Guía de estudio - Regresión Logística

¡Hola! Te damos la bienvenida a esta nueva guía de estudio.

¿En qué consiste esta guía?

La siguiente guía de estudio tiene como objetivo recordar y repasar los contenidos vistos en clase. Esta guía de repaso busca proporcionarte más elementos para comprender y aplicar la regresión logística en problemas de clasificación. A medida que profundices en estos conceptos, podrás abordar problemas más complejos y aplicaciones del mundo real.

¡Vamos con todo!

Tabla de contenidos

¡Comencemos!

Introducción a la Regresión Logística

La regresión logística es una técnica estadística utilizada para modelar y analizar la relación entre una variable dependiente binaria (que tiene dos categorías o resultados posibles, generalmente etiquetados como 0 y 1) y una o más variables independientes. A diferencia de la regresión lineal, que se utiliza para predecir valores numéricos continuos, la regresión logística se emplea para predecir la probabilidad de que una observación pertenezca a una de las dos categorías.

En la regresión logística, se utiliza una función sigmoide (curva S) para transformar la combinación lineal de las variables independientes en un valor que varía entre 0 y 1. Esto representa la probabilidad de que la variable dependiente sea igual a 1. Los coeficientes de regresión se utilizan para ponderar la contribución de cada variable independiente en el modelo.

¿Por qué es importante la regresión logística?

La regresión logística es una técnica importante en el campo de la inteligencia artificial y el machine learning (AI/ML). Los modelos ML son programas de software que puede entrenar para realizar tareas complejas de procesamiento de datos sin intervención humana. Los modelos de ML creados mediante regresión logística ayudan a las organizaciones a obtener información procesable a partir de sus datos empresariales. Pueden usar esta información para el análisis predictivo a fin de reducir los costos operativos, aumentar la eficiencia y escalar más rápido. Por ejemplo, las empresas pueden descubrir patrones que mejoran la retención de los empleados o conducen a un diseño de productos más rentable.

A continuación, enumeramos algunos beneficios del uso de la regresión logística en comparación con otras técnicas de ML.

Simplicidad

Los modelos de regresión logística son matemáticamente menos complejos que otros métodos de ML. Por lo tanto, pueden implementarse incluso si nadie en un equipo tiene una profunda experiencia en ML.

Velocidad

Los modelos de regresión logística pueden procesar grandes volúmenes de datos a alta velocidad porque requieren menos capacidad computacional, como memoria y potencia de procesamiento. Esto los hace ideales para que las organizaciones que están empezando con proyectos de ML obtengan ganancias rápidas.

Flexibilidad

Puede usar la regresión logística para encontrar respuestas a preguntas que tienen dos o más resultados finitos. También puede usarlo para preprocesar datos. Por ejemplo, puede ordenar los datos con un amplio rango de valores, como las transacciones bancarias, en un rango de valores más pequeño y finito mediante la regresión logística. A continuación, puede procesar este conjunto de datos más pequeño mediante el uso de otras técnicas de ML para obtener un análisis más preciso.

Visibilidad

El análisis de regresión logística ofrece a los desarrolladores una mayor visibilidad de los procesos de software internos que otras técnicas de análisis de datos. La solución de problemas y la corrección de errores también son más fáciles porque los cálculos son menos complejos.

Regresión lineal y regresión logística.

Las principales diferencias entre la regresión lineal y la regresión logística se encuentran en el tipo de variable dependiente, la forma de la relación funcional, el objetivo del modelo, la interpretación de los coeficientes y las métricas de evaluación utilizadas.

	Regresión lineal	Regresión logística
Tipo de variable dependiente	Numérica y continua, como la temperatura, el ingreso o la edad.	Categórica o binaria, como sí/no, 0/1 o verdadero/falso.
Función de relación	Utiliza una relación lineal entre las variables independientes y la variable dependiente.	Utiliza una curva sigmoide (forma de "S") para modelar la relación entre las variables independientes y la probabilidad de pertenecer a una categoría.
Objetivo	Predecir valores numéricos y medir la relación lineal entre las variables.	Determinar la probabilidad de pertenecer a una categoría.
Coeficientes de regresión	Indican el cambio en la variable dependiente por cada unidad de cambio en la variable independiente correspondiente.	Indican el cambio en el logaritmo de las probabilidades por cada unidad de cambio en la variable independiente correspondiente.
Evaluación	Se utilizan métricas como el error cuadrático medio (MSE) o	Se utilizan métricas como la precisión, la sensibilidad, la

	el coeficiente de determinación (R^2) para medir la precisión de las predicciones numéricas.	especificidad, el F1-score y la curva ROC-AUC para medir la capacidad de clasificación y la calidad del ajuste.

