Regresión Logística con el Método de Descenso del Gradiente Estocástico

Adrián Rodríguez, Santiago Villarreal, Luis Carlos Bernal, Emiliano Ramírez, Armando Apellaniz

ITAM

12 de mayo de 2021

Regresión logística

En estadística, una regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables explicativas.

Las regresiones logísticas se utilizan comúnmente en en las áreas de ciencias médicas y sociales.

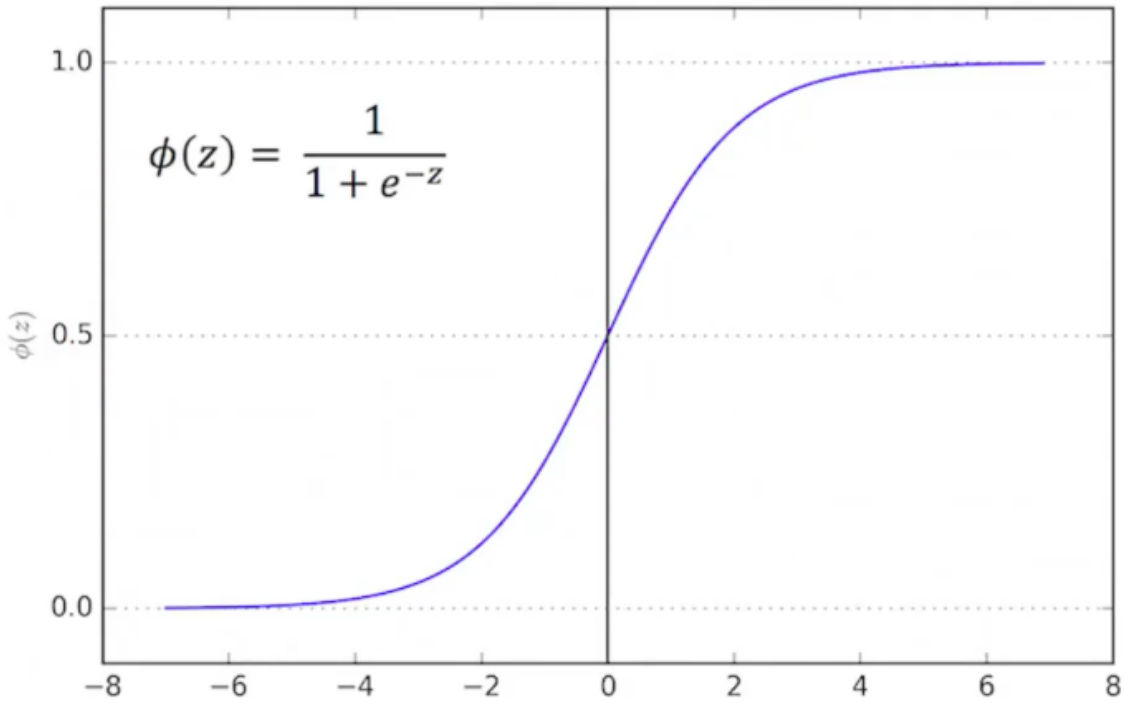
En el presente trabajo exhibiremos una forma de implementar la regresión logística con el método de Descenso del Gradiente Estocástico en python.

¿Cómo se realiza la regresión logística?

La regresión logística utiliza una función logística para asignar a qué categoría pertenece un objeto a partir de cierto valor. Formalmente, la función estima la siguiente probabilidad (donde C es la categoría de una observación y X el vector de datos):

$$P(C|X) = \phi(\beta^t x) \tag{1}$$

Donde β es el vector de coeficientes y ϕ es la función logística.



Ecuación del modelo

Como el modelo busca realizar una clasificación binaria, la distribución de probabilidad indicada para describir estos fenómenos es la Bernoulli. La relación lineal que propicia el modelo es:

$$ln(\frac{p}{1-p}) = \beta^t x \quad \Longrightarrow \quad \frac{p}{1-p} = e^{\beta^t x} \tag{2}$$

Para el caso particular de este problema, la ecuación para la predicción queda dada por:

$$\hat{y} = \frac{1}{1 + e^{-\beta^t x}}$$

Donde:
$\hat{y}$ es el valor estimado que puede ser 0 o 1 (redondeado)

Descenso de gradiente

Descenso de Gradiente es un algoritmo para encontrar el mínimo de una función de clase $\mathcal{C}^2$ siguiendo la dirección de su gradiente.

