivalente a poder particionar dualmente y mediante

combinaciones irrestrictas los elementos de A, como resultado de la relación de A y S dentro de X, que es el conjunto que los contiene a ambos. Por supuesto, la interpretación concreta de esa relación de A y S en X dependerá no solamente de la interpretación de la función matemática que las une (localizada en (Stover, Shattered Set, 2020)) sino también del hecho natural o social concreto del que se trate[30].

Los conjuntos altamente fragmentados aparecen en el contexto de la Ley de los Grandes Números Uniforme, que establece la convergencia uniforme de la distribución hacia una media igual a cero bajo determinadas condiciones específicas que puede verificar el lector en (Wikipedia, 2020).

II.XXXVIII. Dimensión de Vapnik-Chernonenkis (Dimensión VC)

Como se señala en (Wikipedia, 2020), es una medida de la capacidad (complejidad, poder expresivo, riqueza o flexibilidad) de un conjunto de funciones que se pueden aprender mediante un algoritmo de clasificación estadística binaria (como los representados mediante funciones booleanas). Se define como la cardinalidad del mayor conjunto de puntos que el algoritmo puede fragmentar. Originalmente fue definido por Vladimir Vapnik y Alexey Chervonenkis. De manera informal, *la capacidad de un modelo de clasificación está relacionada con lo complicado que puede ser*.

[30] Nótese que por sí misma la interpretación de la función que liga al conjunto A con el conjunto S contribuiría a robustecer fundamentalmente la parte general-abstracta de la definición, pero es necesario el contexto científico de estudio (específicamente la naturaleza de las variables estudiadas -las cuales modelan un hecho natural o social de interés-) para robustecer fundamentalmente la parte general-concreta.

Por ejemplo, considere el umbral de un polinomio de alto grado: si el polinomio se evalúa por encima de cero, ese punto se clasifica como positivo, de lo contrario, como negativo. Un polinomio de alto grado puede ser ondulado, por lo que puede ajustarse bien a un conjunto determinado de puntos de entrenamiento. Pero se puede esperar que el clasificador cometa errores en otros puntos, porque es demasiado ondulado.

Tal polinomio tiene una gran capacidad. Una alternativa mucho más simple es establecer el umbral de una función lineal. Es posible que esta función no se ajuste bien al conjunto de entrenamiento porque tiene poca capacidad.

II.XXXIX. Máquinas de Vectores de Soporte (MVS)

Como se señala en (Wikipedia, 2020), son un conjunto de algoritmos de aprendizaje supervisado cuyas raíces históricas se localizan en la investigación teórica y aplicada llevada a cabo por Vladimir Vapnik y su equipo de trabajo en los laboratorios AT&T. Además, como se señala en (Pang-Ning, Steinbach, & Kumar, 2014, pág. 256), es una técnica que tiene sus raíces teóricas en la Teoría del Aprendizaje Estadístico (que es un abordaje filosófico y teórico de Vapnik de la Teoría del Aprendizaje Computacional -busca ser, por consiguiente, una metateoría-) y que ha mostrado resultados empíricos prometedores en muchas aplicaciones, desde reconocimiento de escritos a manos hasta categorización de texto. La filosofía subyacente a este enfoque de Aprendizaje Automático es la que orquesta a la Teoría Vapnik-Chervonenkis (Teoría VC), desarrollada entre 1960 y 1990 (según (Wikipedia, 2020)), que busca entender el proceso de aprendizaje automático desde el punto de vista de la Estadística y del Análisis Funcional (el subcampo de las Matemáticas que estudia el comportamiento de objetos matemáticos en espacios de funciones), por lo que la Teoría VC está íntimamente relacionada con la Teoría del Aprendizaje Estadístico (es una rama de Teoría del Aprendizaje Estadístico, por decirlo de alguna manera) y con los procesos empíricos (en Teoría de la Probabilidad, son procesos estocásticos que describen la

proporción de objetos en un sistema, cuando este sistema se encuentra en un estado determinado).

Las MVS también han presentado buenos resultados al trabajar con conjuntos de datos de alta dimensionalidad y no es presa del problema de la maldición de la dimensionalidad anteriormente descrito. Uno de los aspectos que caracterizan a este enfoque de Aprendizaje Automático que en este los límites de decisión se expresan mediante subconjuntos de los conjuntos de entrenamiento creados, conocidos como *vectores de soporte*.

II.XL. Vector Prototipo

Como se señala en (Duval, What does prototype mean in clustering?, 2016), en el contexto de los árboles de decisión y el Aprendizaje Automático, un prototipo o vector prototipo es un elemento del espacio de datos que representa a un grupo de elementos. En el contexto del análisis de grupos, un grupo prototipo es aquel grupo que sirve para caracterizar al grupo en su totalidad, es decir, los elementos del grupo prototipo sirven para caracterizar a la totalidad del grupo de estudio. Este concepto teórico tiene sus raíces históricas en dos localizaciones. La primera de sus raíces se encuentra en una ponencia dentro de la vigésima sexta Conferencia Internacional de Aprendizaje Automático realizada en Montreal (Canadá) en el año 2009 a cargo de los investigadores Zhang, Kwok y Parvin en julio de 2009, como puede verificarse en (Zhang, Kwok, & Parvin, 2009) -fuente en la que se registra el 6 de julio de 2009- . La segunda de ellas se ubica en la investigación titulada "Classification by Set Cover: The Prototype Vector" publicada el 15 de julio de 2009 por Jacob Bien y Robert Tibshirani (co-autor de obras sobre Aprendizaje Estadístico, una de las cuales ha sido citada en esta investigación) y enviada a Arvix un 17 de agosto de 2009 para su publicación, como puede verificarse (Bien & Tibshirani, 2009).

II.XLI. Cuantización Vectorial

Según (Wikipedia, 2020), la cuantización vectorial es una técnica clásica de cuantización para el procesamiento de señales que permite modelar las funciones de densidad de probabilidad mediante la distribución de probabilidad de los vectores prototipo.

II.XLII. Análisis de Grupos por K-Medias

Como se señala en (Wikipedia, 2020), el *análisis de grupos por k-medias* es un método de cuantización vectorial, surgido originalmente en el estudio del procesamiento de señales, que busca particionar n observaciones en k grupos en los que cada observación pertenece al grupo con la media (que representa el centro o centroide -desde la perspectiva geométrica- del grupo) más cercana, de lo que se desprende que tales medias sirven de prototipo del grupo.

II.XLIII. Clasificador

Con lo planteado con antelación, resulta simple comprender la definición de *clasificador*. Según (Asiri, 2018), un clasificador es "(…) the process of predicting the class of given data points. Classes are sometimes called as targets/labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). For example, spam detection in email service providers can be identified as a classification problem. This is s binary classification since there are only 2 classes as spam and not spam. A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known spam and non-spam emails have to be used as the training data. When the classifier is trained accurately, it can be used to detect an unknown email (…) Classification belongs to the category of supervised learning where the targets also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc."

II.XLIV. Algoritmo de Maximización de Expectativas (Expectation-Maximization Algorithm)

II.XLIV.I. Generalidades

Como puede verificarse en (Dempster, Laird, & Rubin, 1977, pág. 1), el término "Datos incompletos" en su forma general implica la existencia de dos espacios muestrales e $\mathcal{Y}$ y $\mathcal{X}$, así como un mapeo de $\mathcal{X}$ a $\mathcal{Y}$ con la característica de poder asignar varios elementos de $\mathcal{X}$ a uno de $\mathcal{Y}$. Los datos observados y son realizaciones de $\mathcal{Y}$. El correspondiente x en $\mathcal{X}$ no es observado directamente, sino indirectamente a través de y. Más específicamente, se asume que existe una función $x \to y(x)$ que realiza su mapeo de $\mathcal{X}$ a $\mathcal{Y}$, así como también que la única información que el investigador posee sobre x es que se encuentra en $\mathcal{X}(y)$, el subconjunto de $\mathcal{X}$ determinado por la ecuación $y = y(x)$, en donde y es el conjunto de datos observado (del que dispone el investigador). En la investigación citada, los autores llaman al conjunto de datos $x \in \mathcal{X}$ *datos completos* (aunque en ciertos ejemplos x incluye lo que tradicionalmente se conoce como parámetros, según palabras de los propios autores en el lugar referido).

Así, los autores postulan una familia de densidades muestrales $f(x|\Phi)$ que depende de los parámetros Φ y derivan la familia de densidades muestrales que les corresponde, la cual adopta la forma $g(y|\Phi)$. La especificación de los datos completos $f(\dots|\dots)$está relacionada con la especificación de los datos incompletos $g(\dots|\dots)$ mediante el siguiente operador:

$$g(y|\Phi) = \int_{\mathcal{X}(y)} f(x|\Phi)dx$$

El algoritmo EM, planteado por Dempster, Laird y Rubin en 1977 (en la investigación referida), está diseñado para encontrar el valor de Φ que maximiza $g(y|\Phi)$ dada una y observada, pero hace tal cosa mediante un empleo esencial de la familia asociada $f(x|\Phi)$.

Una forma simple de expresar el papel de este algoritmo se puede localizar en (Maklin, 2019). Como se verifica en la fuente citada anteriormente, para estimar los parámetros de cada grupo gaussiano (de cada clúster cuyos elementos siguen una distribución Normal), es decir, para estimar la media, la varianza y las ponderaciones o pesos de cada distribución Normal que modela los datos contenidos en nada clúster estudiado es necesario primero saber qué muestra pertenece a qué grupo o clúster gaussiano para precisamente poder estimar así tales parámetros. Ese es precisamente el papel fundamental que desempeña el *algoritmo de maximización de expectativas* o *algoritmo EM*. Como se expone en (Dempster, Laird, & Rubin, 1977, pág. 1), cada iteración del algoritmo EM involucra dos pasos los cuales fueron llamados por los autores como "expectation step" (paso de expectativa -en referencia a la esperanza matemática-) y "maximization step" (paso de maximización, en referencia a la maximización de la verosimilitud de los valores deseados), llamados también por ellos de forma abreviada "E-Step" (Paso E) y "M-Step" (Paso M), respectivamente.

El proceso de iteración del algoritmo es explicado por sus creadores en (Dempster, Laird, & Rubin, 1977, pág. 4). Suponga el lector que $f(x|\Phi)$ posee una familia exponencial regular[31] de la forma:

$$f(x|\Phi) = \frac{b(x)\exp(\Phi \mathrm{t(x)}^T)}{a(\Phi)}$$

[31] Más adelante en su investigación señalarán los autores que "The term regular means here that Φ is restricted only to an r − dimensional convex set Ω such that (2.1) defines a density for all Φ in Ω" (Dempster, Laird, & Rubin, 1977, pág. 4).

En donde Φ denota al vector de parámetros $1 \times r$, $t(x)$ denota al vector de estadísticos suficientes de los datos completos ("(…) denotes a $1 \times r$ vector of *completa-data* sufficient statistics") y T como potencia denota la matriz transpuesta del conjunto de datos de los que dispone el investigador. La ecuación anterior es a la que en la fuente original que se ha citado se etiqueta como la ecuación (2.1).

Así, los autores presentan una caracterización simple del algoritmo EM que exponen puede ser usualmente utilizada cuando la ecuación (2.1) de la fuente citada se sostiene. Suponga entonces el lector que $\Phi^{(p)}$ denota el valor actual de Φ después de p ciclos realizados por el algoritmo EM. El siguiente ciclo a realizar por el algoritmo en cuestión puede ser descrito mediante dos pasos:

II.XLIV. II. Paso E

Estima los estadísticos suficientes de la data-completa (datos completos, "complete-data") $t(x)$ a través de encontrar:

$$t^{(p)} = E\left(t(x)\middle|y, \Phi^{(p)}\right)$$

II.XLIV. III. Paso M

Determina $\Phi^{(p+1)}$ como la solución de las ecuaciones:

$$E(t(x)|\ \Phi) = t^{(p)}$$

La ecuación anterior es la forma familiar (la forma matemática bajo la que usualmente aparecen) de las ecuaciones de máxima verosimilitud para la estimación de máxima verosimilitud dado el conjunto de datos proveniente de una familia exponencial regular (en el sentido ya definido). Esto es, si se asume que $t^{(p)}$ representa los estadísticos suficientes estimados de una x observada proveniente de (2.1), es decir, estimada mediante (2.1), entonces las ecuaciones de máxima verosimilitud de la forma aquí presentadas usualmente definen el estimador de máxima verosimilitud de Φ.

Como puede observarse, la mecánica general a la que obedece este algoritmo es, para el caso de grupos (clúster) cuyos elementos se distribuyen normalmente, la siguiente:

1. Se parte de parámetros estocásticos que siguen una distribución de probabilidad Normal. En este paso se utilizan los datos disponibles para estimar los valores de las variables faltantes en el modelo (recordar que se hace en el contexto de un conjunto de datos incompletos).
2. Se repite ese proceso hasta alcanzar la convergencia, es decir, hasta alcanzar el valor óptimo de los parámetros, en donde tal óptimo es determinado con base en la configuración matemática previamente expuesta. Aquí, en función de los valores generados en el paso anterior, los datos completos se utilizan para actualizar los valores de los parámetros, realizando este proceso iterativamente hasta alcanzar el óptimo descrito.

II.XLV. Distribución de Probabilidad Multinomial (Distribución de Probabilidad de Bernoulli Generalizada)

Como se señala en (Murphy, 2012, pág. 35), la distribución binomial puede ser usada para modelar los resultados de lanzamientos de una moneda. Para modelar los resultados de lanzar un dado de $K - lados$ es posible utilizar la conocida como *distribución de probabilidad multinomial* o *distribución de Bernoulli generalizada*, también conocida en algunos contextos como *distribución de probabilidad categórica* (puesto que es un caso particular de la primera, como se verá más adelante). Esta distribución discreta se define como se presenta a continuación. Sea $x = (x_1, x_2, \dots, x_K)$ un vector estocástico, en donde x_j es el número de el lado j del dado aparece tras el lanzamiento. Entonces x posee la siguiente distribución de masa de probabilidad:

$$\mathrm{Mu}(x|n, \boldsymbol{\theta}) \triangleq \frac{n!}{x_1!\, x_2! \dots x_K!}$$

En donde θ_j es la probabilidad de que aparezca el lado j, mientras que

$$\binom{n}{x_1, x_2, \ldots, x_K} \triangleq \frac{n!}{x_1!\, x_2! \ldots x_K!}$$

Es el *coeficiente multinomial* (que expresa el número de formas de dividir el conjunto de tamaño $n = \sum_{k=1}^{K} x_K$ en subconjuntos con tamaños que van desde x_1 hasta x_K).

Finalmente, es necesario explicar el significado del signo ≜. Este es, como se señala en (Duval, Who first defined the "equal-delta" or "delta over equal" (≜) symbol?, 2016), "The symbol ≜ is sometimes used in mathematics (and physics) for a definition. It is instantiated for instance in the Unicode Character 'DELTA EQUAL TO' (U+225C). The notation $t \triangleq m$ (often) means: "t is defined to be m" or "t is equal by definition to m" (often under certain conditions). In a similar sense, some uses $:=$ or $=:$ (…)"

Generalizando su interpretación, esta distribución discreta describe los posibles valores que puede tomar una variable estocástica de las $K - ésimas$ categorías posibles, con la probabilidad de cada categoría especificada por separado. Como se señala en (Wikipedia, 2020), no existe un orden subyacente innato de estos resultados, pero a menudo se adjuntan etiquetas numéricas para facilitar la descripción de la distribución (por ejemplo, 1 a K). La distribución categórica $K - dimensional$ es la distribución más general sobre un evento que puede ocurrir de K maneras. Cualquier otra distribución discreta sobre un espacio muestral de tamaño K es un caso especial. Los parámetros que especifican las probabilidades de cada resultado posible están limitados solo por el hecho de que cada uno debe estar en el rango de 0 a 1, y todos deben sumar 1, puesto que las probabilidades oscilan entre cero y uno.

La distribución categórica es la generalización de la distribución de Bernoulli para una variable aleatoria categórica, es decir, para una variable discreta con más de dos resultados posibles, como la tirada de un dado, generalizándose esta idea para un dado de K lados. Por otro lado, la distribución categórica es un caso especial de la distribución multinomial, ya que con ella se obtienen las probabilidades de

resultados potenciales de una sola extracción en lugar de varias extracciones[32]. Así, se hace alusión a que existe una cantidad (dígase m) de muestras de tamaño arbitrario y que se extrae únicamente una de tales muestras.

Familias Exponenciales y Distribución Conjugada A Priori

Con base en lo planteado por (Hoff, 2009, pág. 51), una familia exponencial de un parámetro es todo modelo cuyas densidades de probabilidad pueden ser expresadas como $p(y|\phi) = h(y)c(\phi)e^{\phi t(y)}$, en donde ϕ es el parámetro desconocido (llamado también *parámetro de la familia*) y $t(y)$ es el estadístico suficiente en función de las observaciones. En la expresión anterior, tanto h como c son funciones de distribución de probabilidad conocidas, en concreto, $h(y)$ es la distribución de probabilidad que sigue el conjunto de observaciones y $c(\phi)$ es la distribución de probabilidad que sigue el parámetro ϕ. Finalmente, puede observarse que $e^{\phi t(y)}$ es una combinación exponencial del parámetro ϕ y del estadístico suficiente $t(y)$ en función del conjunto de datos. La notación utilizada en la fuente citada, aunque menos familiar, no es en lo absoluto metafísica, ya que ϕ es lo que usualmente se denota con θ para hacer referencia al parámetro o conjunto de parámetros, como el lector perfectamente sabe.

