TEORÍA DEL APRENDIZAJE ESTADÍSTICO

ISADORE NABI

I. Definición General de Aprendizaje Estadístico

Según (James, Witten, Hastie, & Tibshirani, 2013, pág. 17), en esencia, el aprendizaje estadístico hace referencia a un conjunto de abordajes para predecir una función f a partir de conjuntos de variables dependientes e independientes. Complementariamente, se señala en (Hastie, Tibshirani, & Friedman, 2017, págs. xi-xii) que esto desempeña en la actualidad un papel fundamental en áreas como la agricultura, la industria, el almacenamiento de datos (que dio origen a la minería de datos), en la bioinformática, en la medicina y en muchos otros campos del conocimiento humano, en donde los modelos estadísticos son utilizados exclusivamente con fines predictivos partiendo de un determinado conjunto de datos, con independencia de otros factores. A esto se le conoce como *aprender de los datos*. De la evolución de los procesos de aprendizaje de datos surge el campo multidisciplinario conocido como *Aprendizaje Automático* (*Machine Learning*, en inglés). Así, los objetivos de este campo consisten en clasificar y predecir conjuntos de datos, para lo cual utilizan el marco teórico de la estadística matemática.

Como señalan (StackExchange Data Science, 2016) y (StackOverFlow, 2014), en general, la estadística se preocupa más por inferir parámetros (lo que implica validar que estadísticos muestrales se corresponden con sus versiones poblaciones), mientras que, en el aprendizaje automático, la predicción y la clasificación son el objetivo final.

Con respecto a la predicción, las ciencias de la estadística y el aprendizaje automático comenzaron a resolver casi el mismo problema desde diferentes perspectivas. Básicamente, la estadística asume que los datos fueron producidos por un determinado modelo estocástico. Así, desde una perspectiva estadística, se asume un modelo y, dados varios supuestos, se tratan los errores y se infieren los parámetros del modelo y otras cuestiones.

El aprendizaje automático nace bajo una visión informática de la manipulación de los datos. Por ello, los modelos son algorítmicos y, por lo general, se requieren muy pocas suposiciones con respecto a los datos. Es por ello que, como se adelantó, también usa, al igual que la Estadística, las herramientas del análisis funcional, como por ejemplo al construir los espacios de hipótesis (en el mismo sentido en que fueron planteados por Jerzey Neyman y Egon Pearson), así como también al hablar del sesgo de aprendizaje (conocido también como sesgo de inducción, sesgo de aprendizaje automático o sesgo de inteligencia artificial).

Lo anterior ha contribuido en buena medida a que, a pesar de que las dos ciencias no parecieran terminar de converger en términos gnoseológicos por su diferente aparentemente diferente genética filosófica (el espíritu conceptual bajo el cual nacieron), metodológicamente cada vez existe una mayor convergencia, expresada en que ambas comparten cada vez mayor cantidad de conocimientos y técnicas comunes. Existe por supuesto una base material a este hecho, la cual radica en que en que los problemas que enfrentaban tenían en común que podían ser resueltos mediante la determinación de tal o cual patrón geométrico del conjunto de datos (como se adelantó al introducir el marco teórico de los GLM), sin embargo, en los albores del aprendizaje automático esta compatibilidad de instrumentos no fue tan marcada, como ahora que el aprendizaje automático tiende a abordarse cada vez más desde una perspectiva estadística. Complementariamente, el aprendizaje se clasifica en aprendizaje supervisado, aprendizaje no supervisado, aprendizaje en línea y aprendizaje por refuerzo. Ejemplos de aprendizaje no supervisado incluyen el análisis de agrupaciones y asociaciones.

II. Positivos y Negativos en la Predicción / Clasificación

En el contexto de los modelos de variables dicotómicas, conocidos también como modelos de respuesta binaria, existen dos tipos de positivo y dos tipos de negativo.

- *Verdaderos Positivos:* cuando el valor de las observaciones es “Sí” y el valor predicho por el modelo es “Sí”.
- *Falso Positivo:* cuando el valor de las observaciones es “No” y el valor predicho por el modelo es “Sí”.

- *Verdadero Negativo:* cuando el valor de las observaciones es "No" y el valor predicho por el modelo es "No".
- *Falso Negativo:* cuando el valor de las observaciones es "Sí" y el valor predicho por el modelo es "No".