Para poder implementarlo, es necesario tener una expresión analítica de la función f, así como de su derivada ∇f. En cada iteración, se verifica el tamaño del paso (dado por $\alpha > 0$) para optimizar el método y generar el menor costo computacional. El paso está dado por $x_{k+1} = x_k + \alpha P_k$.

Descenso de gradiente aplicado a regresión logística

El método de descenso de gradiente aplicado a una regresión logística, permite que en cada iteración, el vector β de coeficientes se actualice. Sea $p = \phi(\beta^t x)$. Entonces por definición de la distribución Bernoulli:

$$\begin{cases} ln(p) & y = 1 \\ ln(1-p) & y = 0 \end{cases}$$

El método de gradiente estocástico actualiza los coeficientes después de cada iteración utilizando la siguiente ecuación:

$$\beta = \beta + l_r(y - \hat{y})\hat{y}(1 - \hat{y})x$$

Donde:
β es el vector de coeficientes.
l_r es un parámetro que el usuario configura manualmente, el cual limita la corrección de cada coeficiente cada vez que se actualiza.
$y - \hat{y}$ es el error de la predicción.

Análisis de datos

A continuación se hará un análisis de predicción en tres bases de datos. Recordar que, por la definición de regresión logística, dicha predicción solamente puede tomar los valores 0 o 1. La interpretación de estos valores va a variar en cada base de datos.

Para la aplicación del algoritmo en los datos establecimos el cross-validation de 5-folds con un tamaño de paso o learning rate de 0.1 y un número de exposición del algoritmo con las bases de entrenamiento de 100.

Se recopilará la información necesaria para aplicar la regresión logística con descenso de gradiente estocástico. Esto permitirá estimar los coeficientes y realizar la predicción sobre la variable objetivo en cuestión.

Primero veremos una breve explicación del código realizado.

Descripción del código

i) Se desarrollaron funciones para leer los datos y darles el formato necesario.

ii) Se realizó una función para hacer predicciones, se utilizará para evaluar los posibles candidatos para coeficientes del modelo.

iii) Se hizo una función para estimar los coeficientes de la regresión logística usando el método de gradiente de descenso estocástico.

iv) Una función implementará la validación cruzada, esto es, construiremos y evaluaremos k veces el modelo y estimaremos el desempeño como forma de "entrenarlo" y "probarlo" para darle una mayor precisión.

v) Implementamos el algoritmo de regresión logística para las distintas bases de datos.

Bloque de código de la función de Descenso del Gradiente Estocástico

```python
# Estima los coeficientes de la regresión logit utilizando el método del gradiente estocástico
def coefs_sgd(train, step, num_epoch):
    coef = [0.0 for i in range(len(train[0]))]
    #epoch es el número de veces que nuestro algoritmo correra sobre las bases de entrenamiento para actualizar los coeficientes
    for epoch in range(num_epoch):
        for fila in train:
            #generamos las predicciones usando nuestra base de entrenamiento
            yhat = predict(fila, coef)
            #obtenemos el error de nuestra predicción
            error = fila[-1] - yhat
            #actualizamos nuestro coeficiente del intercepto condicionado al tamaño de paso que establecimos arbitrariamente
            coef[0] = coef[0] + step * error * yhat * (1.0 - yhat)
            #actualizamos nuestros coeficientes condicionado al tamaño de paso que establecimos arbitrariamente
            for i in range(len(fila)-1):
                coef[i + 1] = coef[i + 1] + step * error * yhat * (1.0 - yhat) * fila[i]
    #regresamos coeficientes actualizados por el entrenamiento
    return coef
```

Figure: Función de Gradiente Estocástico de nuestro algorítmo

Implementación del código

A continuación daremos una breve descripción de cada base de datos y del código aplicado a dichos datos; y presentaremos los resultados obtenidos.

Clasificación de tumores de mama

Utilizamos la base de datos de la Universidad de Wisconsin (descargada a través de Kaggle). En la base, además del diagnóstico del tumor, encontramos variables de tamaño, de forma y de simetría del tumor. Entre varias de las 32 variables encontramos altos niveles de correlación, que podían representar un problema para el proyecto, por lo cual corrimos un algoritmo Lasso en ellas, para reducir la dimensionalidad de la base de datos y nos quedamos con las variables:

- Promedio de puntos cóncavos en el tumor
- Medición más grande del radio del tumor
- La desviación estándar de los valores en la escala de grises (textura).
- Medida de uniformidad del tumor
- Cantidad de puntos cóncavos en el tumor
- Medida de simetría del tumor

Clasificación de tumores de mama: Resultados y Conclusiones

Al aplicarle el algoritmo de regresión logística estimando los coeficientes con gradiente estocástico, los resultados de precisión de predicción que obtuvimos fueron alentadores:

$$\text{Scores} = \left[97.35\%, 94.69\%, 94.69\%, 96.43\%, 97.35\%\right]$$

Lo cual nos da una precisión promedio de predicción de 96.11%.

