U N I V E R S I D A D D E G U A D A L A J A R A

CENTRO UNIVERSITARIO DE CIENCIAS EXACTAS E INGENIERÍA
DEPARTAMENTO DE CIENCIAS COMPUTACIONALES

Seminario de Solución de Problemas de Sistemas Basados en Conocimiento

Práctica No. 8
Regresión Logística

Nombre: Hurtado González Edgar Arturo
Código: 212597894

Introducción

La regresión logística es un método estadístico utilizado para modelar la relación entre una variable dependiente categórica y una o más variables independientes. Aunque su nombre incluye "regresión", en realidad se usa para problemas de clasificación.

En lugar de predecir valores continuos como en la regresión lineal, la regresión logística estima la probabilidad de que una observación pertenezca a una categoría específica. Utiliza la función logística para transformar la salida de una combinación lineal de variables independientes a un valor entre 0 y 1, que representa la probabilidad.

Se basa en el concepto de odds ratio (razón de probabilidades) para modelar la relación entre las variables independientes y la probabilidad de que ocurra un evento.

La interpretación se hace a través de coeficientes, que indican el cambio en el logaritmo de la odds ratio por cada unidad de cambio en la variable independiente, manteniendo las demás constantes.

Es ampliamente utilizado en diversas áreas, como ciencias sociales, medicina, biología y negocios, para problemas de clasificación binaria o multinomial, como la predicción de enfermedades, análisis de riesgos crediticios, entre otros.

La regresión logística es una técnica fundamental en estadística y aprendizaje automático, especialmente en problemas de clasificación. Aquí hay más detalles:

Función Logística:

La función logística es el núcleo de la regresión logística. Transforma la combinación lineal de las variables independientes usando la fórmula:

$$P\,(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}}$$

Donde:

- $P\,(Y = 1 \mid X)$ es la probabilidad condicional de que la variable dependiente Y sea 1 dado un conjunto de variables independientes X.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes que la regresión logística estima.
- $X_1, X_2, \ldots, X_n$ son las variables independientes.

Coeficientes y Odds Ratio:

Los coeficientes de la regresión logística se utilizan para interpretar la influencia de las variables independientes en la variable dependiente. Un coeficiente positivo indica que a medida que la variable independiente aumenta, la probabilidad de pertenecer a la categoría 1 también aumenta, y viceversa.

La interpretación se realiza a través de odds ratio. Por ejemplo, si el odds ratio asociado a una variable es 2, significa que por cada unidad de aumento en esa variable, la odds (la probabilidad de éxito dividida por la probabilidad de fracaso) se duplica.

Entrenamiento y Evaluación:

El modelo de regresión logística se entrena utilizando técnicas como la maximización de la verosimilitud, que busca encontrar los coeficientes que mejor se ajusten a los datos observados.

Se evalúa mediante métricas como precisión, sensibilidad, especificidad, área bajo la curva ROC (Receiver Operating Characteristic), entre otras, dependiendo del contexto y los requisitos del problema.

Regularización y Variaciones:

A veces, la regresión logística puede beneficiarse de técnicas de regularización como la penalización L1 (Lasso) o L2 (Ridge) para evitar el sobreajuste y mejorar la generalización del modelo.

Existen variaciones de la regresión logística, como la regresión logística multinomial para problemas de clasificación con más de dos categorías, y la regresión logística ordinal para variables dependientes ordenadas.

Aplicaciones:

La regresión logística se aplica en una amplia gama de campos, como medicina para predecir enfermedades, en marketing para la segmentación de clientes, en finanzas para el análisis de riesgos crediticios y en la industria para el control de calidad, entre otros.

Desarrollo

El código realiza varias tareas relacionadas con la regresión logística y la evaluación del modelo. Aquí está la descripción de la lógica del código:

1. *Importación de bibliotecas y definición de funciones:*
   - Importa las bibliotecas necesarias como pandas, numpy, matplotlib, sklearn, y seaborn.
   - Define una función sigmoid(x) que calcula la función sigmoidal.
   - Define funciones para entrenar un modelo de regresión logística, calcular gradientes, hacer predicciones y calcular la matriz de confusión manualmente.