Por otro lado, según lo planteado en (Congdon, 2006, pág. 152), el estado general de las distribuciones conjugadas a priori en el contexto de los modelos jerárquicos puede describirse como se hace a continuación. La sucesión de puntos de datos (observaciones) y sus subyacentes valores verdaderos $(y_i|\theta_i), i = 1, 2, \dots, n$ son idénticamente distribuidos. La distribución de la densidad de las observaciones y_i, dado θ_i, es $P(y_i|\theta_i)$. En una segunda etapa, mediante el estudio de la función de densidad de probabilidad que gobierna al parámetro θ_i puede especificarse una

[32] En esta parte se lee en la última fuente citada de forma textual: "The categorical distribution is the generalization of the Bernoulli distribution for a categorical random variable, i.e. for a discrete variable with more than two possible outcomes, such as the roll of a die. On the other hand, the categorical distribution is a special case of the multinomial distribution, in that it gives the probabilities of potential outcomes of a single drawing rather than multiple drawings."

media común (bajo intercambiabilidad) o puede involucrar medias diferentes definidas por una regresión sobre los predictores X_i. Esta segunda etapa de la mixtura de densidad está gobernada por los hiperparámetros $\Lambda = \lambda_1, \lambda_2, \dots, \lambda_L$, cuya densidad se especifica en la tercera etapa. Más formalmente, un modelo jerárquico de 3 etapas posee los siguientes componentes:

1) Condicionada por $\{\theta_1, \theta_2, \dots, \theta_n\}$, los elementos del conjunto de observaciones y_i son independientes, con densidades $P(y_i|\theta_i)$, las cuales son independientes de θ_j, para $j \neq i$, y de Λ.
2) Condicionados por Λ, los valores verdaderos θ_i son extraídos de la misma función de densidad de probabilidad $g(\theta|\Lambda)$.
3) Los hiperparámetros Λ poseen su propia función de densidad $h(\Lambda)$.

II.XLVI. Distribución de Dirichlet

Como se señala en (Wikipedia, 2020), la distribución de Dirichlet es la distribución conjugada a priori de la distribución categórica (una distribución de probabilidad discreta genérica con un número dado de resultados posibles) y la distribución multinomial (la distribución sobre los conteos observados de cada categoría posible en un conjunto de observaciones distribuidas categóricamente). Esto significa que, si un punto de datos tiene una distribución categórica o multinomial, y la distribución a priori del parámetro de distribución (el vector de probabilidades que genera el punto de datos) se distribuye según una función de densidad de probabilidad Dirichlet, entonces la distribución a posteriori (o distribución posterior) del parámetro de estudio también es se distribuye según una función de densidad de probabilidad de Dirichlet. De manera intuitiva, en tal caso, a partir del conocimiento que posee el investigador sobre el parámetro antes de observar el punto de datos, es posible actualizar su conocimiento con base base al punto de datos y terminar con una nueva distribución de la misma forma que la anterior. Esto significa que es posible que el investigador actualice sucesivamente su

conocimiento sobre un parámetro incorporando nuevas observaciones, una a la vez, sin encontrarse con dificultades matemáticas.

Según lo documentado por (Wang Ng, Tian, & Tang, 2011, pág. 37), "Whenever a multivariate observation is a set of proportions, called *compositional data* by Aitchison (1986), the Dirichlet family of distributions is usually the first candidate employed for modeling the data. n Bayesian inference for multinomial data, the Dirichlet distribution is the conjugate prior distribution so that the posterior distribution is also a Dirichlet distribution. Applications of the distribution are so many and so diverse that here we can name only a few. For example, Wilks (1962) used it in theoretical analysis to derive the distribution function of a set of order statistics, Theil (1975) used it to model random rational behavior in consumption expenditures, and Spiegelhalter et al. (1994) used it to study the frequencies of congenital heart disease. In biology, the Dirichlet distribution is used to represent proportions of amino acids when modeling sequences with hidden Markov models (Sjölanderet al., 1996) or with allelic frequencies (Lavalet al., 2003). In text-mining, it is adopted to model topic probabilities (Bleiet al., 2006)."[33]

II.XLVII. Optimización y Conceptos Relacionados

La optimización consiste en el proceso de encontrar valores de máximos o mínimos (sean absolutos, relativos o condicionados) que representan, según sea el caso de aplicación, la selección del mejor elemento (o decisión), con respecto a algún criterio de referencia, de un conjunto de elementos disponibles.

Por otro lado, la restricción de una función $f(x)$ es otra función $g(x)$ definida en un subconjunto del dominio de $f(x)$. La función $f(x)$ es a su vez una extensión de $g(x)$. La restricción de una función se obtiene al reducir su dominio. Así, los máximos o mínimos encontrados en algún problema de optimización pueden ser algunas de las siguientes opciones:

[33] Cursivas añadidas por el autor de esta investigación.

1) Máximo absoluto, que es el valor máximo que toma una función en la totalidad de su dominio.
2) Mínimo absoluto, que es el valor mínimo que toma una función en la totalidad de su dominio.
3) Máximo relativo o local, que es el valor máximo que toma una función dentro de una región particular del dominio de la función.
4) Mínimo relativo o local, que es el valor mínimo que toma una función dentro de una región particular del dominio de la función.
5) Máximo condicionado, que es el valor máximo de una función no sobre cualquier punto de su dominio sino sobre un subconjunto del dominio de la función, es decir, consiste en encontrar un valor máximo sujeto a la condición de que el punto donde se produce pertenezca a un cierto conjunto. Tal condición viene dada por la función que funge como restricción.
6) Mínimo condicionado, que es el valor mínimo de una función no sobre cualquier punto de su dominio sino sobre un subconjunto del dominio de la función, es decir, consiste en encontrar un valor mínimo sujeto a la condición de que el punto donde se produce pertenezca a un cierto conjunto. Tal condición viene dada por la función que funge como restricción.

Por su parte, el gradiente es una operación vectorial, que opera sobre una función escalar, para producir un vector cuya magnitud es la máxima razón de cambio de la función en el punto donde se realiza el cálculo del gradiente y que apunta en la dirección de ese gradiente. También puede ser definido como la colección de todas las derivadas parciales de una función en forma de vector. Para una función de dos variables, su gradiente será $\nabla f(x, y) = \begin{pmatrix} f_x \\ f_y \end{pmatrix}$.

En general, un operador puede ser definido como un símbolo que indica que debe ser llevada a cabo una operación especificada sobre un cierto número de operandos. Este símbolo y la operación que lleva aparejada, a grandes rasgos, permite transformar una función en otra función. La derivada es un ejemplo de operador,

pues transforma una función f a una función f'. Los diferentes tipos de derivada representan diferentes tipos de operadores diferenciales, como se muestra en la figura presentada a continuación.

Figura 13

Nombre	Símbolo	Ejemplo
Derivada	$\frac{d}{dx}$	$\frac{d}{dx}(x^2) = 2x$
Derivada parcial	$\frac{\partial}{\partial x}$	$\frac{\partial}{\partial x}(x^2 - xy) = 2x - y$
Gradiente	∇	$\nabla(x^2 - xy) = \begin{bmatrix} 2x - y \\ -x \end{bmatrix}$

Fuente:

Por otro lado, el determinante es un operador que le asigna a una matriz un número real único, mientras que un punto crítico es cualquier valor en que la función no es diferenciable[34] o cuando su derivada es cero (punto estacionario en términos de una variable[35]). El valor de la función en el punto crítico es un valor crítico de la función.

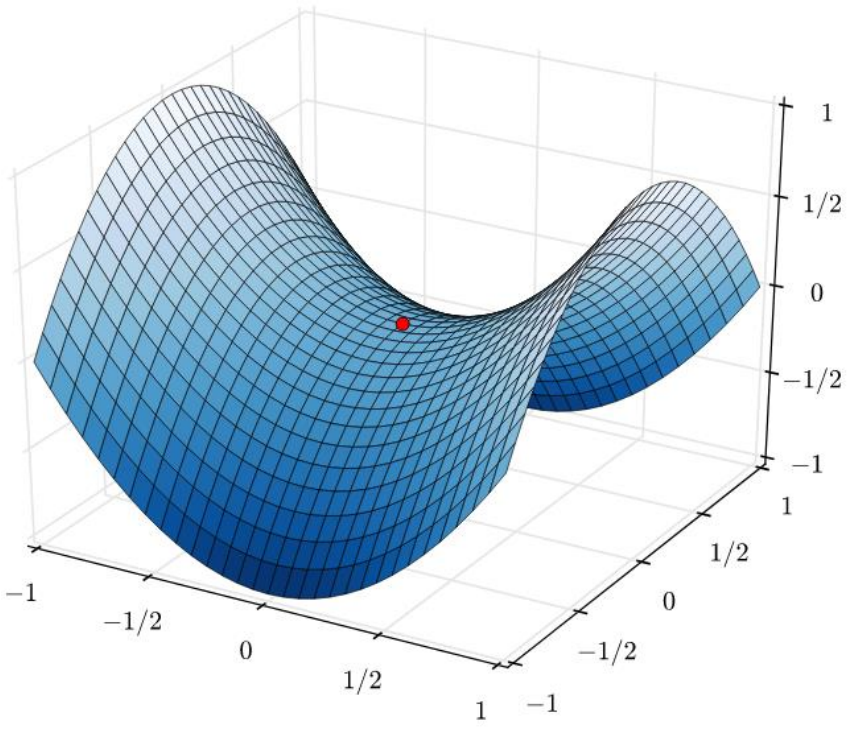
Así, un punto estacionario de una función de varias variables reales, es un punto en donde se anulan simultáneamente todas sus derivadas parciales. Si la función f es diferenciable, los puntos donde tiene un extremo están entre los puntos estacionarios de la función, mientras que un punto de silla es el punto sobre una superficie en el que la pendiente es cero, pero no se trata de un extremo relativo. Es el punto sobre una superficie en el que la elevación es máxima en una dirección y mínima en la dirección perpendicular.

---

[34] Al hablar de diferenciabilidad en $\mathbb{R}$ solamente se requería que la función $f(x)$ fuera derivable en el punto (es decir, que su derivada evaluada en el punto existiera), por lo cual, una función derivable y una función diferenciable eran sinónimos; sin embargo, al pasar de $\mathbb{R}$ a $\mathbb{R}^n$ ya no solo se requiere que la función sea derivable parcialmente en un entorno de cualquier punto y que sea continua en el punto.

[35] Si la función f es derivable y tiene un extremo local en un punto, ese punto estará entre sus puntos estacionarios.

Figura 14



Fuente: (Wikipedia, 2017).

II.XLVIII. Gradiente Descendiente

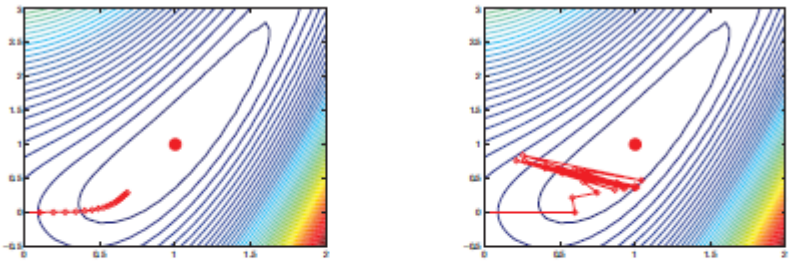
Como se señala en (Murphy, 2012, pág. 247), quizás el algoritmo más simple para realizar optimización sin restricciones es el conocido como *gradiente descendiente*, conocido también como *descenso más empinado*. Este algoritmo puede ser escrito como se hace a continuación:

$$\theta_{k+1} = \theta_k + \eta_k g_k$$

En donde η_k es la *tasa de aprendizaje* o el tamaño del paso. La cuestión de interés principal en el algoritmo del gradiente descendiente es cómo configurar la tasa de aprendizaje, que por motivos de la delimitación de la investigación no realizará un análisis de tal configuración en esta investigación, pues por sí misma esa temática casi seguramente algún día conformará una investigación independiente. Sin embargo, si el lector consulta (Murphy, 2012, págs. 247-249) puede hacerse una

idea general de los aspectos relacionados con la configuración de la tasa de aprendizaje de un modelo estadístico.

Figura 15



(a) (b)

Figure 8.2 Gradient descent on a simple function, starting from $(0, 0)$, for 20 steps, using a fixed learning rate (step size) η. The global minimum is at $(1, 1)$. (a) $\eta = 0.1$. (b) $\eta = 0.6$. Figure generated by `steepestDescentDemo`.

Fuente: (Murphy, 2012, pág. 247).

Núcleo (Kernel)

Como se señala en (Nabi, Algunas Reflexiones Sobre la Distribución Binomial Negativa II (Un Análisis Teórico y Aplicado), 2020), el concepto teórico de *núcleo* suele encontrarse en la literatura en inglés como *kernel*. Desde el punto de vista de la lengua inglesa, la palabra kernel es una modernización de la palabra *cyrnel* perteneciente al inglés antiguo, la cual significa "semilla" y, además, a nivel de la lexicología en habla inglesa empleada en Botánica, la palabra "core" se comprende como un contenedor de semillas); sin embargo, esta palabra tiene su origen en el idioma alemán y ahí su traducción es directamente *núcleo*. Según (Wikipedia, 2020) tiene distintos significados según los campos de aplicación y según la Wikipedia en ruso (Википедия, 2020) es un concepto aplicado específicamente en Estadística Matemática, en Econometría, en Estadística Bayesiana, Estadística No-Paramétrica y la Teoría del Reconocimiento de Patrones verificando también (al igual que en versión en inglés) que su significado varía según su campo de aplicación. La Wikipedia en francés (Wikipedia, 2020) contribuye a terminar de clarificar la

imagen cuando explica que los núcleos son parte de los denominados estimadores kernel (que sirven para estimar la función de densidad de probabilidad de una variable aleatoria, en la regresión paramétrica para estimar las esperanzas condicionales y en series de tiempo para estimar la densidad espectral. Por lo tanto, resulta evidente que este caso de estudio es a nivel de regresiones paramétricas en las que se estiman esperanzas condicionales (en otra investigación se extenderán estos resultados a la regresión binomial negativa 2) y en Estadística Bayesiana. En este contexto se habla del núcleo de una masa o de una densidad y es la forma que toma la masa o la densidad en la cual todos los factores que no son funciones de ninguna variable del dominio son omitidos de la masa o de la densidad, según corresponda. Este concepto busca capturar la intuición de "la esencia de la función", su *núcleo*.

II.XLVIX. Parámetros e Hiperparámetros

Como se señala en (Paul, 2018), en el contexto del Aprendizaje Automático, un parámetro de un modelo es una variable de configuración que es interna al modelo y cuyo valor se puede estimar a partir de los datos. Sin embargo, los parámetros son mucho más que eso y antes de explicar qué son los hiperparámetros en el contexto del Aprendizaje Automático, es de importancia fundamental explicar en su sentido más general qué son los parámetros, es decir, en cualquier contexto del que se trate. Para ello, estudiaremos los parámetros en el contexto de las ecuaciones diferenciales y las ecuaciones en diferencias.

Como ya el lector sabrá, una ecuación en diferencias se distingue de una ecuación diferencial solo por el hecho en que se suscita en tiempo discreto, mientras que la segunda lo hace en tiempo continuo; además, una ecuación en diferencias puede concebirse como un acercamiento por métodos numéricos a una solución aproximada de una ecuación diferencial, lo cual se realiza mediante la ruptura de la continuidad del dominio de la función en intervalos discretos; por ello, bastará

con aproximarse epistemológicamente a las ecuaciones diferenciales para comprender por inferencia de la misma manera las ecuaciones en diferencias.

Así, solo resta establecer brevemente las diferencias entre epistemología y epistémica, la cual según (Ramírez, 2017) consiste en que la epistémica se ocupa de los contextos históricos, filosóficos y culturales en los cuales se desarrolla un determinado estilo de pensamiento, mientras que la epistemología versa sobre la galería de posiciones y disposiciones para construir el objeto de estudio, es decir, conlleva a preguntarse qué es conocer, quién conoce y qué es lo conocido.

En los textos de Matemática Pura, se define una ecuación diferencial como "una ecuación que liga la variable independiente x, la función incógnita $y = y(x)$ y sus derivadas $y', y'', \dots, y^{(n)}$, es decir, una ecuación de la forma $F(y', y'', \dots, y^{(n)}) = 0$. En otras palabras, se llama *ecuación diferencial* una ecuación en la que figura la derivada o diferencial de la función incógnita." (Kiseliov, Krasnov, & Makarenki, 1973, pág. 9), sin embargo, la definición anterior no posee mayor valor epistemológico en términos científicos, es decir, concibiendo a la Matemática Pura como una ciencia instrumental, esta definición por sí misma no esclarece como las ecuaciones diferenciales permiten a las ciencias entrar en contacto con las fuentes, las formas y los métodos de conocimiento de la realidad o, dicho de otra manera, no esclarece como las ecuaciones diferenciales representan un camino al conocimiento.

Así, deberá empezar por plantearse una ecuación diferencial como aquella igualdad matemática que liga la variable independiente x conocida, una variable dependiente del valor de x desconocida de la forma y = y(x), es decir, una función incógnita de x, así como las tasas de cambio instantáneas de la variable dependiente respecto a la variable independiente.

Las ecuaciones diferenciales se presentan en aquellas circunstancias en que al analizar un fenómeno natural o social se desconoce la forma funcional que adopta

la variable cuyo comportamiento se desea explicar y se conoce únicamente la variable explicativa y las tasas de cambio instantáneas de la función incógnita respecto a la variable explicativa en cuestión o bien, puede darse el caso que aparezca la variable que se desea explicar, pero no de tal forma en que sea posible analizarla directamente, puesto que no se encuentra aislada de sus tasas de cambio y de la(s) forma(s) en que aparece(n) su(s) variable(s) explicativa(s). En otras palabras, una ecuación diferencial sirve para encontrar la forma funcional de variable que se desea explicar mediante la variable explicativa y sus tasas de cambio instantáneas respecto a la última.

