Práctica no. 2: Regresión logística

Jair Antonio Bautista Loranca
a01365850@itesm.mx
Tecnológico de Monterrey
Monterrey, N.L., México

Maximiliano Zambada Camacho
a01570146@itesm.mx
Tecnológico de Monterrey
Monterrey, N.L., México

RESUMEN

. La regresión logística es una técnica de Machine Learning para clasificar los registros de un conjunto de datos. En este tipo de regresión, se usa una o más variables independientes para predecir un resultado, es decir la variable dependiente. La técnica es análoga a la regresión lineal, pero intenta predecir un campo objetivo categórico en lugar de uno numérico.

. En la regresión lineal, se intenta predecir variables de valor continuo, mientras que en regresión logística se predicen variables binarias (Si/No, Verdadero/Falso). En esta última, las variables dependientes deben ser continuas, por lo que si son categóricas hay que transformarlas en algún valor continuo. La regresión logística puede ser usada tanto para clasificación binaria como para multiclase.

ACM Reference Format:
Jair Antonio Bautista Loranca and Maximiliano Zambada Camacho. 2021. Práctica no. 2: Regresión logística. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1. INTRODUCCIÓN

Después de haber realizado los análisis de las diferentes bases de datos con regresión lineal, pudimos ver la facilidad de realizar estos estudios. Al observar las gráficas generadas, así como los coeficientes que se consiguieron después del análisis, y compararlos con los obtenidos en el análisis con Gradiente Descendiente, encontramos que fue más preciso con las funciones de SKLearn. Ahora para la implementación del análisis de regresión logística esperamos que se obtengan resultados precisos, y poder acercarnos lo más posible con nuestra Gradiente Descendente.

. Para esto implementaremos primeramente la regresión logística en los datos de los archivos que nos fueron proporcionados, para más adelante implementar nuestra versión de Gradiente Descendiente y llegar a un análisis similar.

2. CONCEPTOS PREVIOS

. Si una variable cualitativa con dos niveles se codifica como 1 y 0, matemáticamente es posible ajustar un modelo de regresión lineal por mínimos cuadrados $\beta_0 + \vec{\beta}X$. El problema de esta aproximación es que, al tratarse de una recta, para valores extremos del predictor, se obtienen valores de Y menores que 0 o mayores que 1, lo que entra en contradicción con el hecho de que las probabilidades siempre están dentro del rango [0,1].

. Para evitar estos problemas, la regresión logística transforma el valor devuelto por la regresión lineal ($\beta_0 + \vec{\beta}X$) empleando una función cuyo resultado está siempre comprendido entre 0 y 1. Existen varias funciones que cumplen esta descripción, una de las más utilizadas es la función logística (también conocida como función sigmoide):

.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

. Para valores de x muy grandes positivos, el valor de e^{-x} es aproximadamente 0, por lo que el valor de la función sigmoide es 1. Para valores de x muy grandes negativos, el valor e^{-x} tiende a infinito por lo que el valor de la función sigmoide es 0.

. Sustituyendo la x de la ecuación 1 por la función lineal ($\beta_0 + \beta_1 X$) se obtiene que:

$$P(Y = k|X = x) = \frac{1}{1 + e^{-\beta_0 + \beta_1 X}} = \frac{\beta_0 + \beta_1 X}{1 + e^{\beta_0 + \beta_1 X}}$$

. donde $P(Y = k|X = x)$ puede interpretarse como: la probabilidad de que la variable cualitativa Y adquiera el valor k (el nivel de referencia, codificado como 1), dado que el predictor X tiene el valor x.

. Para el entrenamiento requerimos un conjunto de entrenamiento. Debido a que cada entrada a nuestro modelo consiste en un vector $\vec{x}$ entonces podemos formar el conjunto de entrenamiento como la matriz X de dimensiones $n \times m$. Donde cada renglón corresponde a un vector de entrada para el modelo. Además requerimos del vector $\vec{y}$ que corresponde a los resultados históricos que tenemos del modelo.

$$X = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}, \ \vec{y} = \begin{pmatrix} \vec{y}_1 \\ \vec{y}_2 \\ \vdots \\ \vec{y}_n \end{pmatrix}$$

. Para entrenar nuestro modelo con los datos que tenemos necesitamos una función de costo que determina que tan aproximado es nuestro a los datos que tenemos. Usualmente se utiliza la función de error cuadrático medio que para nuestro modelo quedaría de la siguiente manera.

$$\text{MSE}(\vec{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left(h(\vec{\beta}, \vec{x}_i) - \vec{y} \right)^2$$

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

. Existe una forma cerrada para minimizar el error cuadrático medio para nuestro modelo, sin embargo no es computacionalmente viable. Por ello recurrimos al método de gradiente para encontrar el $\vec{\beta}$ que minimice el error para nuestros datos.

