Práctica 2

Luis Fernando Lomelín Ibarra
a01177015@itesm.mx
Tecnológico de Monterrey
Ingeniería en tecnologías computacionales
Monterrey, NL, México

David Alejandro Martínez Tristán
a01610267@itesm.mx
Tecnológico de Monterrey
Ingeniería en tecnologías computacionales
Monterrey, NL, México

ABSTRACT

En el presente documento se exponen los resultados de utilizar la técnica de regresión logística para predecir la clasificación de una entidad con varios atributos entre dos posibles categorías.

La práctica utilizó dos datasets, cada uno de 10,000 instancias y varios atributos. El primero es de usuarios de tarjetas de crédito que, dependiendo de si son estudiantes, su saldo promedio y su ingreso, se predice si cumplirá con sus deudas. El segundo es de personas que, según su altura y su peso, se predice si es hombre o mujer.

Se elaboró un script de Python que lee los datos de entrada y utiliza dos modelos diferentes: uno provisto por la librería Scikit-Learn y otro implementado manualmente que aplica gradiente descendiente.

Respecto al dataset de deudas de tarjeta de crédito, el modelo entrenado con la librería de Scikit-Learn obtuvo una baja precisión de 0.22034 y su matriz de confusión demuestra que clasifica correctamente a quienes cumplen sus deudas. Por otra parte, el modelo de gradiente descendiente obtuvo una precisión mucho menor de 0.06790 y su matriz de confusión que no es precisa para clasificar a quienes incumplen sus deudas.

Respecto al dataset de género, el modelo entrenado con la librería de Scikit-Learn obtuvo una alta precisión de 0.91318 para identificar el género y su matriz de confusión indica que es preciso tanto para hombres como para mujeres. Por otra parte, el modelo de gradiente descendiente obtuvo una precisión aún mayor de 0.95812 y su matriz de confusión indica que es preciso para clasificar a mujeres.

1. INTRODUCCIÓN

Durante la tercera semana del curso, se presentaron dos nuevas técnicas para entrenar un modelo de aprendizaje automático: la **regresión lineal con múltiples salidas** y la **regresión logística**.

La regresión lineal con múltiples salidas es utilizada cuando se desea predecir una matriz de salidas a partir de una matriz de atributos. Consiste en el proceso de realizar una regresión lineal de forma iterativa.

La regresión logística es utilizada para predecir la probabilidad de que una salida pertenezca a una de dos categorías en particular. Aplica una función sigmoide que permita mapear todos los valores de entrada en un valor de salida que se encuentre en el rango [0, 1]. De esta manera, los valores de entrada se pueden clasificar entre dos clases dependiendo del rango en el que se encuentre su valor de salida [0, 0.5) o [0.5, 1].

2. CONCEPTOS PREVIOS

- Regresión multivariable
- Regresión logística
- Función sigmoide
- Gradiente descendiente
- Matriz de confusión
- Tasa de aprendizaje
- Criterio de paro

3. METODOLOGÍA

La práctica utiliza un script elaborado con Python y librerías externas. Se leen los datos de dos datasets de 10,000 instancias cada uno, cuyo objetivo es predecir una variable de salida a partir de diferentes variables de entrada. Para poder utilizar mejor la regresión logística primero se normalizan los datos de cada dataset. Después, cada dataset es partido de manera aleatoria en una proporción 80%-20% para generar el training set y el test set.

Posteriormente, se entrena un primer modelo utilizando la regresión logística de Scikit-Learn y luego con un segundo modelo utilizando el gradiente descendiente con función sigmoide, el cual fue programado manualmente.

Finalmente se calcula la tasa de precisión y la matriz de confusión de cada uno. Para el caso del gradiente descendiente, se reporta también el vector beta obtenido, el criterio de paro y la tasa de aprendizaje alfa empleada.

3.1 Dataset de deudas de tarjeta de crédito

Conjunto de 10,000 instancias cuyo objetivo es predecir cuáles clientes van a incumplir con la deuda de su tarjeta de crédito. Utiliza las variables descritas a continuación:

- student: indica si el cliente es un estudiante
- balance: saldo promedio del cliente
- income: ingreso promedio del cliente

3.2 Dataset de identificación de género

Conjunto de 10,000 instancias cuyo objetivo es predecir el género de una persona a partir de su peso y altura. Utiliza las variables descritas a continuación:

- height: altura de la persona
- weight: peso de la persona

4. RESULTADOS

4.1 Dataset de deudas de tarjeta de crédito

Después de entrenar a los modelos se obtuvieron los siguientes resultados con los set de pruebas. Con el modelo de Scikit Learn se obtuvo una precisión de 0.22034. Esto nos quiere decir que de los no tiene una buena precisión para detectar cuando una persona va a incumplir con una tarjeta. Para analizar más a detalle los resultados se generó la siguiente matriz de confusión.