Lo anterior implica que, al analizar un fenómeno de cualquier índole, es posible modelar el comportamiento de la variable que explica el fenómeno, las tasas instantáneas en que el fenómeno en cuestión cambia respecto su variable explicativa, pero se desconoce la forma funcional que este fenómeno adopta (que permite su análisis directo) y con la cual puede ser modelado su comportamiento. Por tanto, resulta trivial conocer las circunstancias generales que suscitan que el investigador se encuentre con una ecuación diferencial, ya que esto ocurrirá ante ausencia de información sobre el fenómeno estudiado; por el contrario, lo fundamental radica en que a través de las mismas el investigador puede descubrir la forma funcional que modela el comportamiento de la variable que desea explicar, del fenómeno de estudio. Así, una ecuación diferencial es aquella ecuación de la cual se deduce una función desconocida que permita su análisis directo a través sus tasas de cambio instantáneas de orden n-ésimo y de los diferentes comportamientos observables de la variable explicativa.

Cabe mencionar que el investigador no se encontrará en la naturaleza ni en la sociedad con una ecuación diferencial, sino únicamente con los elementos que permiten construirla y deducir de ella la función que modela el comportamiento del fenómeno específico de estudio. En otras palabras, el investigador, puesto en un escenario en que no conoce cómo se comporta un fenómeno y, por tanto, no

pudiendo tampoco modelar dicho comportamiento, pero conociendo las tasas de cambio instantáneas del fenómeno respecto a la variable que lo explica y pudiendo modelar el comportamiento de dicha variable explicativa (e inclusive, una función de la variable a explicar que no permita un análisis directo), construye una ecuación que le permita deducir el modelo de comportamiento del fenómeno de estudio.

II.XLVIX.I. Solución General de una Ecuación Diferencial

En términos de lo anteriormente expuesto, la solución general de una ecuación diferencial es la forma funcional del fenómeno analizado que satisface la construcción teórica y matemática realizada del mismo a través de los distintos comportamientos observables de la variable explicativa y sus tasas de cambio instantáneas (la ecuación diferencial) para cualesquiera parámetros arbitrarios, es decir, para una $n-ésima$ cantidad los mismos.

II.XLVIX. II. Solución Particular de una Ecuación Diferencial

Por su parte, la solución particular de una ecuación diferencial se obtiene de la solución general, específicamente cuando en ella se asignan valores específicos a los $n-ésimos$ parámetros obtenidos al encontrar la función que modela el comportamiento del fenómeno analizado.

II.XLVIX.III. Solución Singular de una Ecuación Diferencial

Básicamente, una solución de una ecuación diferencial es singular cuando no es posible obtenerla mediante la sustitución de valores específicos para las variables y parámetros en la solución general, es decir, no es una solución particular, pero satisface la ecuación diferencial[36].

[36] Tal solución puede obtenerse a través de diversos mecanismos, por ejemplo, planteando un sistema de ecuaciones con la ecuación diferencial y la solución general para luego eliminar los parámetros, tras lo cual se obtiene una solución que no puede obtenerse de la solución general (es decir, no es una solución particular), pero satisface la ecuación diferencial.

En la matemática de conocimiento popular, una función es aquella en que el valor de una variable (dependiente) se encuentra en función o depende del valor que tome otra variable (independiente); sin embargo, tal definición daría lugar a concebir la variable dependiente como función de los parámetros, lo cual formalmente no es así.

En Análisis Matemático, una función hace referencia a una ley de composición[37] que asigna a cada elemento de un conjunto A un único elemento de un conjunto B. Así, para el caso de $y = f(x)$, a cada elemento de x le corresponde un único elemento de y. Ahora bien, supóngase el caso univariante en el que se tiene la función $y = kf(x)$, dado un valor específico de k, a cada elemento de x le corresponde un único elemento de y; a su vez, dado x a cada valor específico de k le corresponde un único valor de y. Dicho de otra forma, el parámetro k podría verse como la variable independiente de una función de $\mathbb{R}$ en el conjunto de funciones de cualquier tipo, la cual asocia a cada valor de k una función, es decir: $\big(y = kf(x)\big) = g(k)$, donde $k \in \mathbb{R}$. Lo anterior quiere decir que k también puede concebirse como una variable independiente auxiliar.

Por tanto, una solución singular de una ecuación diferencial en términos epistemológicos no es otra cosa que asumir que el fenómeno a explicar depende única y exclusivamente de sus variables explicativas fundamentales, sus tasas de cambio instantáneas y (si fuese el caso) de alguna forma funcional de sí mismo que

[37] Una ley de composición es aquella operación binaria que da lugar a distintas estructuras algebraicas, la cual es de carácter interno (ley de composición interna) si la operación binaria asigna a todo par ordenado cuyos elementos pertenecen a un conjunto determinado, un tercer elemento que pertenece también a dicho conjunto, por ejemplo, las suma entre dos números naturales (será siempre un número natural) o la multiplicación entre dos números racionales (será siempre un número racional), así como la unión y la intersección de dos conjuntos, es decir, la formación de un nuevo conjunto que incluya todos los elementos de los conjuntos unidos (sin repeticiones) y la formación de un conjunto que incluya solo los elementos que los conjuntos intersecados tienen en común, respectivamente, mientras que es de carácter externo (ley de composición externa) si los dos operandos no pertenecen al mismo conjunto. Se ampliará al respecto en los anexos.

no permita que sea analizado directamente, es decir, suponer un escenario en ausencia de variables explicativas auxiliares.

Ahora bien, una vez establecido lo anterior, se deduce que la unicidad de una solución a determinada ecuación diferencial no es otra cosa que la ausencia de soluciones singulares, es decir, que toda solución que satisfaga la ecuación diferencial se encuentre contenida dentro de la solución general. Por tanto, la unicidad de una solución consiste en que el fenómeno analizado no pueda ser explicado sin tomar en cuenta tanto sus variables explicativas fundamentales, como sus variables explicativas auxiliares, lo que implica a su vez la importancia fundamental de estas últimas variables, que es precisamente a los que se conocen como parámetros.

Como señala (Paul, 2018), los parámetros son requeridos para:

1) Hacer las predicciones del modelo.
2) Sus valores definen la habilidad del modelo para resolver el problema para el cual ha sido planteado.
3) Se estiman o se extraen de los datos.
4) A menudo, su configuración no es realizada por el investigador de forma manual.
5) A menudo, se guardan como parte del modelo aprendido.

“So, your main take away from the above points should be parameters are crucial to machine learning algorithms. Also, they are the part of the model that is learned from historical training data. Let’s dig it a bit deeper. Think of the function parameters that you use while programming in general. You may pass a parameter to a function. In this case, a parameter is a function argument that could have one of a range of values. In machine learning, the specific model you are using is the function and requires parameters in order to make a prediction on new data.

Whether a model has a fixed or variable number of parameters determines whether it may be referred to as "parametric" or "nonparametric"." (Paul, 2018)[38].

II.XLVIX. IV. Parámetros e Hiperparámetros como Variables Auxiliares Generadas del Conjunto de Datos

Así, en el contexto del Aprendizaje Automático y con base en lo expuesto en el contexto de las ecuaciones diferenciales como en lo expuesto por (Paul, 2018), es posible entender a los hiperparámetros como variables auxiliares cuya configuración es externa al modelo y cuyos valores no pueden ser estimados del conjunto de datos. Las demás características fundamentales de los hiperparámetros son:

1) A menudo se utilizan en procesos para ayudar a estimar los parámetros del modelo.
2) A menudo los especifica el investigador.
3) A menudo se pueden establecer mediante heurística[39].
4) A menudo se ajustan a un problema de modelado predictivo determinado. Esto significa que no es posible generalizar a gran escala la estructura matemática de los procesos heurísticos (que emplean funciones heurísticas) y que, por consiguiente, es casi seguro que no existirán demostraciones matemáticas en el marco de la Teoría Axiomática de Conjuntos ZFC.

No es posible conocer el mejor valor para un hiperparámetro de un modelo en un determinado problema de estudio. Se pueden utilizar reglas generales, copiar

[38] Cursivas añadidas por el autor de la presente investigación.
[39] Como se señala en (Wikipedia, 2020), una heurística (del griego εὑρισκω "encuentro, descubro") es una técnica diseñada para resolver un problema más rápidamente cuando los métodos clásicos son demasiado lentos, o para encontrar una solución aproximada cuando los métodos clásicos no permiten encontrar ninguna solución exacta. Esto se logra intercambiando la optimización, la integridad, la exactitud (accuracy) o la precisión (precision) por velocidad. En cierto modo, puede considerarse un atajo a los métodos clásicos. Así, una función heurística, también llamada simplemente heurística, es una función que clasifica las alternativas en los algoritmos de búsqueda en cada paso de bifurcación en función de la información disponible para decidir qué bifurcación seguir. Por ejemplo, puede aproximarse a la solución exacta.

valores utilizados en otros estudios o buscar el mejor valor mediante prueba y error. Cuando un algoritmo de aprendizaje automático se ajusta a un problema específico, esencialmente está ajustando los hiperparámetros del modelo para descubrir los parámetros del modelo que dan como resultado las predicciones de mejor calidad. Como señalan en la fuente citada "According to a very popular book called "Applied Predictive Modelling" - "Many models have important parameters which cannot be directly estimated from the data. For example, in the K-nearest neighbor classification model … This type of model parameter is referred to as a tuning parameter because there is no analytical formula available to calculate an appropriate value."

Ejemplos de la necesidad del empleo de hiperparámetros son:

1) La tasa de aprendizaje para entrenar una red neuronal.
2) Los hiperparámetros C y sigma para máquinas de vectores de soporte.
3) La k en los vecindarios $k - más\ cercanos$ (en inglés "$k - nearest$ neighbors").

Como se señala en (Paul, 2018), la mejor manera de pensar en los hiperparámetros es en términos de la configuración de un algoritmo que se puede ajustar para optimizar el rendimiento, del mismo modo que puede girar las perillas de una radio AM para obtener una señal clara. Al crear un modelo de aprendizaje automático, se presentarán opciones de diseño sobre cómo definir la arquitectura del modelo, *i.e.*, su estructura fundamental, su estructura interna. A menudo, no sabe de inmediato cuál debería ser la arquitectura óptima para un modelo determinado y, por lo tanto, el investigador puede requerir de explorar una variedad de posibilidades. "In a true machine learning fashion, you'll ideally ask the machine to perform this exploration and select the optimal model architecture automatically."

Como señalan (minerals; enterML; Lakshmi Prasad Y; Dynamic Stardust; Manju Savanth; Prhld, 2018) en una sección de discusión titulada "What is the difference between model hyperparameters and model parameters?":

"You call something a 'hyperparameter' if it cannot be learned within the estimator directly. However, 'parameters' is more general term. When you say 'passing the parameters to the model', it generally means a combination of hyperparameters along with some other parameters that are not directly related to your estimator but are required for your model (…) Model parameters are the properties of the training data that are learnt during training by the classifier or other ML (Machine Learning) model. For example, in case of some NLP (Natural Language Processing) task: word frequency, sentence length, noun or verb distribution per sentence, the number of specific character n-grams per word, lexical diversity, etc. Model parameters differ for each experiment and depend on the type of data and task at hand (…) Hyper-parameters are those which we supply to the model, for example: number of hidden Nodes and Layers, input features, Learning Rate, Activation Function etc., in Neural Network, while Parameters are those which would be learned by the machine, like Weights and Biases."

Matemáticamente hablando, un modelo de aprendizaje automático M con parámetros e hiperparámetros toma la siguiente forma:

$$Y \approx M_{\mathcal{H}}(\Phi|D)$$

En donde Φ representa los parámetros y $\mathcal{H}$ son los hiperparámetros. Por su parte, D es el conjunto de datos de entrenamiento y Y es el conjunto de datos de salida ("output data") o etiquetas de clase ("class labels") en el caso de una tarea de clasificación. El objetivo durante el entrenamiento del modelo es encontrar los parámetros $\widehat{\Phi}$ que optimizan la función de pérdida $\mathcal{L}$ que el investigador ha especificado. Puesto que tanto el modelo M y la función de pérdida $\mathcal{L}$ se basan en $\mathcal{H}$, entonces los consecuentes parámetros Φ también dependen de los hiperparámetros $\mathcal{H}$.

Los hiperparámetros $\mathcal{H}$ no aprenden durante el entrenamiento, sin embargo, de ello no debe entenderse que sus valores permanecen inmutables. Típicamente, los hiperparámetros son fijos y se piensa en ellos como el modelo M, en lugar del modelo $M_{\mathcal{H}}$. En este contexto, los hiperparámetros pueden ser también considerados como parámetros a priori.

La raíz de la confusión usual entre parámetros e hiperparámetros radica, según se señala en la discusión citada, en el uso de $M_{\mathcal{H}}$ y la modificación de los hiperparámetros $\mathcal{H}$ durante la rutina de entrenamiento, en adición obviamente con los parámetros Φ. Potencialmente existen múltiples motivaciones para modificar $\mathcal{H}$ durante el entrenamiento. Un ejemplo de ello podría ser el cambiar la tasa de aprendizaje durante el entrenamiento para mejorar la velocidad y/o la estabilidad de la rutina de optimización.

La distinción importante a realizar es que, mientras que la predicción de la etiqueta Y_{pred} está basada en los parámetros Φ del modelo, mientras que los hiperparámetros $\mathcal{H}$ no. Sin embargo, esta distinción tiene salvedades y, en consecuencia, las líneas se difuminan. Considérese, por ejemplo, la tarea de agrupamiento, específicamente el modelado con mixtura gaussiana (GMM, por su nombre en inglés). Los parámetros establecidos aquí son $\Phi = \{\bar{\mu}, \bar{\sigma}\}$, en donde $\bar{\mu}$ es el conjunto de N grupos (clúster) y $\bar{\sigma}$ es el conjunto de N desviaciones estándar, correspondientes a N kernel gaussianos, es decir, N kernel de funciones de densidad de probabilidad Normal. Precisamente el hiperparámetro en este ejemplo es el número de clúster N, por lo que $\mathcal{H} = \{N\}$.

Típicamente, la validación de grupos (clúster) es utilizada para determinar N *a priori*, usando una pequeña submuestra del conjunto de datos D. Sin embargo,

podría modificarse el algoritmo de aprendizaje de los modelos de mixtura gaussiana para modificar el número de kernel N durante el entrenamiento, con base en algún criterio. En tal escenario, el hiperparámetro N se convierte en parte del conjunto de parámetros $\Phi = \{\bar{\mu}, \bar{\sigma}, M\}$.

No obstante, resulta prudente señalar que el resultado o valor pronosticado para un punto d en un espacio de entrenamiento D (espacio conformado por el conjunto de datos de entrenamiento) está basado en el modelo $GMM(\bar{\mu}, \bar{\sigma})$, sin incluir N. Esto significa que cada uno de los N kernel gaussianos aportará algún valor de verosimilitud para d con base en la distancia de d de su respectiva media μ y de su propia desviación σ (aquí entra en juego la función de pérdida definida previamente). El parámetro N no está involucrado explícitamente aquí, por lo que es claramente argumentable de que no es en realidad un parámetro del modelo, en su entendido formal. En suma, la distinción entre parámetros e hiperparámetros está matizada debido a la forma en que los profesionales los utilizan al diseñar el modelo M y función de pérdida $\mathcal{L}$.

Como se señala en la misma fuente, "(…) a parameter is learned by the model, whereas the hyperparameter is specified by us (….) Model Parameters are something that a model learns on its own. For example, 1) Weights or Coefficients of independent variables in Linear regression model. 2) Weights or Coefficients of independent variables SVM. 3) Split points in Decision Tree. Model hyper-parameters are used to optimize the model performance. For example, 1) Kernel and slack in SVM. 2) Value of K in KNN. 3) Depth of tree in Decision trees. (…)", sin embargo, es necesario agregar que "They don't necessarily have anything to do with optimizing a model. Hyperparameters are just parameters to the model building process (…) Model parameters are estimated from data automatically and model hyperparameters are set manually and are used in processes to help estimate model parameters. Model hyperparameters are often referred to as parameters because they are the parts of the machine learning that must be set

manually and tuned. Basically, parameters are the ones that the "model" uses to make predictions etc. For example, the weight coefficients in a linear regression model. Hyperparameters are the ones that help with the learning process. For example, number of clusters in K-Means, shrinkage factor in Ridge Regression. They won't appear in the final prediction piece, but they have a large influence on how the parameters would look like after the learning step."

Finalmente, cabe destacar que la definición aquí expuesta sobre los hiperparámetros contiene inclusive a la de Andrew Ng (en el contexto del Aprendizaje Profundo), quien define los hiperparámetros invariablemente como la tasa de aprendizaje α cuyo significado variará según el contexto de aplicación, pudiendo significar desde el número de iteraciones, el número de variables ocultas, el número de unidades a la derecha de cero, entre otras, como puede verificarse en (Ng, 2020).

II.L. Modelo de Mixtura

En Estadística, un modelo de mixtura es un modelo probabilístico utilizado para representar la presencia de subpoblaciones dentro de una población general, sin requerir que un conjunto de datos observados identifique la subpoblación a la que pertenece una observación individual. Formalmente, un modelo de mixtura corresponde a la mixtura de probabilidad que representa la distribución de probabilidad las observaciones de la población general, como puede verificarse en (Wikipedia, 2020). Sin embargo, mientras que los problemas asociados con las mixturas de probabilidad se relacionan con la derivación de las propiedades de la población general a partir de las de las subpoblaciones, los modelos de mixtura se utilizan para hacer inferencias estadísticas sobre las propiedades de las subpoblaciones dadas solo observaciones sobre la población agrupada, sin información de identidad de la subpoblación. Algunas formas de implementar modelos de mixtura involucran pasos que atribuyen identidades de subpoblaciones postuladas a observaciones individuales (o ponderaciones hacia

tales subpoblaciones), en cuyo caso estos pueden considerarse como tipos de aprendizaje no supervisado o procedimientos de agrupamiento. Sin embargo, no todos los procedimientos de inferencia implican tales pasos.