. Este método consiste en lo siguiente, dado que el costo está en función del vector $\vec{\beta}$. Entonces podemos utiliza el gradiente negativo para encontrar la dirección a donde se minimiza la función. Por lo tanto podemos minimizar el costo por medio del siguiente algoritmo.

$$\nabla J(\vec{\beta}_j) = X^T(\vec{\mu} - \vec{y})$$

. En este algoritmo, el i-ésimo elemento del vector $\vec{\mu}$, es $\mu_i = \frac{1}{(1+e^{-\vec{\beta}_j^T \cdot \vec{x}_i}}$

3. METODOLOGÍA

La metodología a seguir durante el proceso de desarrollo de la práctica fue una metodología colaborativa de igual manera que el resto de las prácticas. Ambos integrantes trabajamos de manera paralela en los pasos de la práctica para poder crear los códigos necesarios. Desde el análisis en regresión logística de los datasets de Default y el de género, se fue realizando en un proceso de prueba y error.

. Para el análisis que se realizó con regresión logística, fue un proceso bastante sencillo y fluido, al ya tener conocimiento del uso de las funciones de las librerías utilizadas. Con la práctica conseguida en el primer laboratorio, fue bastante sencillo, utilizando el mismo proceso para generar las gráficas, y ahora las matrices de confusión y análisis de Clasificación.

. Se decidió que para la creación de las matrices, así como del reporte de clasificación, se vería mejor visualmente, por lo que se realizó la gráfica de las mismas.

. Finalmente, ya que se tiene esto realizado, lo único que queda pendiente es realizar la gráfica para género, para la cual lo que se hizo fue primeramente, graficar utilizando el set de Train de las variables, para ver cómo se distribuyen dentro del plano 2D. Después de esto, se procedió a realizar ahora la gráfica para el test de Train junto con las predicciones conseguidas con la regresión logística.

. Como último paso, para el código de la Gradiente Descendente, se siguieron los siguientes pasos. Primeramente, así como para la práctica pasada, fue un proceso donde se tuvo que ir implementando poco a poco las funciones para poder conseguir el algoritmo con la fórmula aplicada correctamente.

4. RESULTADOS

4.1. Análisis de Regresión Logística

Por medio de los resultados obtenidos en el análisis de regresión logística, pudimos conseguir las gráficas necesarias para poder observar el comportamiento de las variables al realizar las predicciones, así como su precisión:

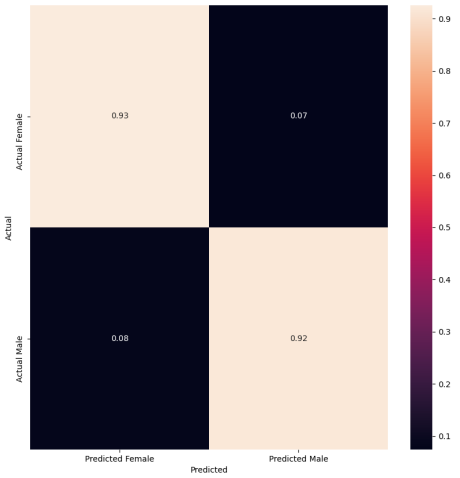

Figura 1: `genero.txt`
Confusion Matrix de Género por Regresión Logística

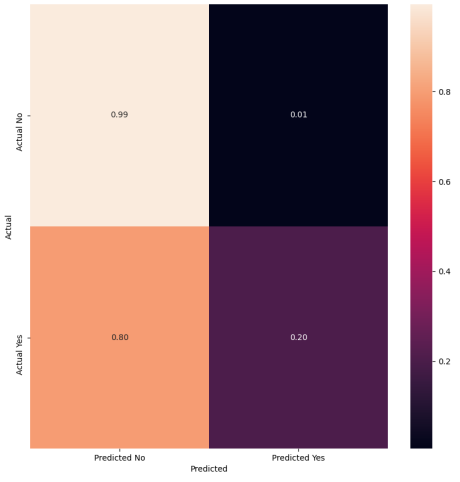

Figura 2: `Default.txt`
Confusion Matrix de Default por Regresión Logística

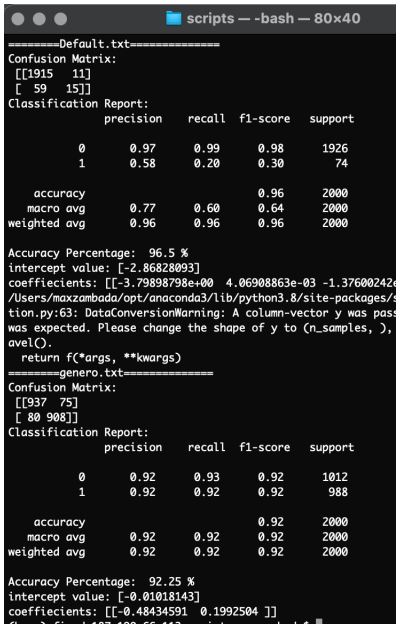

Figura 3: `Análisis de Precisión`
Análisis de ambos datasets con su reporte de clasificación y porcentaje de precisión

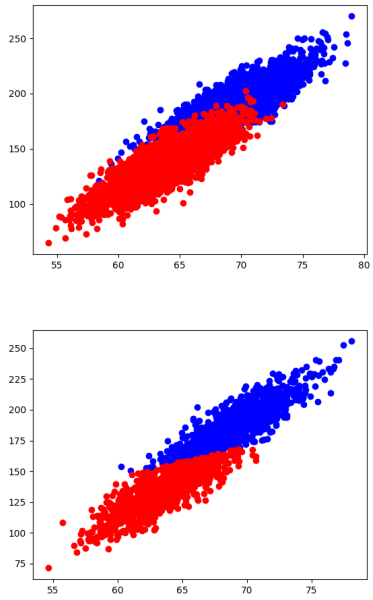

Figura 4: `Gráfica de Regresión Logística de Género`
Comparación del comportamiento de los atributos del dataset contra los conseguidos con la predicción de Regresión Logística

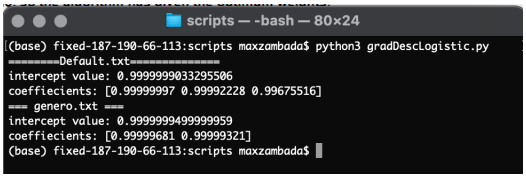

Figura 5: `Análisis por Gradiente Descendente`
Análisis de ambos datasets desde cero, obteniendo el intercepto y los coeficientes correspondientes

. Como podemos ver al analizar nuestros datos, se puede apreciar que se tiene una precisión muy alta, que va de 96.5 % en el dataset de Default, y de 92.25 % en el de género. Esto nos indica que la precisión obtenida al utilizar este modelo de regresión logística es muy alto, consiguiendo predicciones correctas de acuerdo a los datos proporcionados.

. Asimismo, se puede ver en la graficación de Scatter de género cómo se parecen demasiado los valores del dataset con los valores que se predicen al utilizar el modelo de análisis.

. Ahora bien, comparando los resultados obtenidos por la regresión logística y por nuestra implementación del algoritmo de descenso de gradiente, podemos ver que nuestros resultados no son tan precisos como los obtenidos con SKLearn. Consideramos que esto se debe a que nuestra regresión se estanca en un mínimo local a la hora de realizar el análisis de los datos, ya que los coeficientes y el intercepto conseguidos no se acercan a los valores esperados que se consiguieron con la regresión logística implementada con SKLearn, como se puede ver en la figura 3 y la figura 5.

5. CONCLUSIONES Y REFLEXIONES

Por lo discutido en el documento podemos ver que la regresión logística es de igual manera una herramienta muy importante para la creación de modelos de predicción, consiguiendo una precisión muy elevada. Al implementar esta regresión logística manualmente por medio del Descenso de Gradiente, pudimos observar el grado de dificultad y las distintas formulaciones necesarias para conseguir los resultados. Al final, al ser un algoritmo creado desde cero, hay muchos factores que afectan el resultado final.

Jair Antonio. Dados los resultados obtenidos puedo darme cuenta de la utilidad de algorimtos como decenso de gradiente. Debido a que nos permite explorar el espacio de soluciones y de esta manera podemos encontrar una solución adecuada que minimice el error para nuestro modelo. A su vez he logrado comprender el funcionamiento del modelo de regresión logísitca, el cual se trata de un modelo de clasificación que trabaja sobre el modelo generado por la regresión lineal.

Maximiliano Zambada. Al estar implementando el algoritmo desde cero con gradiente descendiente, pude observar todos los procesos que se llevan a cabo con la simple instrucción de SKLearn de LogisticRegression. Los resultados obtenidos me pudieron dar una imagen de cómo se puede predecir de manera efectiva los datos de un dataset, y la utilidad de este modelo más adelante con problemas de predicción.