Esto es importante, porque nos indica que el cancer de mama es diagnosticable sin la necesidad de un procedimiento quirúrgico, con la ayuda de algoritmos de clasificación de aprendizaje de máquina.

Educación Superior

Aplicamos nuestro modelo para analizar la inscripción a la educación superior por parte de niños y niñas que fueron seguidas a través del tiempo en la República Mexicana por un periodo de diez años. Esta encuesta se llama Mexican Family Live Sourvey y fue realizada por el CIDE en conjunto con la Universidad de Duke.
Variables Independientes:

- número de hermanos
- edad de la persona en la última encuesta
- logaritmo del ingreso mensual
- escolaridad de ambos padres

Dependiente: 1 si la persona tiene escolaridad superior, 0 si no la tiene.
Objetivo:
Determinar si una persona tuvo acceso a educación superior con base en las variables dependientes.

Educación Superior: Resultados

Como ya mencionamos, el método de gradiente estocástico utiliza subconjuntos aleatorios de k datos de nuestra base completa para el entrenamiento del algoritmo y actualización de los coeficientes. En este caso, dividimos nuestros datos en 5 bloques del mismo tamaño. Para esta base de datos, obtuvimos los siguientes resultados:

$$\text{Scores} = \left[80.95\%, 73.46\%, 78.9\%, 79.59\%, 78.91\%\right]$$

Precisión Promedio$= 78.36\%$ entre los datos reales y los estimados

Es decir, basándonos en las variables mencionadas, podremos determinar aproximadamente si $\frac{4}{5}$ niños y niñas recibieron educación superior.

Educación Superior: Conclusiones

Este resultado es de bastante utilidad, porque, por ejemplo, en México es difícil tener acceso a información sobre escuelas que se encuentran en áreas rurales.

Los resultados anteriores nos indica que los controles usados son buenas variables predictoras. Es por eso que la rama del Desarrollo Económico se ha dedicado a estudiar a fondo dichas variables y sus efectos en resultados deseables para la sociedad como lo es la inscripción a la educación superior.

Para enriquecer nuestros resultados, opinamos que se podría, agregar variables de tratamiento como el Estado de la República donde se encuentra la vivienda, medio de transporte de la familia o número de televisores o dispositivos con acceso a internet.

Diagnóstico de Diabetes en mujeres de la población indígena Pima

Finalmente, utilizamos una base de datos del National Institute of Diabetes and Digestive and Kidney Diseases que recopila la información fisiológica de mujeres de al menos 21 años de linaje de la población Pima para predecir, para este sector específico, si tienen diabetes o no. Las variables que se usaron como controles son las siguientes.

- Número de embarazos
- Concentración de la glucosa en la plasma
- Presión arterial
- Grosor del tejido epidérmico del triceps
- Secreción de insulina en un lapso de dos horas
- Índice de masa corporal
- Función Diabetes Pedigree
- Edad

Diagnóstico de Diabetes: Resultados y conclusiones

Al aplicarle el algoritmo de regresión logística estimando los coeficientes con gradiente estocástico, los resultados de precisión de predicción que obtuvimos fueron buenos:

$$\text{Scores} = \left[79.73\%, 75.81\%, 73.85\%, 78.43\%, 77.77\%\right]$$

Lo cual nos da una precisión promedio de predicción de 77.12%.

Esta es una buena precisión de predicción, sin embargo creemos que no es tan buena como la de la base de tumores ("comparamos" los resultados de estas dos bases ya que las dos son bases con variables de diagnóstico clínico) pues en la base de tumores hicimos la selección de variables para predicción con el data-driven LASSO que es un método de selección de variables recursivo que se adapta a los datos y pudimos obtener los controles más *ad hoc* para una predicción mientras que en la base de diabetes no filtramos los controles.

Referencias

- How To Implement Logistic Regression, Jason Brownlee, from his blog *Python from Scratch*

  `https://machinelearningmastery.com/implement-logistic-regression-stochastic-gradient-descent-scratch-python/: :text=Gradient%20Descent%20is%20the%20process,downhill%20towards%20the%20minimum%20value`