Como se señala en la última fuente citada, un modelo de mixtura de dimensiones finitas típico es un modelo jerárquico que consta de determinados componentes. Antes de exponer tales componentes, es necesario recordar qué es un modelo jerárquico y para ello se expondrán también sus diferencias con las mixturas de probabilidad: "La diferencia entre un modelo jerárquico y una mixtura consiste en que el primero es una función matemática jerarquizada, i.e., es un instrumento de medición cuyo diseño responde al interés de la humanidad por medir fenómenos naturales y sociales para los que es significativamente relevante el hecho de que las etapas en que ocurre su desarrollo poseen una jerarquía bien-definida (según la teoría científica de referencia), mientras que las mixturas de probabilidad son funciones de distribución de probabilidad en que uno o más parámetros de la distribución de probabilidad (que puede ser univariante o multivariante, pero siempre implica probabilidad condicional y una modelación teórica del proceso de estimación de las distribuciones de probabilidad que responde a un diseño por etapas jerarquizadas) se modelan con base al comportamiento de alguna otra variable aleatoria, que condiciona implícitamente el comportamiento de la variable aleatoria estudiada. Estos modelos buscan capturar la noción de variable latente, variable implícita, variable oculta." (Nabi, Algunas Reflexiones Sobre la Distribución Binomial Negativa II (Un Análisis Teórico y Aplicado), 2020, pág. 34). En suma, las mixturas de probabilidad son un tipo de modelo jerárquico, que es un concepto matemático de carácter más abstracto y general que el de mixtura.

Así, los componentes de los modelos de mixtura típicamente son:

- N variables aleatorias observables, en donde cada una está distribuida según una mezcla de K componentes, con los componentes pertenecientes a la misma familia paramétrica de distribuciones (por ejemplo, todas

normales, todas binomial negativa II, etc.), aunque con diferentes parámetros.

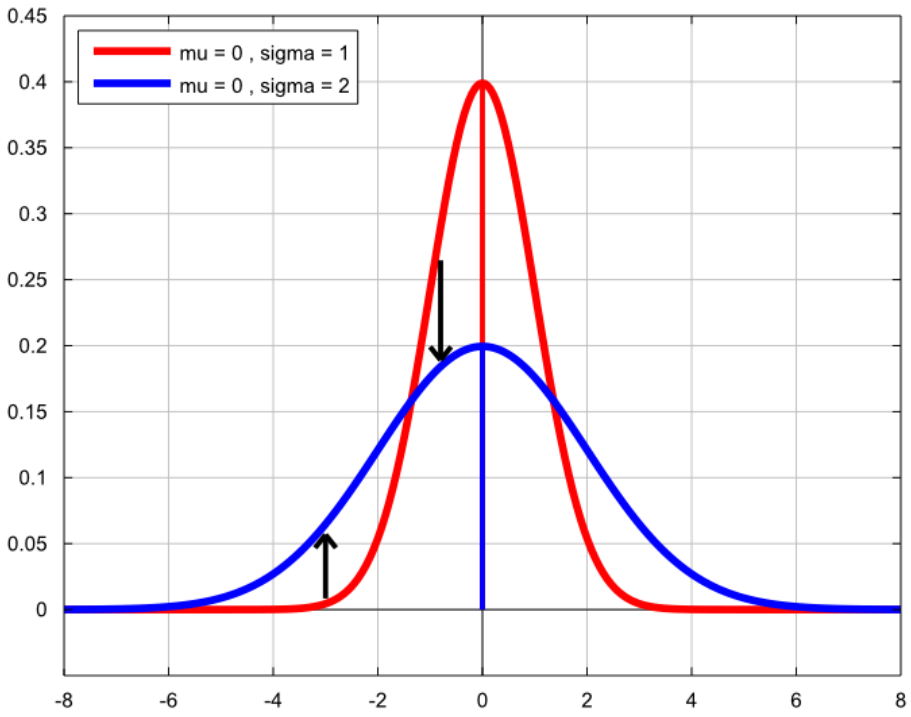
- N variables latentes u ocultas que especifican la identidad del componente de la mixtura en cada observación, cada uno de estos componentes distribuido según una distribución categórica $K - dimensional$.
- Un conjunto K de pesos o ponderaciones de la mixtura, que son probabilidades cuya suma es igual a la unidad.
- Un conjunto K de parámetros, cada uno especificando el parámetro del componente de la mixtura correspondiente. En muchos casos, cada "parámetro" es en realidad un conjunto de parámetros. Por ejemplo, si los componentes de la mixtura son distribuciones gaussianas, habrá una media y una varianza para cada componente. Si los componentes de la mixtura son distribuciones categóricas (por ejemplo, cuando cada observación es una ficha de un alfabeto finito de tamaño V), habrá un vector de probabilidades V cuya suma será igual a la unidad.
- Finalmente, cabe destacar que, en el contexto de la inferencia bayesiana, los pesos y los parámetros de la mixtura serán variables estocásticas y las distribuciones a priori sustituirán a las variables que en inferencia no bayesiana se suelen utilizar. En tal caso, los pesos o ponderaciones se ven típicamente como un vector estocástico de dimensión K extraído de una distribución de Dirichlet (la distribución conjugada a priori de la distribución categórica) y los parámetros se distribuirán de acuerdo con sus respectivos conjugados a priori.

II.LI. Modelos de Mixtura Gaussiana Finita

Como señala (Maklin, 2019), los modelos de mixtura gaussiana pueden ser utilizados para agrupar un conjunto de datos sin etiqueta de forma parecida a como se realiza esta tarea mediante el empleo de $k - medias$. Sin embargo, este último algoritmo de análisis de grupos no tiene en cuenta la varianza, lo que

geométricamente se presenta como la anchura de la curva en forma de campana que se presenta a continuación y el lector identificará inmediatamente:

Figura 16

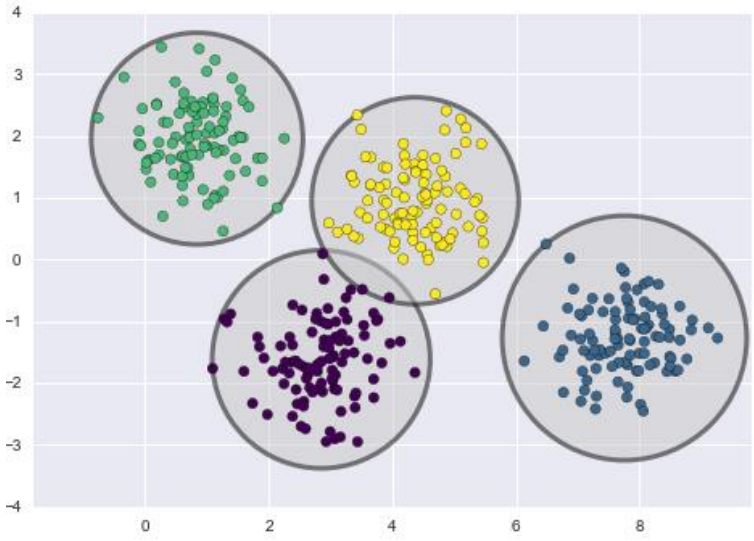


Fuente: (Maklin, 2019).

Cuando se habla de dos dimensiones, la covarianza determina la forma de la distribución. Una forma de pensar en el modelo de $k-medias$ es en que coloca un círculo (o, en dimensiones más altas, una híper-esfera) en el centro de cada grupo

(clúster), cuyo radio (el de cada grupo) está definido por el punto más distante en el grupo del centro del grupo o clúster en cuestión, tal como se observa en la figura presentada a continuación:

Figura 17

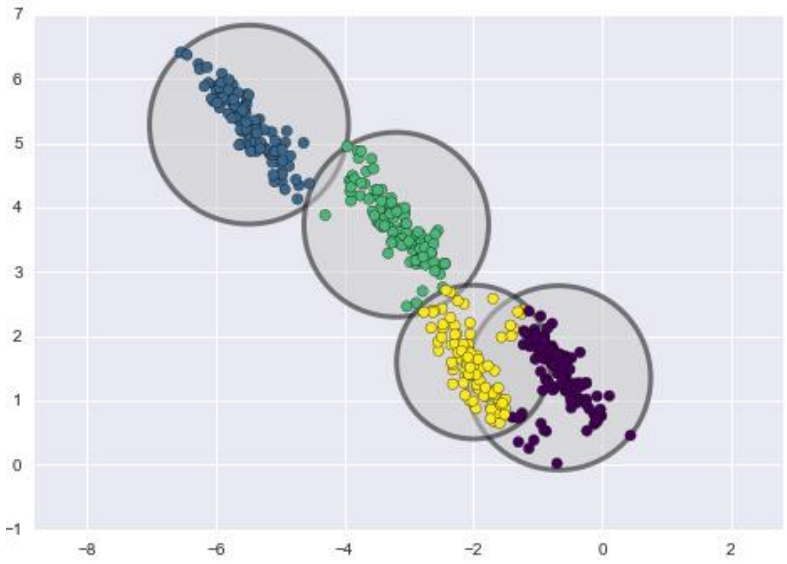


Fuente: (Maklin, 2019).

Por cuestiones puramente geométricas que el lector puede deducir del mismo concepto de radio como longitud característica de un sistema, esta lógica funciona bien cuando el conjunto de datos con el que el investigador trabaja es circular. Sin embargo, cuando tal conjunto de datos toma una forma geométrica diferente el

investigador obtendrá al usar tal algoritmo de análisis de grupos un resultado similar al que se presenta a continuación.

Figura 18

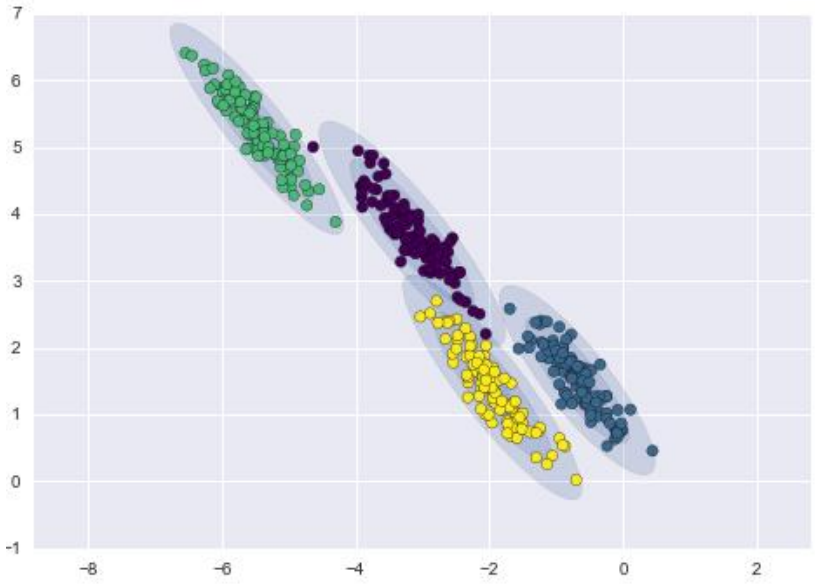


Fuente: (Maklin, 2019).

Como puede observarse en la figura anterior, los datos no quedan muy ajustados cuando la forma geométrica real que toman los datos no es circular, por lo que los modelos de mixtura gaussiana surgen históricamente como solución a esta

deficiencia señalada en los modelos de $k - medias$. En un modelo de mixtura gaussiana, se obtendría con ese mismo conjunto de datos (el empleado en la figura anterior), un resultado como el que se presenta a continuación.

Figura 19



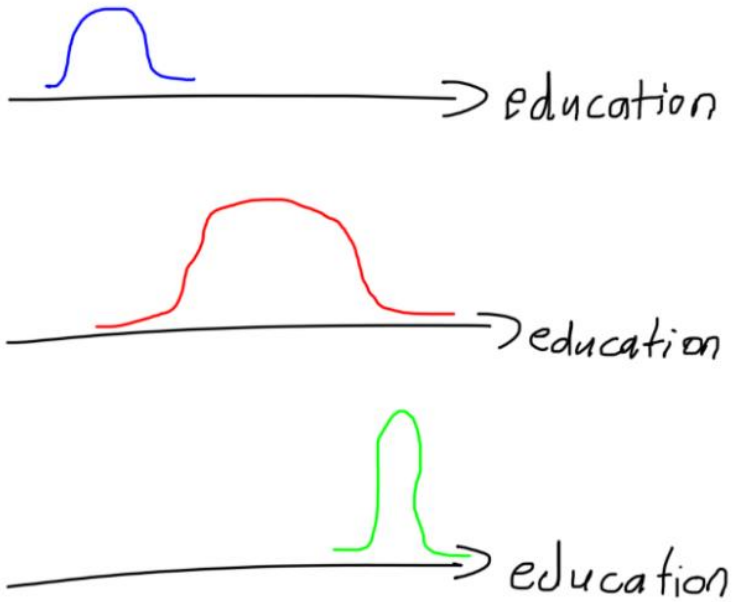
Fuente: (Maklin, 2019).

Como se puede observar, los modelos de mixtura gaussiana manejan de mejor manera los conjuntos de datos cuya forma geométrica tiene la característica de ser sumamente alargada.

La segunda diferencia entre los modelos de mixtura de $k-medias$ y los modelos de mixtura gaussiana es que el primero realiza una clasificación estricta mientras que el segundo realiza una clasificación suave. En otras palabras, el modelo de $k-medias$ nos dice qué punto de datos (observación) pertenece a qué grupo, pero no nos proporcionará las probabilidades de que un punto de datos (observación) determinado pertenezca a cada uno de los posibles grupos.

Como su nombre lo indica, un modelo de mixtura gaussiana implica la mixtura (es decir, la superposición) de múltiples distribuciones gaussianas. Supóngase que se tienen tres distribuciones formadas por muestras de tres clases distintas. En la figura que se presenta a continuación, la distribución Normal en azul es la distribución Normal que sigue el nivel de educación en la clase baja, la de color rojo corresponde al de la clase media y la de color verde a la de la clase alta, según el ejemplo localizado en (Maklin, 2019).

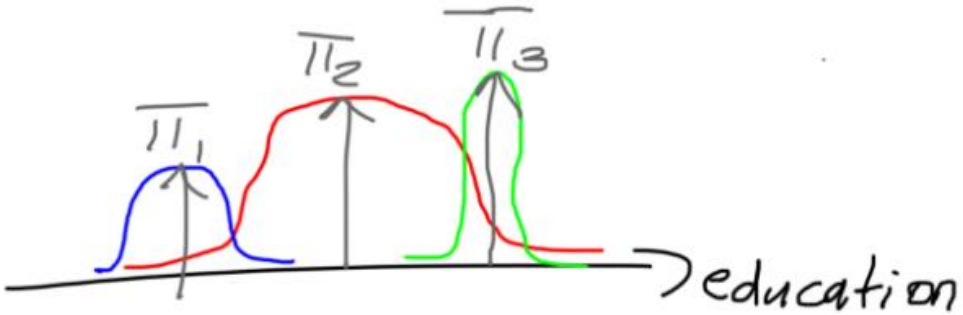
Figura 20



Fuente: (Maklin, 2019).

Sin conocer cuáles muestras provienen de qué clase social el objetivo en este ejemplo es utilizar modelos de mixtura gaussiana para asignar las observaciones (puntos de datos) al grupo (clúster) apropiado. Después de entrenar el modelo, idealmente se terminaría con tres distribuciones en el mismo eje. Luego, dependiendo del nivel de educación de una muestra dada (donde se ubica en el eje), la ubicaríamos en una de las tres categorías, tal como se presenta a continuación.

Figura 21



Fuente: (Maklin, 2019).

Cada distribución se multiplica por un peso o ponderación denotada en la localización referida por π para tener en cuenta el hecho de que no se tienen muestras del mismo tamaño (conjuntos de la misma cardinalidad) en cada categoría. En otras palabras, se podrían haber incluido solo 1,000 personas de la clase alta y 100,000 personas de la clase media, por ejemplo. Puesto que se está

estudiando en el contexto de las probabilidades, la suma de los pesos o ponderaciones $i-$ésimas debe ser igual a la unidad.

Figura 22

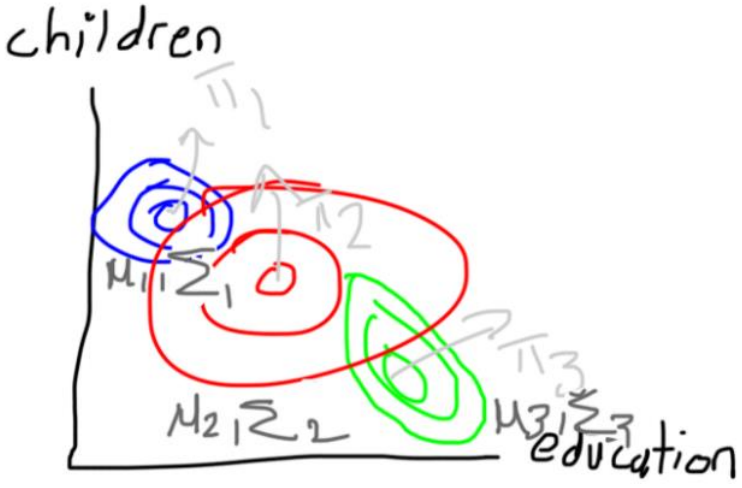
$$\pi = [\underset{\pi_1}{0.31}, \underset{\pi_2}{0.48}, \underset{\pi_3}{0.21}]$$

$$\sum_{K=1}^{K} \pi_K = 1$$

Fuente: (Maklin, 2019).

Si se decide añadir otra dimensión (otra característica) al conjunto de datos como el número de niños, por ejemplo, entonces el resultado final del proceso de agrupamiento podría verse como en la figura presentada a continuación.

Figura 23



Fuente: (Maklin, 2019).

Conviene ahora estudiar el algoritmo de la mixtura gaussiana. En concreto, se desea saber cuál es la probabilidad de que la $i-ésima$ muestra provenga de alguna de las k distribuciones gaussianas (normales).