Real/Predecido	Cumple	Incumple
Cumple	1936	46
Incumple	5	13

Como se puede analizar en la matriz de confusión, el modelo clasifica correctamente a los que cumplen pero no es muy bueno clasificando a los que incumplen.

Usando el mismo set de datos de prueba se obtuvieron los resultados del modelo de gradiente descendiente. La precisión que obtuvo este modelo fue de 0.06790. Este modelo a comparación del de scikit learn es aún más impreciso. Para poder tener un mejor entendimiento de los resultados de este modelo también se generó una matriz de confusión.

Real/Predecido	Cumple	Incumple
Cumple	1929	55
Incumple	12	4

Como se puede apreciar en esta matriz, este modelo de gradiente descendiente no es muy bueno identificando qué persona va a incumplir con la tarjeta de crédito. Esta falla tal vez se pueda mejorar ajustando la tasa de crecimiento del modelo.

Como ambos tienen una precisión baja, tal vez valga la pena buscar más datos para ayudar a la precisión ó tal vez convenga utilizar otro modelo completamente diferente.

4.2 Dataset de identificación de género

Los resultados de los modelos para identificar el género de una persona a partir de su peso y alturas. Para evaluar esto se utilizó las herramientas de la tasa de precisión y la matriz de confusión de Scikit Learn.

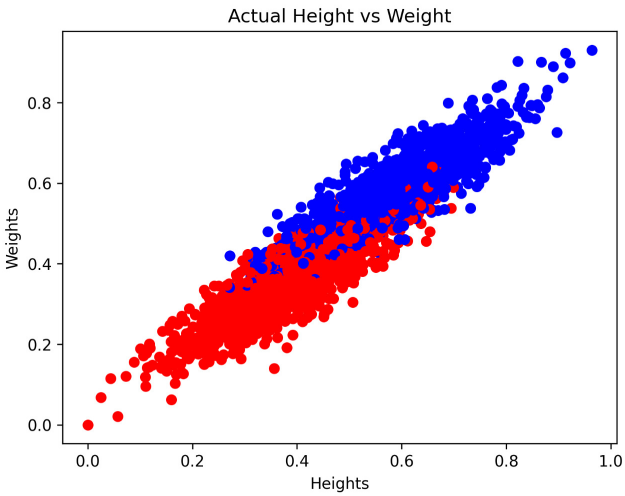

Figura: Relación real peso altura de mujeres (Rojo) y Hombres (Azul)

Con el modelo generado utilizando la librería de Scikit Learn, obtuvo una precisión de 0.91318. Esto indica que el proceso tiene una buena precisión a la hora de identificar el género. Esto se puede apreciar comparando la relación peso-altura de los datos contra los datos predichos.

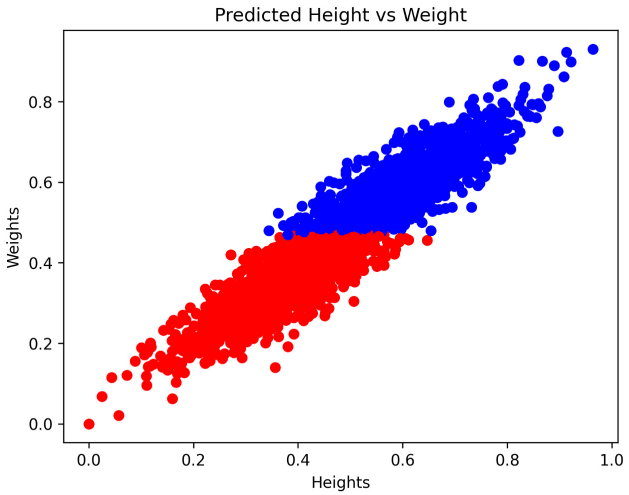

Figura: Relación peso altura de mujeres (Rojo) y Hombres (Azul) predicha por el modelo de Scikit Learn

Para tener una mejor idea de las predicciones se generó la siguiente matriz de confusión:

Real/ Predecido	Hombre	Mujer
Hombre	922	85
Mujer	99	894

Como se puede apreciar en la tabla, el modelo puede identificar de los 2000 sujetos del set de pruebas, el modelo identifica correctamente a 922 hombre, a 894 Mujeres y erra en identificar a 99 mujeres y a 85 hombres.