Véase el siguiente ejemplo en el contexto de la detección de tumores benignos o malignos:

Verdadero positivo (VP): • Realidad: Maligno • Predicción del modelo de AA: Maligno • **Número de resultados de VP: 1**	**Falso positivo (FP):** • Realidad: Benigno • Predicción del modelo de AA: Maligno • **Número de resultados de FP: 1**
Falso negativo (FN): • Realidad: Maligno • Predicción del modelo de AA: Benigno • **Número de resultados de FN: 8**	**Verdadero negativo (VN):** • Realidad: Benigno • Predicción del modelo de AA: Benigno • **Número de resultados de VN: 90**

Fuente: (Google Developers, 2021).

III. Matriz de Confusión

Como se señala en (James, Witten, Hastie, & Tibshirani, 2013, pág. 145), una matriz de confusión compara las predicciones del modelo de aprendizaje estadístico seleccionado con los verdaderos valores contenidos en las observaciones de entrenamiento del conjunto de datos (*default data set*). Los elementos en la diagonal de la matriz representan observaciones cuyos valores fueron correctamente predichos/clasificados por el modelo, mientras que fuera de la diagonal de la matriz se encuentran aquellos valores que fueron inadecuadamente predichos/clasificados. A continuación, se presenta una matriz de confusión para el caso de un análisis de discriminante lineal.

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

TABLE 4.4. *A confusion matrix compares the LDA predictions to the true default statuses for the* 10,000 *training observations in the* `Default` *data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for* 23 *individuals who did not default and for* 252 *individuals who did default.*

Fuente: (James, Witten, Hastie, & Tibshirani, 2013, pág. 145).

Finalmente, cabe decir, con base en (Barrios, 2019), que una matriz de confusión es esencialmente una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado.

IV. Exactitud, Tasa de Error, Sensibilidad, Especificidad, Precisión y Predicción Negativa del Modelo de Aprendizaje

IV.I. Exactitud

Como se señala en (Google Developers, 2021), la exactitud es una métrica para evaluar modelos de clasificación. Informalmente, la exactitud es la fracción de predicciones que el modelo realizó correctamente. Formalmente, la exactitud tiene la siguiente:

$$Exactitud = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones}$$

IV. II. Tasa de Error del Entrenamiento

Como se señala en (James, Witten, Hastie, & Tibshirani, 2013, pág. 37), el abordaje más común para estimar la precisión del modelo de entrenamiento $\hat{f}$ es el coeficiente conocido como *tasa de error del entrenamiento*, equivalente al cociente entre los errores cometidos si se aplica $\hat{f}$ a las observaciones de entrenamiento.

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

En donde $\hat{y}_i$ es la etiqueta de clase predicha (el valor de la variable dependiente pronosticado) para la $i - ésima$ observación usando $\hat{f}$, mientras que y_i es su valor real (el contenido en el conjunto de datos de entrenamiento). En la expresión anterior, $I(y_i \neq \hat{y}_i)$ es la función indicatriz que toma el valor 1 cuando la predicción es incorrecta y toma el valor 0 cuando la predicción es correcta. En términos de positivos y negativos lo anterior equivale a decir:

$$\frac{Falsos\ positivos\ +\ Falsos\ negativos}{Total\ de\ predicciones}$$

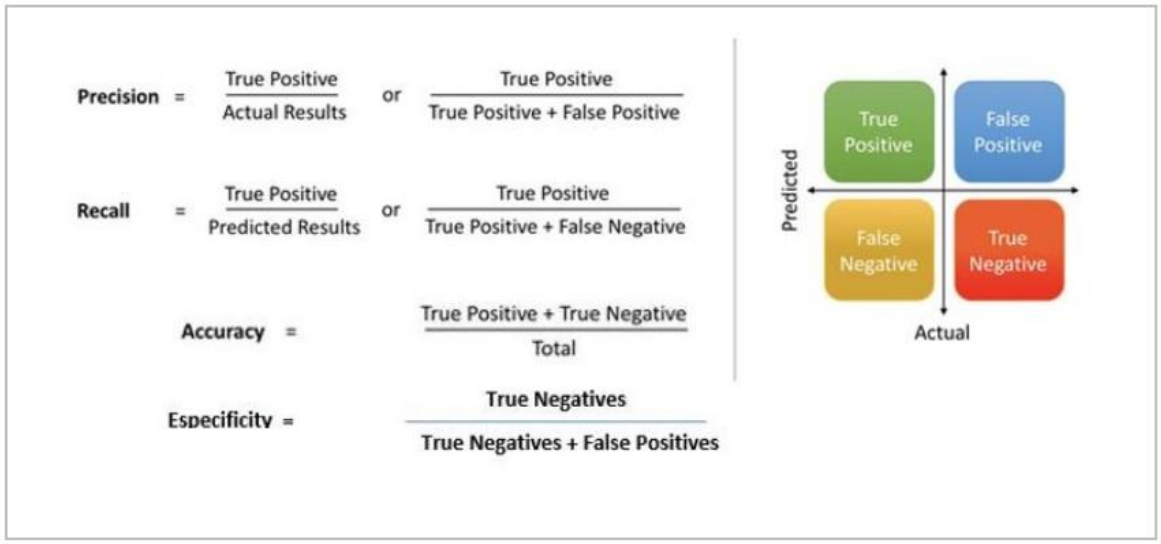
IV. III. Sensibilidad y Especificidad

Como señala (James, Witten, Hastie, & Tibshirani, 2013, pág. 145), el rendimiento específico de la clase también es importante en medicina y biología, donde los términos *sensibilidad* y *especificidad* caracterizan el desempeño de sensibilidad y especificidad un clasificador o prueba de detección. La sensibilidad es el porcentaje de verdaderos positivos que se identifican en relación a los positivos reales totales. Por su parte, la especificidad es el porcentaje de verdaderos negativos que se identifican correctamente en relación a los negativos reales totales.

IV. IV. Precisión (Valor Predictivo Positivo) y Valor Predictivo Negativo

Como señala (Barrios, 2019), el concepto de *precisión* refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión, mayor la precisión. Se representa por la proporción de verdaderos positivos dividido entre todos los resultados positivos (tanto verdaderos positivos, como falsos positivos). Este concepto, como se señala en (Wikipedia, 2021), es conocido a menudo como *valor predictivo positivo*. Su equivalente inverso es el concepto de *valor predictivo negativo*, estimado de forma equivalente, aunque considerando los negativos.

IV. V. Sumario



Fuente: (Barrios, 2019).

Matriz de Confusion		Predicho			
		Negativo	Positivo		
Real	Negativo	a	b	*Verdadero Negativo (True negative rate)*	a/(a+b)
	Positivo	c	d	*Exactitud*	d/(c+d)
		Sensibilidad	*Especifidad*		
		d/(d+c)	a/(a+b)	Precisión=(a+d)/(a+b+c+d)	

Figura 3: Matriz de confusión con otras métricas de evaluación.

a: es el número de predicciones correctas de clase negativa (negativos reales)

b: es el número de predicciones incorrectas de clase positiva (falsos positivos)

c: es el número de predicciones incorrectas de clase negativa (falsos negativos)

d: es el número de predicciones correctas de clase positiva (positivos reales)

Fuente: (Barrios, 2019).

V. Modelos Lineales Generalizados desde la Teoría del Aprendizaje Estadístico

Como señalan (StackExchange Data Science, 2016) y (StackOverFlow, 2014), los modelos lineales generalizados son un desarrollo estadístico. Sin embargo, los nuevos tratamientos bayesianos ponen este algoritmo también en el campo de juego del aprendizaje automático. Entonces creo que ambas afirmaciones podrían ser correctas, ya que la interpretación y el tratamiento de cómo funciona podrían ser diferentes.

La distinción sutil entre modelos estadísticos y modelos de aprendizaje automático es que, en los modelos estadísticos, usted decide explícitamente la estructura de la ecuación de salida antes de construir el modelo. El modelo está construido para calcular los parámetros/coeficientes.