Algoritmo de Modelación con Mixturas Gaussianas Finitas

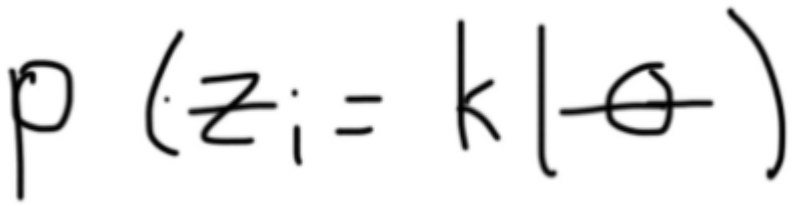
En esta investigación resulta de interés fundamental estudiar la versión 5.2 del paquete estadístico *Mclust* de R, entre cuyos elementos más destacables se encuentra como novedad respecto a versiones anteriores la inclusión de la modelación por mixturas gaussianas finitas, como se verifica (Scrucca, Fop, Murphy, & Raftery, 2016, pág. 291).

Conviene aquí iniciar recordando que para estimar los parámetros de cada grupo gaussiano (de cada clúster cuyos elementos siguen una distribución Normal), es decir, para estimar la media, la varianza y las ponderaciones o pesos de cada distribución Normal que modela los datos contenidos en nada clúster estudiado es necesario primero saber qué muestra pertenece a qué grupo o clúster gaussiano para precisamente poder estimar así tales parámetros. De ahí el papel fundamental que desempeña el algoritmo EM. Una vez funcionando sin inconvenientes el algoritmo EM, el programa estadístico con el que se esté trabajando procederá a realizar el análisis de grupos correspondiente[40].

Como se establece en (Maklin, 2019), lo que se desea conocer es la probabilidad de que la $i-ésima$ muestra proceda de un grupo (clúster) k cuya distribución sea normal. Lo anterior se puede expresar como:

Figura 24

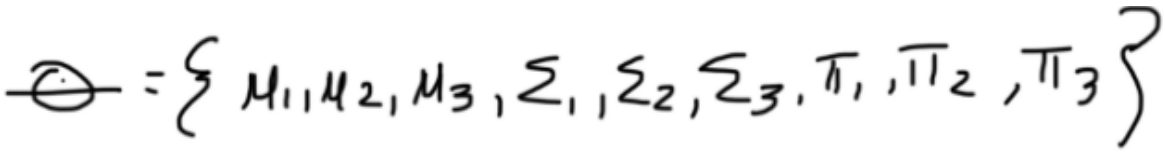
[40] Para el caso de esta investigación será el programa estadístico R, como se verificará más adelante.



Fuente: (Maklin, 2019).

En donde θ representa la media, la covarianza y los pesos o ponderaciones de cada grupo o clúster distribuido normalmente, es decir:

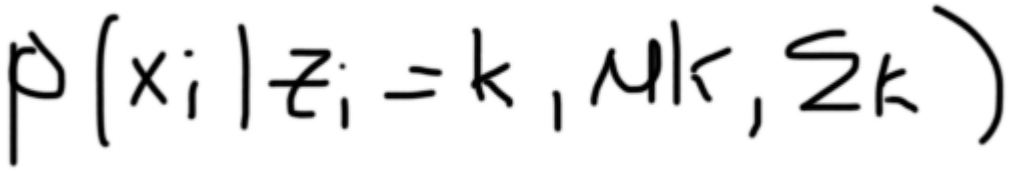
Figura 25



Fuente: (Maklin, 2019).

A continuación, se expresa la probabilidad de observar un punto de datos (una observación) dado que provenga de algún clúster gaussiano k como:

Figura 26



Fuente: (Maklin, 2019).

La versión más familiar de la identidad expuesta en la figura anterior para el contexto de la presente investigación (relacionado con las mixturas gaussianas) puede expresarse formalmente como se expone en (Scrucca, Fop, Murphy, & Raftery, 2016, pág. 291).

Sea $x = \{x_1, x_2, \dots, x_i, \dots, x_n\}$ una muestra de n observaciones independientes e idénticamente distribuida. La distribución de probabilidad de cada observación está especificada por su función de densidad de probabilidad estimada a través de un modelo de mixtura finita con G componentes, el cual toma la forma:

$$f(x_i; \Psi) = \sum_{k=1}^{G} \pi_k f_k(x_i; \theta_i)$$

En la expresión anterior, $\Psi = \{\pi_1, \pi_2, \dots, \pi_{G-1}, \theta_1, \dots, \theta_G\}$ son los parámetros del modelo de mixtura, mientras que $f(x_i; \Psi)$ es el k-ésimo componente de densidad (la k-ésima función de densidad estimada por la mixtura) para la observación x_i con el vector de parámetros θ_k, siendo $(\pi_1, \pi_2, \dots, \pi_{G-1})$ los pesos o ponderaciones de la mixtura (*i.e.*, las probabilidades mismas, tal que $\pi_k > 0, \sum_{k=1}^{G} \pi_k = 1$) y G es el número de componentes de la mixtura.

El lector debe recordar que esta selección óptima que realiza el algoritmo en cada ciclo tiene como orquesta de fondo que lo anima la lógica de la máxima verosimilitud, que debe ser familiar al lector desde los cursos elementales de Estadística. Por ello, la probabilidad se estima a través del método de máxima verosimilitud vía el algoritmo EM y es a causa de eso que Scrucca et al plantean lo que se contiene en la figura presentada a continuación[41]:

Figura 27

Assuming that G is fixed, the mixture model parameters $\boldsymbol{\Psi}$ are usually unknown and must be estimated. The log-likelihood function corresponding to equation (1) is given by $\ell(\boldsymbol{\Psi}; \boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \sum_{i=1}^{n} \log(f(\boldsymbol{x}_i; \boldsymbol{\Psi}))$. Direct maximisation of the log-likelihood function is complicated, so the maximum likelihood estimator (MLE) of a finite mixture model is usually obtained via the EM algorithm (Dempster et al., 1977; McLachlan and Peel, 2000).

Fuente: (Scrucca, Fop, Murphy, & Raftery, 2016, pág. 291).

[41] Junto con cierto contenido de Álgebra Lineal necesario, queda pendiente añadir en los anexos la parte del marco teórico referente únicamente al aspecto teórico de la optimización de estimadores realizada por el Método de Máxima Verosimilitud.

Como se señala en la última fuente citada, el abordaje del análisis de grupos basado en el modelado tiene como característica que cada componente de la mixtura de densidad finita es usualmente asociado a un grupo o clúster. La mayor parte de las aplicaciones de este enfoque analítico asumen que todos los componentes (que son densidades) provienen de la misma familia de distribución paramétrica[42], aunque este no es necesariamente el caso general.

Como señalan Scrucca et al, un modelo de mixtura popular entre los investigadores estadísticos es el modelo de mixtura gaussiana (GMM, por su nombre en inglés), el cual asume de forma multivariable distribuciones normales para cada uno de sus componentes integrantes, *i.e.*, $f_K(x;\ \theta_K) \sim N(\mu_K, \Sigma_K)$. Puesto que los clústeres son elipsoidales, centrados en el vector de medias μ_K, así como también con otras características geométricas tales como el volumen, la forma y la orientación, todas ellas determinadas por la matriz de covarianzas Σ_K.

Una parametrización parsimoniosa[43] de las matrices de covarianza puede ser obtenida a través de las medias provistas por una *eigen descomposición* o *descomposición característica*[44] de la forma $\Sigma_K = \lambda_K D_K A_K D_K$, en donde λ_K es un

[42] Una familia paramétrica se define como aquella familia (colección de objetos en que cada uno está asociado a alguna lista o índice de algún conjunto) que contiene objetos (relacionados entre sí de alguna manera) cuyas diferencias dependen únicamente de los valores que sean escogidos de un conjunto de parámetros.

[43] Una parametrización es el proceso de definir o escoger parámetros. Además, e intentando eludir en la medida de lo posible una discusión sobre Filosofía de la Ciencia, una parametrización parsimoniosa es aquella parametrización que cumple el principio de parsimonia. Este principio se basa en el *principio de la navaja de Occam* (conocido también como ley de brevedad –"law of briefness"-, originalmente *lex parsimoniae* en latín). Lo que dice el principio de Occam es que entre dos respuestas a una pregunta usualmente la explicación más simple es la correcta; Einstein se encargó de criticar brillantemente esto, así como también la realidad ha sido contundente al respecto. Sin embargo, en el contexto "puramente" de la Estadística, una parametrización parsimoniosa es aquella que usa el número óptimo de parámetros para explicar el conjunto de datos de los que dispone el investigador.

[44] Para comprender esto se debe comenzar por comprender algunos conceptos previos. Como se señala en (Wikipedia, 2020), en Lingüística, Neuropsicología y Filosofía del Lenguaje, un *lenguaje*

escalar de control del volumen del cuerpo elipsoidal, A_K es la matriz diagonal que especifica la forma de los contornos de la densidad (los gráficos en forma de disco antes expuestos) cuyo determinante (puesto que puede ser expresada mediante una matriz cuadrada invertible) es igual a la unidad, es decir, al volumen o medida de Lebesgue del Campo Probabilístico de Kolmogórov (que por definición es 1, puesto que las probabilidades van de 0 a 1), mientras que D_K es la matriz ortogonal que determina la orientación del correspondiente cuerpo elipsoidal. En un modelo unidimensional, sólo existen dos modelos:

1) El modelo E para idéntica varianza.
2) El modelo V para varianza variable.

Por otro lado, en la configuración multivariable del modelo, el volumen, la forma y la orientación de los cuerpos elipsoidales (determinado todos ellos por la matriz de

---

natural (o lenguaje ordinario) es un lenguaje que ha evolucionado naturalmente en humanos a través de su uso repetido sin planificación consciente o premeditación; y como se señala en (Wikipedia, 2020), el *procesamiento de lenguaje natural* es un subcampo de la Lingüística, las Ciencias de la Computación y la Inteligencia Artificial que se ocupa de las interacciones entre las computadoras y el lenguaje humano, en particular cómo programar las computadoras para procesar y analizar grandes cantidades de datos del lenguaje natural. A su vez, codificar es el acto de codificar caracteres, lo que según (Wikipedia, 2020) es el método que permite convertir un carácter de un lenguaje natural (como el de un alfabeto o silabario) en un símbolo de otro sistema de representación, como un número o una secuencia de pulsos eléctricos en un sistema electrónico, aplicando normas o reglas de codificación. Como se señala en (Nag, 2016), *representación* (en el contexto del Aprendizaje Automático) es básicamente el espacio de todos los modelos de aprendizaje permitidos (el *espacio de hipótesis o espacio nulo*), que además toma en cuenta el hecho de que el investigador busca expresar el modelo en algún lenguaje formal que puede codificar algunos modelos de aprendizaje con mayor facilidad que otros. En Ciencias de la Computación, un proceso de *canonicalización* (conocido también como estandarización o normalización) es aquel diseñado para convertir datos que tienen más de una representación posible en una forma "estándar", "normal", "canónica" (en relación a las prácticas normales de la comunidad científica de la que se trate -pues puede variar-) para diversos fines, sobre lo que puede ampliarse en (Wikipedia, 2020). Finalmente, según (Wikipedia, 2020), una descomposición característica (a veces llamada *descomposición espectral*) de una matriz (de datos, para este caso) es el proceso de factorización algebraica para representar la matriz de datos en su forma canónica (según lo definido con antelación), lo que en este contexto específico implica que la matriz de datos se representa en términos de sus valores característicos y vectores característicos. Únicamente las matrices diagonalizables (una matriz cuadrada A es diagonalizable es aquella para la cual existe una matriz invertible P y una matriz diagonal D tal que $P^{-1}AP = D$ o de forma equivalente $A = PDP^{-1}$ -en donde P, D no son únicas-) pueden ser representadas de esta manera.

covarianzas) pueden ser restringidas[45] para que sean iguales entre los grupos. Así, dentro de este paquete estadístico disponible de forma gratuita en R, existen 14 modelos gaussianos disponibles con diferentes características geométricas. En la figura presentada a continuación se exponen todos estos modelos, mostrando además su estructura geométrica de distribución de probabilidad, volumen, forma, orientación y los nombres asociados a ellos.

Figura 28

Model	$\boldsymbol{\Sigma}_k$	Distribution	Volume	Shape	Orientation
EII	$\lambda \boldsymbol{I}$	Spherical	Equal	Equal	—
VII	$\lambda_k \boldsymbol{I}$	Spherical	Variable	Equal	—
EEI	$\lambda \boldsymbol{A}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k \boldsymbol{A}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda \boldsymbol{A}_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k \boldsymbol{A}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Variable	Variable

Table 3: Parameterisations of the within-group covariance matrix $\boldsymbol{\Sigma}_k$ for multidimensional data available in the **mclust** package, and the corresponding geometric characteristics.

[45] Que hace alusión a constreñir en el contexto de la Optimización Matemática.

Fuente: (Scrucca, Fop, Murphy, & Raftery, 2016, pág. 292).

A continuación de presentan las elipses de isodensidad (que son descritas por la misma densidad) para cada uno de los 14 modelos gaussianos disponibles en el paquete estadístico estudiado en esta investigación, los cuales han sido obtenidos por descomposición característica o espectral para el caso de tres grupos en un espacio de dos dimensiones.

Figura 29

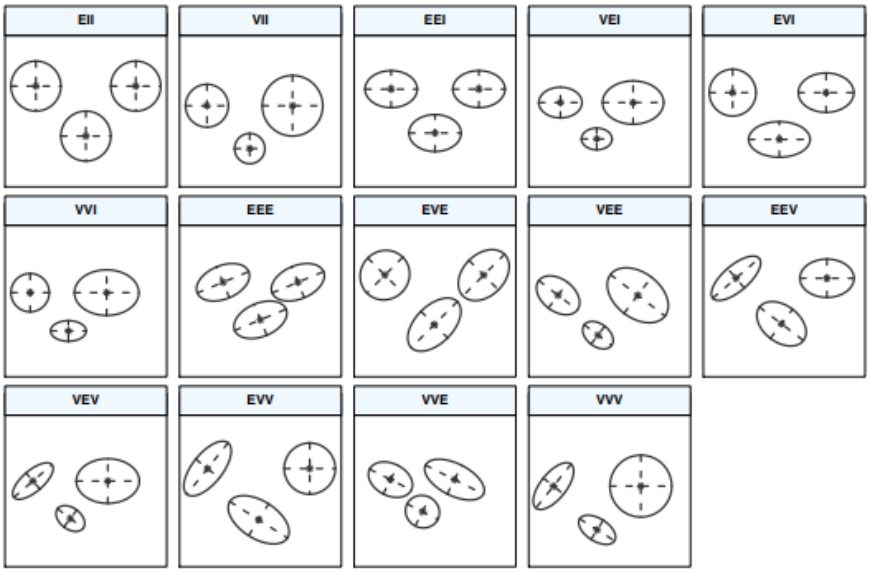


Figure 2: Ellipses of isodensity for each of the 14 Gaussian models obtained by eigen-decomposition in case of three groups in two dimensions.

Fuente: (Scrucca, Fop, Murphy, & Raftery, 2016, pág. 292).

Conviene preguntarse entonces por qué los contornos de probabilidad de las distribuciones multivariantes de carácter gaussiano son elípticos. Como se señala en (Chughes, 2013), todo contorno gaussiano de dos dimensiones concentra las observaciones en algún punto particular (en algún "chichón" o "pico"), con las observaciones alejándose constantemente de ese punto particular y si se grafican las regiones (de uno de esos contornos) que tienen la misma altura en el "chichón" o "pico" mencionado (que implica que se posee la misma distribución de densidad de probabilidad bajo una función de densidad de probabilidad determinada), se forma geométricamente hablando en consecuencia una elipse.

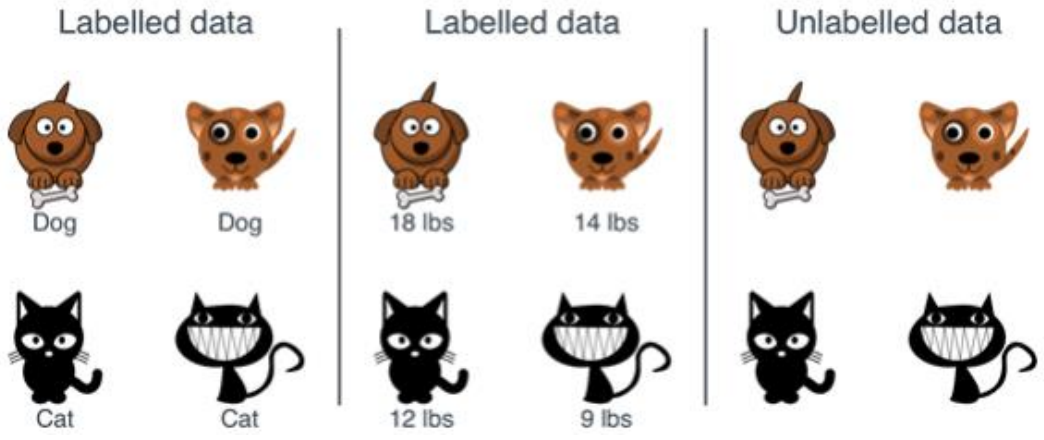
La interpretación probabilística de este hecho geométrico es que, matemáticamente hablando, la elipse de dos dimensiones de eje alineado (simétrica) se expresa de la siguiente manera:

$$1 = \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2}$$

El lector puede notar que el lado izquierdo de la identidad coincide con la medida del Campo Probabilístico de Kolmogórov y a su vez es equivalente a la estructura matemática de la elipse descrita. No será en esta investigación en la que se entrará en un estudio profundo de este hecho, aunque sin lugar a dudas será materia abordada con la profundidad requerida en alguna otra investigación.