Con el modelo de gradiente descendiente se obtuvo una precisión de 0.95812. Como se puede observar el modelo de gradiente descendiente es más preciso en este dataset que el de scikit learn. Esto se puede deber a el parámetro de tasa de aprendizaje de 0.009.

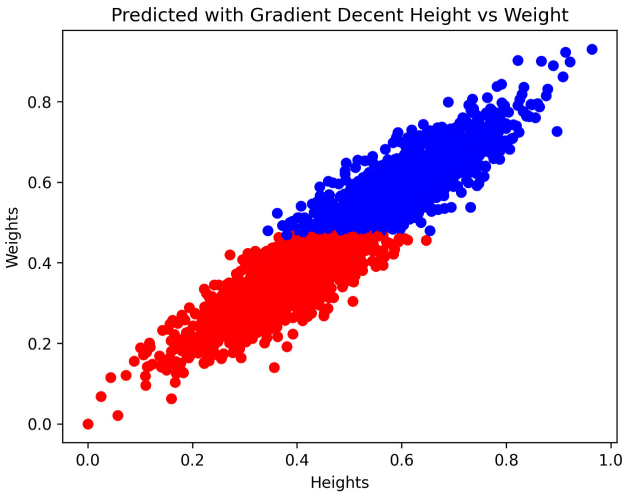

Figura: Relación peso altura de mujeres (Rojo) y Hombres (Azul) predicha por el modelo de Gradiente Decendiente,

Similarmente se generó a partir del modelo la siguiente matriz de confusión.

Real/ Predecido	Hombre	Mujer
Hombre	15	41
Mujer	1006	938

Analizando la matriz de confusión podemos ver porque se dio el incremento de precisión. Como para calcular la precisión se toma en consideración los Verdaderos positivos y los falsos positivos, si lo comparamos con el otro modelo da una mejor precisión. Pero esto no necesariamente significa que el modelo es bueno prediciendo. Se puede decir con buena seguridad que puede identificar a mujeres con mayor facilidad que a hombres, pero no puede identificar tan bien a hombres a partir de su peso y de su altura.

5. CONCLUSIONES Y REFLEXIONES

David:
La presente práctica abrió mi panorama acerca de las predicciones utilizando técnicas de regresión. La regresión multi-salida para predecir una matriz de salidas, mientras que la regresión logística para clasificar entre dos categorías. Además aprendí la importancia de contar con métricas para realizar un análisis más detallado de los resultados obtenidos.
Elaborar este ejercicio abrió mi perspectiva y estimuló mi imaginación acerca de las posibilidades que tienen estos

modelos para resolver problemáticas de interés actual, por lo que estoy entusiasmado de poner a prueba los nuevos conocimientos fuera del salón. No obstante, cabe recordar que se tratan de predicciones que pueden acertar, pero también pueden fallar.

Luis:

Fue una práctica bastante interesante. Ya con la experiencia de la práctica anterior fue mucho más sencillo formar la implementación del código, pero donde se puso más interesante fue en aprender sobre la precisión y la matriz de confusión. De ambos realmente sabía muy poco y fue bastante iluminador aprender de ellos. Me ayudó bastante a entender mucho mejor los resultados de los modelos y de la investigación que hice me hizo entender mejor la importancia de tener múltiples métricas para analizar los datos para realmente comprender los resultados que está sacando el modelo y poder juzgar a base de ellos si realmente el modelo está funcionando como uno quiere.

REFERENCIAS

Long, A., 2018. *Understanding Data Science Classification Metrics in Scikit-Learn in Python*. [online] towards data science. Available at: <https://towardsdatascience.com/understanding-data-science-classification-metrics-in-scikit-learn-in-python-3bc336865019> [Accessed 30 August 2021].

aprendeAI. 2021. *Métricas de Evaluación Clasificación con Scikit Learn*. [online] Available at: <https://aprendeia.com/metricas-de-evaluacion-clasificacion-con-scikit-learn-machine-learning/> [Accessed 30 August 2021].

Scikit-learn.org. 2021. *sklearn.metrics.confusion_matrix — scikit-learn 0.24.2 documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html> [Accessed 30 August 2021].

Narkhede, S., 2021. *Understanding Confusion Matrix*. [online] towards data science. Available at: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62> [Accessed 30 August 2021].

Sharma, P., 2021. *Decoding the Confusion Matrix*. [online] towards data science. Available at: <https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb> [Accessed 30 August 2021].