2. *Preprocesamiento de datos:*
   - Lee los datos desde un archivo CSV llamado 'Employee.csv' usando Pandas.
   - Realiza un etiquetado de las columnas 'Education', 'City', 'Gender', y 'EverBenched' usando LabelEncoder.
   - Extrae las características ('X') y las etiquetas ('Y') del DataFrame.
   - Escala las características usando StandardScaler.
   - Divide los datos en conjuntos de entrenamiento y prueba (80% y 20% respectivamente).

3. *Entrenamiento del modelo:*
   - Entrena un modelo de regresión logística utilizando los datos de entrenamiento.
   - Utiliza el descenso de gradiente para optimizar los coeficientes del modelo.

4. *Predicción y evaluación del modelo:*
   - Realiza predicciones en el conjunto de prueba utilizando los coeficientes aprendidos.
   - Calcula el error cuadrático medio (MSE) entre las etiquetas reales y las predichas.

5. *Matriz de Confusión:*
   - Define una función para calcular la matriz de confusión manualmente.
   - Calcula la matriz de confusión para comparar las predicciones del modelo con las etiquetas reales en el conjunto de prueba.

6. *Visualización de la Matriz de Confusión:*
   - Utiliza seaborn y matplotlib para visualizar la matriz de confusión en un mapa de calor.

7. *Resultados:*
   - Imprime los coeficientes aprendidos, el MSE y muestra una comparación entre valores reales y predichos para las primeras 10 muestras del conjunto de prueba.
   - Muestra la matriz de confusión generada y también imprime las primeras filas del DataFrame.

Conclusión

En esencia, la regresión logística calcula la probabilidad de que la variable dependiente sea igual a 1 (o pertenezca a la categoría de interés) dadas las variables independientes. La salida de la regresión logística está en el rango de 0 a 1, y se interpreta como la probabilidad de pertenecer a una categoría en particular.

Se utiliza en una amplia gama de campos, como la medicina (para predecir la presencia o ausencia de una enfermedad), la industria (para prever el éxito o fracaso de un producto), la investigación social, entre otros, donde se necesita predecir un resultado binario basado en variables explicativas.

El código realiza un proceso de regresión logística para predecir la variable de salida "LeaveOrNot" a partir de las características del conjunto de datos 'Employee.csv' y evalúa el rendimiento del modelo utilizando la matriz de confusión y el error cuadrático medio.

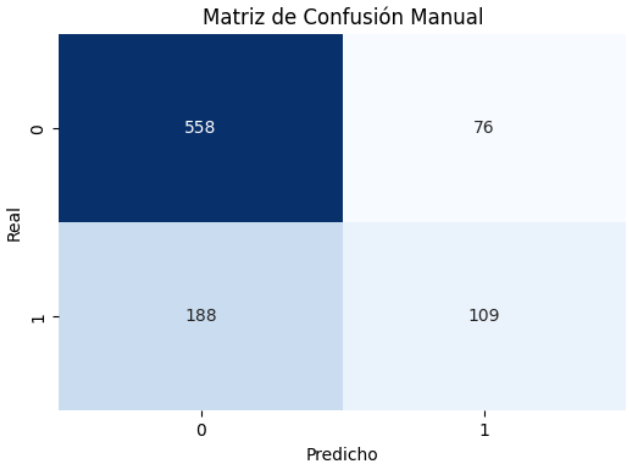
Resultados

Coeficientes: [(-0.6489875) (0.0655483) (0.31292859) (0.28764311) (-0.06062245) (-0.36172419) (0.17749828) (-0.02859144) (-0.22062267)]

MSE: 0.28356605800214824

Valores Predichos: [1 1 1 0 1 1 1 0 0 0]

Valores Reales: [1 1 1 1 0 0 1 1 0 1]



Datos cargados correctamente dentro del código:

	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
0	0	2017	0	3	34	1	0	0	0
1	0	2013	2	1	28	0	0	3	1
2	0	2014	1	3	38	0	0	2	0
3	1	2016	0	3	27	1	0	5	1
4	1	2017	2	3	24	1	1	2	1

Referencias:

¿Qué es la regresión logística? - Explicación del modelo de regresión Logística - AWS. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/logistic-regression/

Torres, L. (2021, 29 enero). ¿En qué consiste la regresión logística? ¿Qué es la regularización? The Machine Learners. https://www.themachinelearners.com/regresion-logistica-regularizacion/

T-Test, Chi-Square, ANOVA, Regression, Correlation. . . (s. f.). https://datatab.es/tutorial/logistic-regression