Lo que se ha denominado *etiqueta*, que más precisamente debería denominarse como *conjunto de datos etiquetados* ("labeled data"), es el conjunto de valores de las predicciones realizadas sobre un hecho natural o social en concreto. Como se señala en (Serrano, 2020, pág. 31), "Normally, if we are trying to predict a feature based on the others, that feature is the label. If we are trying to predict the type of pet, we have (for example cat or dog), based on information on that pet, then that is the label. If we are trying to predict if the pet is sick or healthy based on symptoms and other information, then that is the label. If we are trying to predict the age of the pet, then the age is the label."

Figura 30



Fuente: (Serrano, 2020, pág. 31).

Así, cuando se habla de datos incompletos, se hace referencia a la condición empírica de incompletitud de los datos etiquetados, lo que implica que algunos datos no tengan etiqueta (predicción asignada). Sin embargo, el creador del bootstrapping, Bradley Efron, presenta una definición más amplia de datos incompletos que permite aplicarla a una clase más general de problemas. Según (Hardy & Bryman, 2009, pág. 113), "Missing data are a pervasive problem in almost all areas of empirical research. They arise, for example, during data recording (a datum is omitted), when responses are related to sensitive questions (e.g., age, income, drug use), when measurements of some the variables is too expensive (e.g., measurement may require destroying expensive parts, an interviewer needs to travel a long distance), or when the experiment is run on a group of individuals over a period of time as in clinical studies (e.g., individuals may drop out of the study or not show up for certain visits). Efron (1994) defines *missing data* as a class of problems made difficult by the absence of some part of a familiar data structure. In the examples mentioned above, the missing structure is an observable covariate that is not recorded or observed. Efron's definition of missing data covers a broader class of problems. For example, the latent variables n factor analysis would be considered missing data by his definition. Viewing unobservable (latent) data as missing data can help us gain insight into models and develop computational methodology. For example, Rubin and Thayer (1982) developed and expectation-maximization (EM) algorithm to compute maximum likelihood (ML) estimates for the factor analysis model by treating the latent variables as missing data."

Así, se comprende la relación entre un conjunto de datos incompleto, valores ausentes, datos con etiqueta, datos sin etiqueta y variables ocultas, que son todos los conceptos que surgen al entender a profundidad en qué consiste el análisis de grupos mediante el uso de modelos de mixtura gaussiana.

III. CASOS DE APLICACIÓN CON EL PAQUETE ESTADÍSTICO R

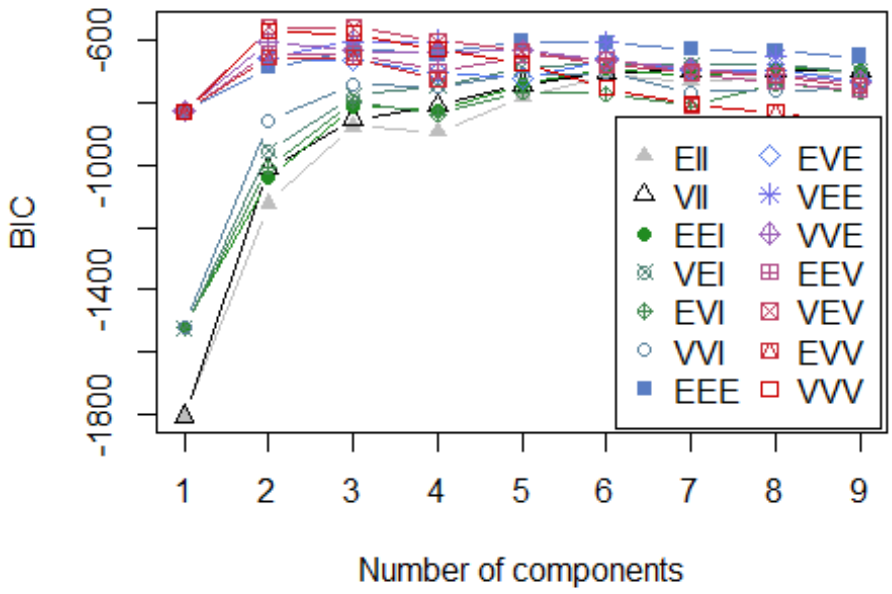
III.I. Base de Datos Iris de R

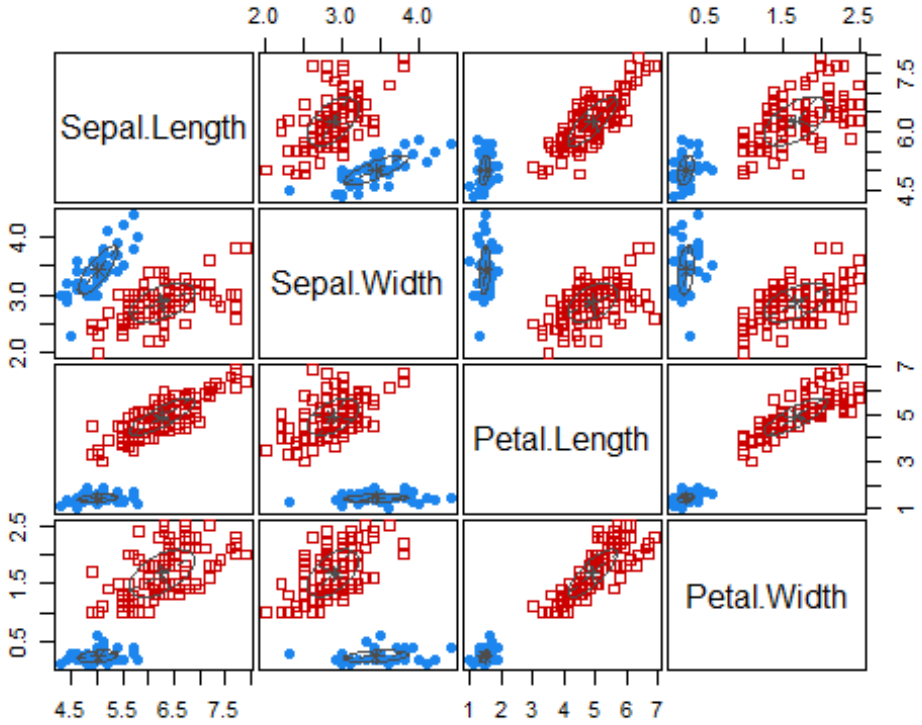
Este es un ejemplo utilizando la base de datos llamada "iris" que viene en R y este ejemplo viene en la viñeta del paquete "mclust".

```
mod1 <- Mclust(iris[,1:4])
summary(mod1)

## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEV (ellipsoidal, equal shape) model with 2 components:
##
##  log-likelihood   n df       BIC       ICL
##       -215.726 150 26 -561.7285 -561.7289
##
## Clustering table:
##   1   2
##  50 100

plot(mod1, what = c("BIC", "classification"))
```






```
data <- mod1$data
BIC <- mod1$BIC
```

III.II. Base de Datos del Banco Mundial en R

Este ejemplo es propio y fue creado utilizando las bases de datos del Banco Mundial. Se pretendía investigar el comportamiento de las siguientes variables: gasto público en educación (% del PIB), el desempleo (% de la población activa total), la tasa de incidencia de la pobreza, sobre la base de la línea de pobreza nacional (% de la población) y la tasa de alfabetización, total de adultos (% de personas de 15 años o más).

```
# Datos de países

Gasto_publico <- read_excel("Gasto_publico.xls", skip = 3)
Alfabetizacion <- read_excel("Alfabetizacion.xls", skip = 3)
Desempleo <- read_excel("Desempleo.xls", skip = 3)
Pobreza <- read_excel("Pobreza.xls", skip = 3)

colnames(Gasto_publico) <- make.names(colnames(Gasto_publico))
colnames(Alfabetizacion) <- make.names(colnames(Alfabetizacion))
colnames(Desempleo) <- make.names(colnames(Desempleo))
colnames(Pobreza) <- make.names(colnames(Pobreza))

Gasto_publico2 <- select(Gasto_publico, Country.Name, X2018)
Alfabetizacion2 <- select(Alfabetizacion, Country.Name, X2018)
Desempleo2 <- select(Desempleo, Country.Name, X2018)
Pobreza2 <- select(Pobreza, Country.Name, X2018)

colnames(Gasto_publico2)[2] <- "Gasto_publico_2018"
colnames(Alfabetizacion2)[2] <- "Alfabetizacion_2018"
colnames(Desempleo2)[2] <- "Desempleo_2018"
```

```r
colnames(Pobreza2)[2] <- "Pobreza_2018"


Datos <- left_join(Gasto_publico2, Alfabetizacion2)
Datos <- left_join(Datos, Desempleo2)
Datos <- left_join(Datos, Pobreza2)
Datos <- mutate(Datos, na = rowSums(Datos[,c(2,3,4,5)], na.rm = FALSE))
unique(Datos$na)
```

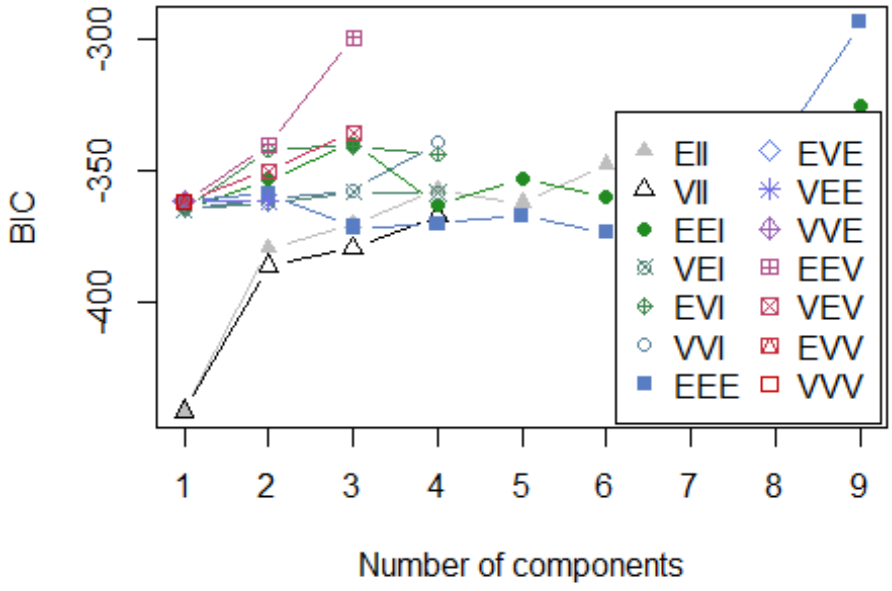
```
##  [1]       NA  94.10224  93.13138 135.66100 135.57436 147.22070 131.92046
##  [8] 108.28567  79.86821 122.02143 111.41296 122.90702 120.18745 107.86083
```

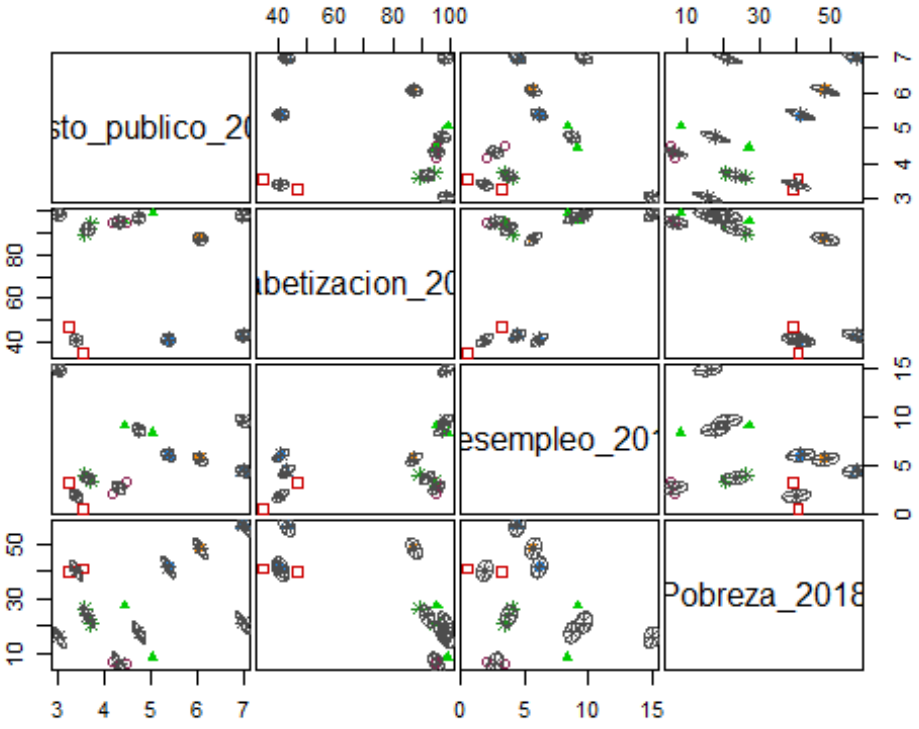
```r
Datos2 <- filter(Datos, !is.na(na))


clust <- Mclust(Datos2[,2:5])
summary(clust)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEE (ellipsoidal, equal volume, shape and orientation) model with 9
## components:
##
##  log-likelihood  n df       BIC       ICL
##       -77.62892 13 54 -293.7651 -293.7727
##
## Clustering table:
## 1 2 3 4 5 6 7 8 9
## 1 2 2 1 1 1 2 2 1
```

```r
plot(clust, what = c("BIC", "classification"))
```


BIC
-300
-350
-400
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV


40
60
80
100
10
30
50
sto_publico_2
betizacion_2
esempleo_20
Pobreza_2018
3
4
5
6
7
0
5
10
15

```r
Datos3 <- left_join(Gasto_publico2, Alfabetizacion2)
Datos3 <- mutate(Datos3, na = rowSums(Datos3[,c(2,3)], na.rm = FALSE))
unique(Datos3$na)
```

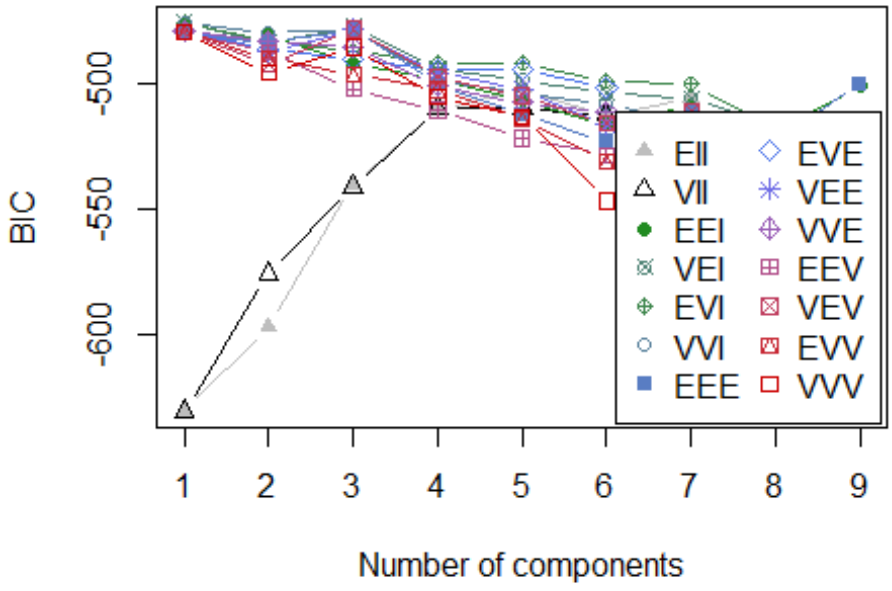
```
##  [1]       NA  45.29637  46.60824  50.42138  80.20224  83.80981  99.55000
##  [8] 104.84236  83.02908  93.27270  65.43010  70.97800  69.19531 101.26146
## [15] 102.39758  86.84122  67.97232  93.83403  79.42107  77.62981  96.15998
## [22]  99.33567  38.59721  73.01050  98.13143  65.69791  76.28934  50.19596
## [29]  92.60102 103.50929  68.98045  69.34274  97.96571  98.99375  74.85978
## [36]  78.65802 103.75145  99.16783  91.36605
```

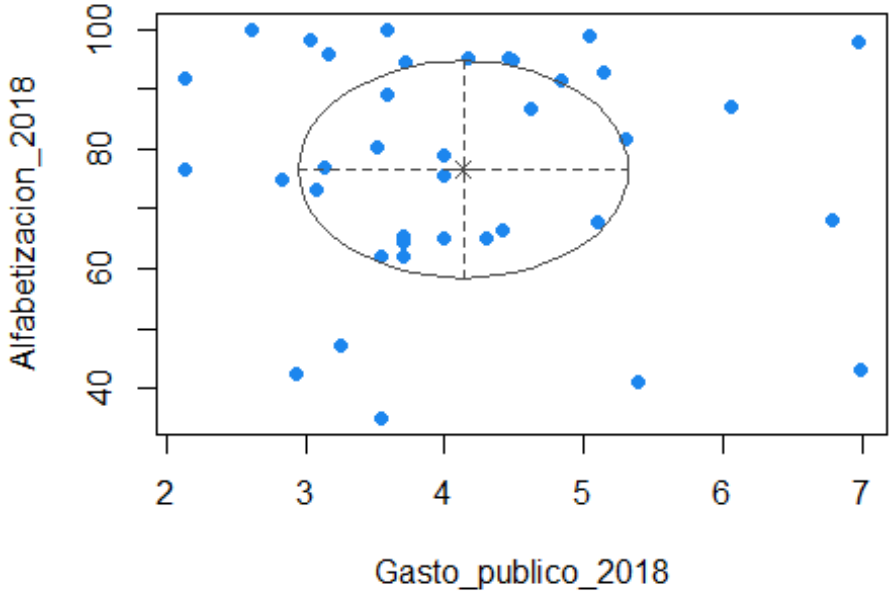
```r
Datos4 <- filter(Datos3, !is.na(na))


clust <- Mclust(Datos4[,2:3])
summary(clust)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust XXI (diagonal multivariate normal) model with 1 component:
##
##  log-likelihood  n df       BIC       ICL
##       -230.3463 39  4 -475.3468 -475.3468
##
## Clustering table:
##  1
## 39
```

```r
plot(clust, what = c("BIC", "classification"))
```