Tómense precisamente los GLM, que son modelos estadísticos y, por consiguiente, útiles para verificar que los modelos estadísticos y las técnicas de aprendizaje automático no son mutuamente excluyentes.

$$y = a_1x_1 + a_2x_2 + a_3x_3$$

Las variables independientes son x_1, x_2 y x_3, mientras que los coeficientes a determinar son a_1, a_2 y a_3. Así, se define la estructura de su ecuación de esta manera antes de construir el modelo y calcule a_1, a_2 y a_3. Si se cree que y está correlacionada de alguna manera con x_2 de forma no lineal, puede probarse una transformación como la siguiente:

$$y = a_1x_1 + a_2(x_2)^2 + a_3x_3$$

Evidentemente, la transformación anterior implica imponer una restricción en términos de la estructura de salida. En el caso de los modelos de aprendizaje automático, rara vez se especifica la estructura de salida y los algoritmos, como los árboles de decisión, son intrínsecamente no lineales y funcionan de manera eficiente. Simplemente se parte de un conjunto de datos con una variable dependiente conocida (etiqueta), se "entrena el modelo" su modelo y luego se aplica al conjunto de datos para intenta predecir un número real, como por ejemplo el precio de una casa[1].

Específicamente, los GLM, así como cualquier metodología estadística de regresión pertenece al aprendizaje supervisado, por cuanto los datos que tiene incluyen tanto la entrada como la salida, por ponerlo en algunos términos. Entonces, por ejemplo, si se tiene un conjunto de datos para, supóngase, las ventas de automóviles en un concesionario. Se tiene, para cada coche, características como marca, modelo, precio, color, descuento, etc., pero también se tiene el número de ventas de cada coche. Si esta tarea no estuviera supervisada, se tendría un conjunto de datos que incluye, tal vez, solo la marca, el modelo, el precio, el color, etc. (no el número real de ventas) y lo mejor que se puede hacer es agrupar los datos. El ejemplo no es perfecto, pero tiene como objetivo transmitir el panorama general. Una buena pregunta que debe hacerse al decidir si un método está supervisado o no es preguntarse "¿Se cuenta con alguna forma de juzgar la calidad de una entrada?". Si se cuenta con datos de regresión lineal, la respuesta es afirmativa. Simplemente se evalúa el valor de la función (en este caso, la función lineal) de los datos de entrada

[1] En este sentido, una aplicación industrial exitosa de los GLM puede encontrarse en http://www.kdd.org/kdd2016/papers/files/adf0562-zhangA.pdf y contribuir a explicar por qué los modelos lineales generalizados son considerados por muchos como una técnica del aprendizaje automático, aun cuando en términos históricos y teóricos no lo son. Se afirma lo anterior puesto que, por ejemplo, muchos afirman que la regresión logística no es en realidad una regresión, lo que justifican planteando que, por lo general, solo se usa para la predicción binaria, lo que es idónea para tareas de clasificación. Por supuesto, estas creencias son refutadas por los mismos orígenes históricos y teóricos de los GLM, los cuales han sido estudiados en esta investigación. Por supuesto, de forma artificial puede concebirse en el contexto del aprendizaje automático como un método de clasificación, aunque durante el entrenamiento lo que hace es predecir si un valor pertenece a una clasificación o no, lo que prueba en última instancia muestra cómo los orígenes antes referidos determinan la naturaleza del método estadístico.

para estimar la salida. No es así en el otro caso. También se supervisa la regresión logística.

VI. Referencias

Barrios, J. (19 de Julio de 2019). *La matriz de confusión y sus métricas* . Obtenido de Health BIG DATA: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Google Developers. (19 de Julio de 2021). *Clasificación: Exactitud*. Obtenido de https://developers.google.com/machine-learning/crash-course/classification/accuracy

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (Segunda ed.). New York: Springer.

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

StackExchange Data Science. (19 de Junio de 2016). *Is GLM a statistical or machine learning model?* Obtenido de https://datascience.stackexchange.com/questions/488/is-glm-a-statistical-or-machine-learning-model

StackOverFlow. (15 de Marzo de 2014). *Supervised Learning, Unsupervised Learning, Regression*. Obtenido de https://stackoverflow.com/questions/22419136/supervised-learning-unsupervised-learning-regression

Wikipedia. (10 de Julio de 2021). *Precision and recall*. Obtenido de Bioinformatics: https://en.wikipedia.org/wiki/Precision_and_recall