BIC
-500
-550
-600
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV


Alfabetizacion_2018
100
80
60
40
2
3
4
5
6
7
Gasto_publico_2018

```r
Datos3 <- left_join(Gasto_publico2, Desempleo2)
Datos3 <- mutate(Datos3, na = rowSums(Datos3[,c(2,3)], na.rm = FALSE))
unique(Datos3$na)
```

```
##  [1]       NA  7.355430  6.526650  5.319970 11.477790 14.073490  9.153840
##  [8]  6.466030  6.488200 13.181050 13.568490 16.610570 13.787592 13.884740
## [15] 17.306060  8.146440  6.562940 11.447300  5.542930 17.399580  6.229520
## [22] 11.715460  7.968844 16.372470  8.368308  9.767569  8.065290 15.995360
## [29] 14.518780 17.993350  7.440950  7.941250  2.814860  5.516120 24.210800
## [36]  8.052057  6.243210  9.121667 30.956191  4.704700  8.417730  4.514490
## [43]  3.416890  8.736850 11.491590 10.336840  7.831260  4.018210  6.452070
## [50]  7.113160  4.353890  9.422262  4.090750 11.360280 11.406630  7.598420
## [57] 16.318390 10.148124 10.463180 18.297070  7.034699  7.448570 11.271950
## [64] 10.463180  5.686440  3.879520 13.383590 24.572780  6.160450 33.078990
## [71] 16.118090 10.958350
```

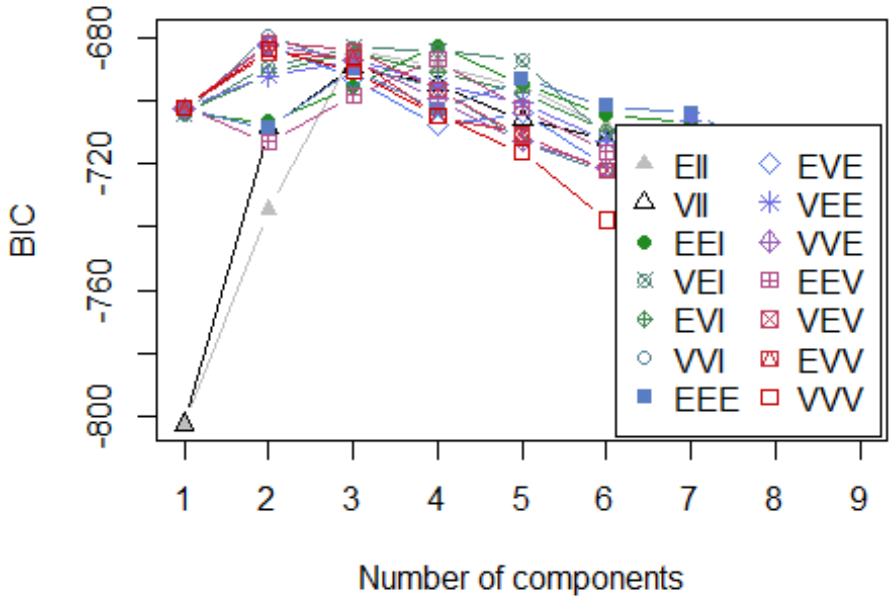
```r
Datos4 <- filter(Datos3, !is.na(na))
```

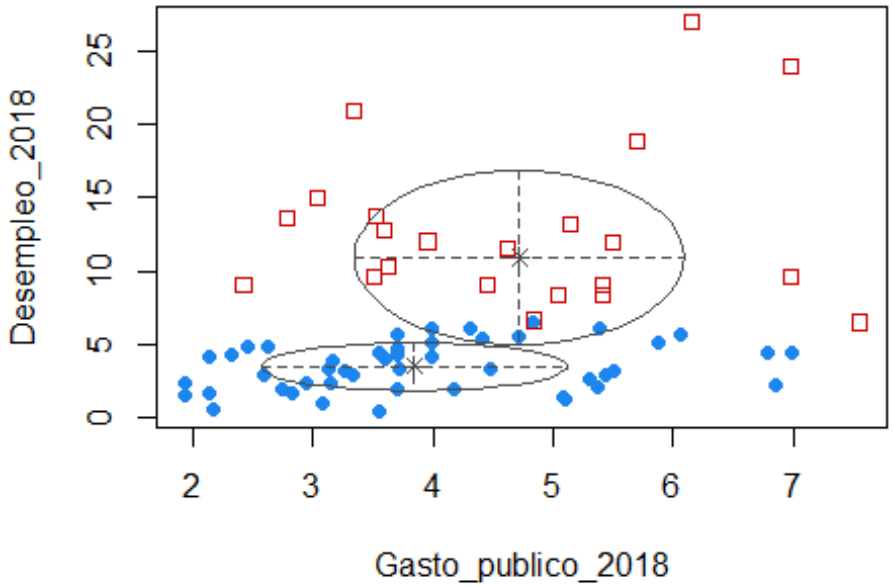
```r
clust <- Mclust(Datos4[,2:3])
summary(clust)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVI (diagonal, varying volume and shape) model with 2 components:
##
##  log-likelihood  n df       BIC       ICL
##       -320.4178 71  9 -679.1997 -696.6834
##
```

```
## Clustering table:
##  1  2
## 49 22

plot(clust, what = c("BIC", "classification"))
```


BIC
-680
-720
-760
-800
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV


Desempleo_2018
0
5
10
15
20
25
2
3
4
5
6
7
Gasto_publico_2018

```r
Datos3 <- left_join(Gasto_publico2, Pobreza2)
Datos3 <- mutate(Datos3, na = rowSums(Datos3[,c(2,3)], na.rm = FALSE))
unique(Datos3$na)
```

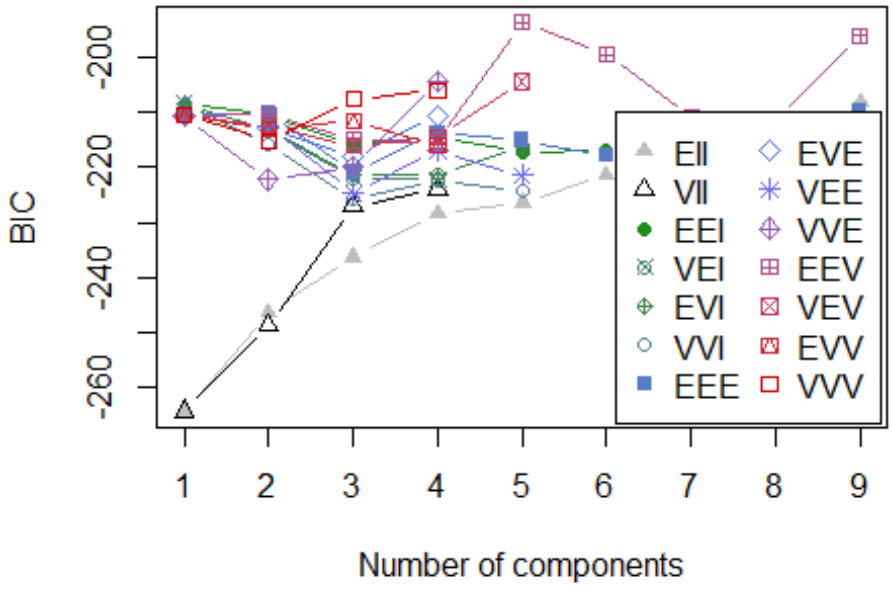
```
##  [1]       NA 46.78379 42.75603 31.45749 28.07857 23.62106 46.01594 54.36746
##  [9] 18.73435 28.43973 10.08126 44.34721 24.22316 63.78963 29.89242 30.09644
## [17] 13.14759 10.86745
```

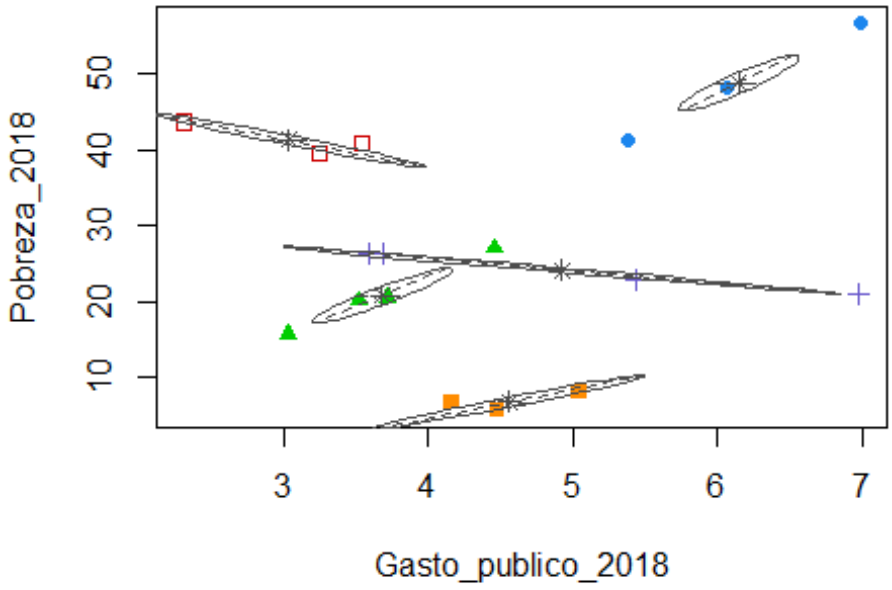
```r
Datos4 <- filter(Datos3, !is.na(na))


clust <- Mclust(Datos4[,2:3])
summary(clust)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEV (ellipsoidal, equal volume and shape) model with 5 components:
##
##  log-likelihood  n df       BIC       ICL
##       -67.09049 17 21 -193.6785 -193.6787
##
## Clustering table:
## 1 2 3 4 5
## 3 3 4 4 3
```

```r
plot(clust, what = c("BIC", "classification"))
```


BIC
-200
-220
-240
-260
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV


Pobreza_2018
10
20
30
40
50
3
4
5
6
7
Gasto_publico_2018

IV. ANEXOS

IV.I. Operador

Símbolo matemático que indica que debe ser efectuada una operación específica sobre un cierto número de variables (números, funciones, vectores, entre otros).

IV.II. Operandos

Variables sobre las cuales se especifica una operación a realizar mediante un operador.

IV.III. Operación Binaria

Operación matemática que para el cálculo de un valor requiere de un operador y dos operandos.

IV.IV. Estructura Algebraica

Es aquella $n - pla$ de la forma $\lambda_1, \lambda_2, \dots, \lambda_n$, en el cual λ_1 es un conjunto preestablecido no nulo y $\lambda_2, \dots, \lambda_n$ es un conjunto de operaciones aplicables a los elementos del conjunto λ_1.

IV.V. Ley de Composición

Aquella operación binaria que da lugar a distintas estructuras algebraicas.

IV.VI. Ley de Composición Interna

Una ley de composición es interna si la operación binaria asigna a todo par ordenado cuyos elementos pertenecen a un conjunto determinado, un tercer elemento que pertenece también a dicho conjunto, por ejemplo, las suma entre dos números naturales (será siempre un número natural) o la multiplicación entre dos números racionales (será siempre un número racional), así como la unión y la intersección de dos conjuntos, es decir, la formación de un nuevo conjunto que incluya todos los elementos de los conjuntos unidos (sin repeticiones) y la

formación de un conjunto que incluya solo los elementos que los conjuntos intersecados tienen en común, respectivamente.

Formalmente se expresa como:

Dado un conjunto A y una operación $\odot$, representado como $2-tupla$ horizontal $(A, \odot)$, $\odot$ es una ley de composición interna en el conjunto A cuando es una función (la ley de composición que asigna a cada elemento de un primer conjunto un único elemento de un segundo conjunto[46]) de la siguiente forma:

$$\odot: A \times A \longrightarrow A$$

$$(a, b) \longmapsto c = a \odot b$$

$$\forall (a, b) \in A \times A, \qquad \exists! \, c \in A: c = a \odot b$$

IV.VII. Ley de Composición Externa

Una ley de composición es externa, si los dos operandos no pertenecen al mismo conjunto.

IV.VII.I. Ley de Composición Externa por la Derecha

Es externa por la derecha si a cada par ordenado (a, b) de $A \times B$ (donde A y B con conjuntos distintos) le asigna un elemento c que pertenece al conjunto A, cumpliéndose que ese elemento asignado al par ordenado (tal elemento pertenece a la operación $A \times B$) es único y a su vez es resultado de operar los elementos del par ordenado.

Formalmente se expresa como:

Dados dos conjuntos A y B, así como una operación $:\cdot$, representado como la $3-tupla$ horizontal $(A, B, \cdot)$:

$$\cdot : A \times B \longrightarrow A$$

[46] Evidentemente, puede tratarse del mismo conjunto.

$$(a, b) \longmapsto c = a \cdot b$$

para la que se define una función que asigna un elemento c que pertenece a A a cada par ordenado (a, b) resultante de la operación $A \times B$, es decir:

$$\forall (a, b) \in A \times B, \qquad \exists! \, c \in A: c = a \cdot b$$

IV.VII. II. Ley de Composición Externa por la Izquierda

Es externa por la derecha si a cada par ordenado (a, b) de $A \times B$ (donde A y B con conjuntos distintos) le asigna un elemento c que pertenece al conjunto B, cumpliéndose que ese elemento asignado al par ordenado (tal elemento pertenece a la operación $A \times B$) es único y a su vez es resultado de operar los elementos del par ordenado.

Formalmente se expresa como:

Dados dos conjuntos A y B, así como una operación $:\circ$, representado como la $3-tupla$ horizontal $(A, B, \circ)$:

$$\circ : A \times B \longrightarrow A$$

$$(a, b) \longmapsto c = a \circ b$$

para la que se define una función que asigna un elemento c que pertenece a A a cada par ordenado (a, b) resultante de la operación $A \times B$, es decir:

$$\forall (a, b) \in A \times B, \qquad \exists! \, c \in A: c = a \circ b$$

Como puede observarse, una *función* puede ser una ley de composición interna o externa, dependiendo de la relación que establezca a través del operador entre los operandos y del valor único asignado a dicha operación y según las funciones definidas entre determinados conjuntos y sus correspondientes elementos, se obtendrán diversas estructuras algebraicas. La estructura algebraica de interés en esta investigación es el espacio $n-euclideano$, que no es más que un espacio vectorial que cumple determinadas condiciones adicionales, en el cual se encuentran definidas una ley de composición interna llamada *suma* para los

elementos del conjunto y una ley de composición externa llamada *producto por un escalar* definida entre dicho conjunto y otro conjunto llamado campo o cuerpo, el cual tiene dos leyes de composición interna llamadas *adición* y *multiplicación*, las cuales a su vez cumplen con determinadas propiedades y en el cual existen determinados elementos[47].

[47] Si el lector desea ampliar al respecto, puede consultar en (Nabi, Un Modesto Ensayo Sobre las Funciones de Probabilidad y la Esperanza Matemática en Espacios de Probabilidad, 2017).

V. REFERENCIAS

Adamchik, V. S. (17 de Noviembre de 2020). *Binary trees*. Obtenido de School of Computer Science - Carnegie Mellon University: https://www.cs.cmu.edu/~adamchik/15-121/lectures/Trees/trees.html

Amazon Web Services. (12 de Noviembre de 2020). *Entrenamiento de modelos de ML*. Obtenido de Guía para desarrolladores: https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/training-ml-models.html

Asiri, S. (11 de Junio de 2018). *What is classification?* Obtenido de Towards Data Science: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

Bayes, T. (23 de Diciembre de 1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 370-418.

Bellman, R. (1972). *Dynamic Programming* (Sexta Impresión ed.). New Jersey: Princeton University Press.

Bhattacharjee, J. (27 de Octubre de 2017). *Some Key Machine Learning Definitions.* Obtenido de Technology at Nineleaps: https://medium.com/technology-nineleaps/some-key-machine-learning-definitions-b524eb6cb48#:~:text=Label%3A%20Labels%20are%20the%20final,to%20one%20or%20more%20labels.

Bien, J., & Tibshirani, R. (17 de Agosto de 2009). *Classification by Set Cover: The Prototype Vector Machine.* Obtenido de Machine Learning: https://arxiv.org/pdf/0908.2284.pdf

Chughes, M. (22 de Enero de 2013). *Why probability contours for the multivariate Gaussian are elliptical?* Obtenido de Probability Basics: https://www.michaelchughes.com/blog/2013/01/why-contours-for-multivariate-gaussian-are-elliptical/

Congdon, P. (2006). *Bayesian Statistical Modelling* (Segunda ed.). West Sussex: John Wiley & Sons.

DeGroot, M., & Schervish, M. (2012). *Probability and Statistics.* Boston: Pearson Education.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Royal Statistic Society*, 1-38.

Dey, D., & Rao, C. (2005). *Bayesian Thinking, Modeling and Computation.* Amsterdam, Provincia de Holanda Septentrional, Países Bajos: Elsevier.

Durkheim, É. (2009). *Las Reglas del Método Sociológico y Otros Escritos* (Cuarta ed.). Madrid: Alianza Editorial.

Duval, L. (1 de Febrero de 2016). *What does prototype mean in clustering?* Obtenido de Signal Processing -Stack Exchange: https://dsp.stackexchange.com/questions/28593/what-does-prototype-mean-in-clustering

Duval, L. (4 de Septiembre de 2016). *Who first defined the "equal-delta" or "delta over equal" (≜) symbol?* Obtenido de History of Science and Mathematics: https://hsm.stackexchange.com/questions/5166/who-first-defined-the-equal-delta-or-delta-over-equal-triangleq-symbol

Ellison, H. (14 de Noviembre de 2020). *Mad moments, or first verse attempts by a born natural, etc.* Obtenido de https://books.google.co.cr/books?id=nzRcAAAAcAAJ&redir_esc=y

Frost, J. (9 de Noviembre de 2020). *Fitted Values*. Obtenido de Statistics By Jim: https://statisticsbyjim.com/glossary/fitted-values/#:~:text=A%20fitted%20value%20is%20a,the%20fitted%20value%20is%2020.

Ganegedara, T. (13 de Noviembre de 2020). *Intuitive Guide to Understanding Decision Trees*. Obtenido de Towards Data Science: https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-understanding-decision-trees-adb2165ccab7

Gibbons, R. (1992). *Un Primer Curso de Teoría de Juegos.* Barcelona: Antoni Bosch.

Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., & Caihuelas Quiles, R. (2017). *Minería de Datos. Modelos y Algoritmos.* Barcelona: Editorial UOC.

Google Developers. (10 de Febrero de 2020). *Conjuntos de entrenamiento y prueba: Separación de datos*. Obtenido de Curso Intensivo de Aprendizaje Automático: https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data?hl=es-419

Google Developers. (5 de Marzo de 2020). *Sobreajuste (overfitting)*. Obtenido de Curso intensivo de aprendizaje automático: https://developers.google.com/machine-learning/crash-course/glossary?hl=es-419#s

Guru99. (14 de Noviembre de 2020). *Supervised vs Unsupervised Learning: Key Differences*. Obtenido de Guru99: https://www.guru99.com/supervised-vs-unsupervised-learning.html

Hardy, M., & Bryman, A. (Edits.). (2009). *The Handbook of Data Analysis.* London: Sage Publications Ltd.

Havens, A. (25 de Febrero de 2019). *Limits and Continuity for Multivariate Functions .* Obtenido de Department of Mathematics. University of Massachusetts, Amhers: https://people.math.umass.edu/~havens/LimContBivar.pdf

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2005). *Introducción a la Minería de Datos.* Madrid: Pearson Educación.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods.* New York: Springer.

Hsu, H. (2011). *Signal and Systems* (Segunda ed.). New York: McGraw-Hill.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Introduction to Statistical Learning with Applications in R.* New York: Springer.

Jaynes, E. (2003). *Probability Theory. The Logic of Science.* Cambridge: Cambridge University Press.

Kiseliov, A., Krasnov, M., & Makarenki, G. (1973). *Problemas de Ecuaciones Diferenciales Ordinarias.* Moscú: Mir.

Kuhn, T. (2004). *La Estructura de las Revoluciones Científicas* (Octava ed.). Buenos Aires: Fondo de Cultura Económica.

Lathi, B. P., & Green, R. (2018). *Linear Systems and Signals* (Tercera ed.). New York: Oxford University Press.

Lehmann, E. (1959). *Testing Statistical Hypothesis.* Toronto: John Wiley and Son.

Liang, S. (Marzo de 2011). *Multivariable Epsilon-Delta Limit Definitions*. Obtenido de Wolfram Demonstrations Project: https://demonstrations.wolfram.com/MultivariableEpsilonDeltaLimitDefinitions/

Maklin, C. (15 de Julio de 2019). *Gaussian Mixture Models Clustering Algorithm Explained*. Obtenido de Towards Data Science: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

minerals; enterML; Lakshmi Prasad Y; Dynamic Stardust; Manju Savanth; Prhld. (4 de Junio de 2018). *What is the difference between model hyperparameters and model parameters?* Obtenido de Data Science - Stack Exchange:

https://datascience.stackexchange.com/questions/14187/what-is-the-difference-between-model-hyperparameters-and-model-parameters

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective* (Primera ed.). Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

Nabi, I. (2017). *Un Modesto Ensayo Sobre las Funciones de Probabilidad y la Esperanza Matemática en Espacios de Probabilidad.* Documento Inédito.

Nabi, I. (2020). *Algunas Reflexiones Sobre la Distribución Binomial Negativa II (Un Análisis Teórico y Aplicado).* Documento Inédito.

Nag, D. (16 de Octubre de 2016). *Machine Learning = Representation + Evaluation + Optimization*. Obtenido de Medium: https://medium.com/@devnag/machine-learning-representation-evaluation-optimization-fc7b26b38fdb

Ng, A. (16 de Noviembre de 2020). *What are hiperparameters*. Obtenido de Redes neuronales y aprendizaje profundo: https://www.coursera.org/lecture/neural-networks-deep-learning/parameters-vs-hyperparameters-TBvb5

Pan, W., Shen, X., & Liu, B. (2013). Cluster Analysis: Unsupervised Learning via Supervised Learning with a Non-convex Penalty. *Journal of Machine Learning Research*, 1865-1889.

Pang-Ning, T., Steinbach, M., & Kumar, V. (2014). *Introduction to Data Mining.* Essex: Pearson Education Limited.

Paul, S. (15 de Agosto de 2018). *Hyperparameter Optimization in Machine Learning Models*. Obtenido de datacamp: https://www.datacamp.com/community/tutorials/parameter-optimization-machine-learning-models?utm_source=adwords_ppc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=3326

PowerData. (30 de Diciembre de 2016). *Calidad de datos en minería de datos a través del preprocesamiento*. Obtenido de Data Quality: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/calidad-de-datos-en-mineria-de-datos-a-traves-del-preprocesamiento

Python Programming. (14 de Noviembre de 2020). Obtenido de Python Programming Tutorials: https://pythonprogramming.net/features-labels-machine-learning-tutorial/

Ramírez, D. (8 de Noviembre de 2017). *Planeamiento para la Investigación: Unidad de Posgrados FCA*. Obtenido de Epistemologías, Epistemia y Epistemología: Conceptos Empleados en la Investigación Social: http://sisbib.unmsm.edu.pe/bibvirtualdata/publicaciones/plan_invest/n1_2004/a03.pdf

Ranganathan, S., Nakain, K., Schönback, C., & Gribskov, M. (2019). *ENCYCLOPEDIA OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY* (Primera ed., Vol. I). (M. Cannataro, B. Gaeta, & M. Asif Khan, Edits.) Radarweg, Amsterdam, Netherlands: Elsevier.

Real Academia Española. (15 de Noviembre de 2020). *Genética*. Obtenido de Diccionario de la Lengua Española: https://dle.rae.es/gen%C3%A9tico#J4a2bti

Real Academia Española. (15 de Noviembre de 2020). *Huella*. Obtenido de Diccionario de la Lengua Española: https://dle.rae.es/huella?m=form

Real Academia Española. (13 de Noviembre de 2020). *informar*. Obtenido de Definición: https://dle.rae.es/informar?m=form

Roberts, S. J. (1997). Parametric and non-parametric unsupervised cluster analysis. *Patter Recognition*, 261-272.

Rosental, M. M., & Iudin, P. F. (1971). *DICCIONARIO FILOSÓFICO.* San Salvador: Tecolut.

Russell, K. (29 de Enero de 2014). *University of Manitoba*. Obtenido de Hypothesis testing: http://home.cc.umanitoba.ca/~krussll/stats/hypothesis-testing.html

Salian, I. (8 de Agosto de 2018). *SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?* Obtenido de NVIDIA Blog: https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/#:~:text=In%20a%20supervised%20learning%20model,and%20patterns%20on%20its%20own.

Schwarz, G. (1978). ESTIMATING THE DIMENSION OF A MODEL. *The Annals of Statistics, VI*(2), 461-464.

Scott, D. W. (2015). *Multivariate Density Estimation. Theory, Practice and Visualization.* New Jersey: John Wiley & Sons.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (Agosto de 2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal, 8*(1), 289-317. Recuperado el 16 de Noviembre de 2020, de https://journal.r-project.org/archive/2016/RJ-2016-021/RJ-2016-021.pdf

Serrano, L. G. (2020). *Grokking Machine Learning.* New York: Manning Publications. Obtenido de https://livebook.manning.com/book/grokking-machine-learning/2-1-what-is-the-difference-between-labelled-and-unlabelled-data-/v-4/35

Shashua, A. (23 de Abril de 2009). *Introduction to Machine Learning.* Obtenido de Hebrew University of Jerusalem: https://www.cs.huji.ac.il/~shashua/papers/class11-PAC2.pdf

Silverman, B. W. (1998). *Density Estimation for Statistics and Data Analysis.* Boca Raton: Chapman & Hall.

Steele, J. M. (21 de Junio de 2015). *Ah, but the Term "Shattered Set"...* Obtenido de Shattered Sets --- A Story of a Colorful Term: http://www-stat.wharton.upenn.edu/~steele/Rants/ShatteredSets.html

Steele, M. J. (Junio de 21 de 2015). *Combinatorial Entropy and Uniform Limit Laws.* Obtenido de A dissertation submitted to the department of Mathematics and the committee on graduate studies of Stanford Univeristy in partial fulfillment of the requirements for the degree of Doctor of Philosophy: http://www-stat.wharton.upenn.edu/~steele/Publications/PDF/SteeleThesis75.pdf

Stover, C. (23 de Octubre de 2020). *Concept*. Obtenido de Wolfram MathWorld: https://mathworld.wolfram.com/Concept.html

Stover, C. (23 de Octubre de 2020). *Shattered Set*. Obtenido de Wolfram MathWorld: https://mathworld.wolfram.com/ShatteredSet.html

Villegas Barahona, G. (2018). *Modelo estadístico pedagógico para la toma de decisiones administrativas y académicas con impacto en el mejoramiento continuo del rendimiento de los estudiantes universitarios, basado en los métodos de selección CUR.* Universidad de Salamanca, Departamento de Estadística. Salamanca: Universidad de Salamanca. Obtenido de https://gredos.usal.es/handle/10366/139405/statistics

Wang Ng, K., Tian, G.-L., & Tang, M.-L. (2011). *Dirichlet and Related Distributions* (Primera ed.). West Sussex: John Wiley & Sons.

Wikipedia. (25 de Abril de 2017). *Punto de Silla*. Obtenido de Wikipedia: https://upload.wikimedia.org/wikipedia/commons/1/1e/Saddle_point.svg

Wikipedia. (11 de Noviembre de 2020). *Absolute continuity*. Obtenido de Functional Analysis: https://en.wikipedia.org/wiki/Absolute_continuity

Wikipedia. (11 de Noviembre de 2020). *Análisis de componentes principales*. Obtenido de Aprendizaje Automático: https://es.wikipedia.org/wiki/An%C3%A1lisis_de_componentes_principales

Wikipedia. (11 de Noviembre de 2020). *Aprendizaje automático*. Obtenido de Inteligencia Artificial: https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico

Wikipedia. (8 de Noviembre de 2020). *Arity.* Obtenido de Abstract Algebra: https://en.wikipedia.org/wiki/Arity

Wikipedia. (30https://en.wikipedia.org/wiki/Bayes_estimator de Octubre de 2020). *Bayes Estimator*. Obtenido de Bayesian statistics.

Wikipedia. (30 de Octubre de 2020). *Bayesian Information Criterion*. Obtenido de Bayesian statistics: https://en.wikipedia.org/wiki/Bayesian_information_criterion

Wikipedia. (25 de Octubre de 2020). *Boolean function*. Obtenido de Mathematical logic: https://en.wikipedia.org/wiki/Boolean_function

Wikipedia. (10 de Septiembre de 2020). *Canonicalization*. Obtenido de Computing terminology: https://en.wikipedia.org/wiki/Canonicalization

Wikipedia. (13 de Noviembre de 2020). *Categorical distribution*. Obtenido de Categorical data: https://en.wikipedia.org/wiki/Categorical_distribution

Wikipedia. (9 de Noviembre de 2020). *Codificación de caracteres*. Obtenido de Lenguaje natural y Ciencias de la Computación: https://es.wikipedia.org/wiki/Codificaci%C3%B3n_de_caracteres

Wikipedia. (4 de Noviembre de 2020). *Curse of dimensionality*. Obtenido de Numerical Analysis: https://en.wikipedia.org/wiki/Curse_of_dimensionality

Wikipedia. (27 de Octubre de 2020). *Decision Rule*. Obtenido de Decision Theory: https://en.wikipedia.org/wiki/Decision_rule

Wikipedia. (2 de Noviembre de 2020). *Density Estimation*. Obtenido de Estimation of Densities: https://en.wikipedia.org/wiki/Density_estimation#:~:text=In%20probability%20and%20statistics%2C%20density,unobservable%20underlying%20probability%20density%20function.

Wikipedia. (10 de Septiembre de 2020). *Dirichlet distribution*. Obtenido de Multivariate continuous distributions: https://en.wikipedia.org/wiki/Dirichlet_distribution

Wikipedia. (10 de Noviembre de 2020). *Eigendecomposition of a matrix*. Obtenido de Matriz theory: https://en.wikipedia.org/wiki/Eigendecomposition_of_a_matrix

Wikipedia. (8 de Noviembre de 2020). *Estimation*. Obtenido de Estimation Theory: https://en.wikipedia.org/wiki/Estimation

Wikipedia. (8 de Noviembre de 2020). *Gradient*. Obtenido de Rate: https://en.wikipedia.org/wiki/Gradient

Wikipedia. (10 de Noviembre de 2020). *Ground Truth*. Obtenido de Optical character recognition: https://en.wikipedia.org/wiki/Ground_truth

Wikipedia. (11 de Noviembre de 2020). *Heuristic (computer science)*. Obtenido de Heuristic algorithms: https://en.wikipedia.org/wiki/Heuristic_(computer_science)

Wikipedia. (11 de Noviembre de 2020). *k-means clustering*. Obtenido de Cluster analysis algorithms: https://en.wikipedia.org/wiki/K-means_clustering

Wikipedia. (14 de Noviembre de 2020). *Law of Large Numbers.* Obtenido de Probability theorems: https://en.wikipedia.org/wiki/Law_of_large_numbers#:~:text=Uniform%20law%20of%20large%20numbers,-Suppose%20f(x&text=Then%20for%20any%20fixed%20%CE%B8,pointwise%20(in%20%CE%B8)%20convergence.

Wikipedia. (11 de Noviembre de 2020). *Límite de una función*. Obtenido de Análisis Real - Funciones: https://es.wikipedia.org/wiki/L%C3%ADmite_de_una_funci%C3%B3n

Wikipedia. (29 de Octubre de 2020). *Loss Function*. Obtenido de Statistical Inference: https://en.wikipedia.org/wiki/Loss_function

Wikipedia. (14 de Noviembre de 2020). *Máquinas de vectores de soporte*. Obtenido de Redes neuronales artificiales: https://es.wikipedia.org/wiki/M%C3%A1quinas_de_vectores_de_soporte

Wikipedia. (29 de Octubre de 2020). *Mean Squared Error*. Obtenido de Least Squares: https://en.wikipedia.org/wiki/Mean_squared_error

Wikipedia. (29 de Octubre de 2020). *Minimum Mean Squared Error*. Obtenido de Signal Estimation: https://en.wikipedia.org/wiki/Minimum_mean_square_error#Definition

Wikipedia. (29 de Septiembre de 2020). *Mixture model*. Obtenido de Cluster analysis: https://en.wikipedia.org/wiki/Mixture_model

Wikipedia. (16 de Noviembre de 2020). *Natural language*. Obtenido de Natural language processing: https://en.wikipedia.org/wiki/Natural_language

Wikipedia. (20 de Noviembre de 2020). *Natural language processing*. Obtenido de Computational linguistics: https://en.wikipedia.org/wiki/Natural_language_processing

Wikipedia. (10 de Noviembre de 2020). *Quantization*. Obtenido de Digital Signal Processing: https://en.wikipedia.org/wiki/Quantization_(signal_processing)

Wikipedia. (8 de Noviembre de 2020). *Sample (Statistics)*. Obtenido de Survey Methodology : https://en.wikipedia.org/wiki/Sample_(statistics)

Wikipedia. (7 de Noviembre de 2020). *Signal processing*. Obtenido de Telecommunication theory: https://en.wikipedia.org/wiki/Signal_processing

Wikipedia. (8 de Noviembre de 2020). *Statistical Classification*. Obtenido de Machine Learning: https://en.wikipedia.org/wiki/Statistical_classification

Wikipedia. (20 de Septiembre de 2020). *Statistics*. Obtenido de Kernel: https://en.wikipedia.org/wiki/Kernel_%28statistics%29

Wikipedia. (13 de Septiembre de 2020). *statistiques*. Obtenido de Noyau : https://fr.wikipedia.org/wiki/Noyau_(statistiques)

Wikipedia. (11 de Noviembre de 2020). *Vapnik–Chervonenkis dimension*. Obtenido de Dimension - Statistical classification: https://en.wikipedia.org/wiki/Vapnik%E2%80%93Chervonenkis_dimension

Wikipedia. (4 de Julio de 2020). *Vapnik–Chervonenkis theory*. Obtenido de Computational Learning Theory: https://en.wikipedia.org/wiki/Vapnik%E2%80%93Chervonenkis_theory

Wikipedia. (31 de Octubre de 2020). *Vector Quantization*. Obtenido de Lossy compression algorithms: https://en.wikipedia.org/wiki/Vector_quantization

Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford: Oxford University Press.

Yanchick, B. (13 de Noviembre de 2020). *An intuitive explanation of the decision tree algorithm*. Obtenido de Towards Data Science: https://towardsdatascience.com/building-an-intuition-for-the-decision-tree-algorithm-75e0786e86d

Zhang, K., Kwok, J. T., & Parvin, B. (6 de Julio de 2009). *Prototype Vector Machine for Large Scale Semi-Supervised Learning*. Obtenido de Departmentof Computer Science and Engineering - Hong Kong University of Science and Technology : https://www.cse.ust.hk/~jamesk/papers/icml09.pdf

Zhou, V. (10 de Noviembre de 2020). *A Simple Explanation of Gini Impurity*. Obtenido de Blog: https://victorzhou.com/blog/gini-impurity/

Википедия. (13 de Septiembre de 2020). *статистика*. Obtenido de Ядро: https://ru.wikipedia.org/wiki/%D0%AF%D0%B4%D1%80%D0%BE_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)