Variables en la Regresión Logística

En la regresión logística, se trabajan con dos tipos principales de variables: la variable dependiente (también llamada variable de respuesta) y las variables independientes (predictoras o explicativas). A continuación, se describen estos tipos de variables:

Variable Dependiente (Variable de Respuesta):

- **Binaria:** En la mayoría de los casos, la variable dependiente en la regresión logística es binaria, lo que significa que solo tiene dos categorías o resultados posibles. Ejemplos comunes incluyen sí/no, 0/1, aprobado/reprobado, verdadero/falso, etc.
- **Ordinal**: En algunos casos, la variable dependiente puede ser ordinal, lo que significa que tiene categorías ordenadas pero no necesariamente equidistantes. Por ejemplo, una variable de satisfacción del cliente con categorías como "insatisfecho", "neutral" y "satisfecho" podría ser una variable ordinal. Aunque no es el enfoque principal de la regresión logística, existen técnicas relacionadas para manejar este tipo de datos.

Variables Independientes (Predictoras o Explicativas):

- **Categóricas:** Las variables independientes pueden ser categóricas, es decir, variables que representan categorías o grupos. Ejemplos incluyen el género (hombre/mujer), la ciudad de residencia, la categoría de producto, etc. En la regresión logística, estas variables suelen codificarse mediante técnicas como one-hot encoding.
- **Numéricas:** También se pueden tener variables independientes numéricas, que representan cantidades o medidas continuas. Ejemplos de variables numéricas incluyen la edad de una persona, el ingreso anual, la temperatura, etc.
- **Interacciones:** En algunos casos, se pueden incluir interacciones entre variables independientes para capturar relaciones complejas. Por ejemplo, la interacción entre el género y la edad en un modelo de predicción de compras en línea.

Función sigmoide

La función sigmoide, también conocida como función logística, es una función matemática que se utiliza en la regresión logística y otros contextos para transformar valores continuos en un rango entre 0 y 1. La función sigmoide toma la forma de una curva "S" y su fórmula matemática es:

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

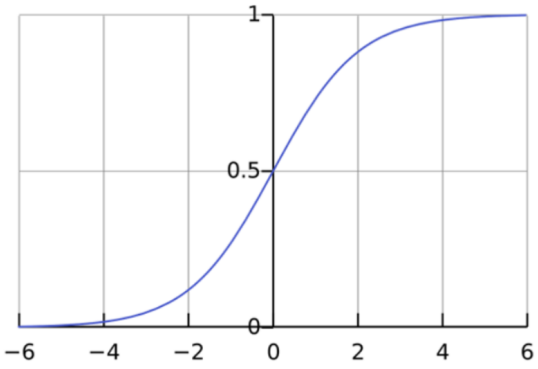

Figura 1: Gráfico de la función sigmoide

Fuente: Desafío Latam

La función sigmoide tiene varias propiedades importantes:

- **Rango Limitado:** La función sigmoide produce valores en el rango de 0 a 1. Esto la hace especialmente útil para modelar probabilidades, ya que los valores por encima de 0.5 pueden interpretarse como una probabilidad mayoritaria de que ocurra un evento (clase 1), mientras que los valores por debajo de 0.5 indican una probabilidad mayoritaria de que no ocurra (clase 0).

- **Suavidad:** La función sigmoide es suave y continua, lo que la hace adecuada para técnicas de optimización y cálculo en algoritmos de entrenamiento.

- **Monotonía:** La función sigmoide es monótona, lo que significa que siempre se incrementa o siempre se decrementa a medida que x cambia. Esto es importante para la interpretación de coeficientes en la regresión logística.

- **Asíntotas Horizontales:** La función sigmoide tiene dos asíntotas horizontales: y = 0 cuando e y = 1.

Evaluación del modelo

Métricas de evaluación

Una primera consideración al evaluar un modelo es considerar cuánto "acierta" en sus predicciones. De manera similar a lo que vimos en los test de hipótesis, definiremos una clase "positiva" (generalmente asociada al valor 1) y con ello

- Verdaderos positivos (VP): corresponde a los valores que son clasificados como positivos por el modelo, y efectivamente lo son en la realidad.
- Falsos positivos (FP): corresponde a los valores que son clasificados como positivos por el modelo, y en la realidad eran negativos.

- Verdaderos negativos (VN): corresponde a los valores que son clasificados como negativos por el modelo, y efectivamente lo son en la realidad.
- Falsos negativos (FN): corresponde a los valores que son clasificados como negativos por el modelo, y en la realidad eran positivos.

Lo anterior nos hace definir las siguientes métricas:

Accuracy (Exactitud)

Se calcula como (VP + VN) / (VP + VN + FP + FN), y mide la proporción de todas las predicciones que son correctas. Es una métrica global que evalúa la capacidad del modelo para clasificar correctamente tanto los ejemplos positivos como los negativos.

Es útil cuando las clases están equilibradas en el conjunto de datos, pero puede ser engañosa en conjuntos de datos con desequilibrio de clases.

Precision (Precisión)

Se calcula como VP / (VP + FP), y se enfoca en la proporción de ejemplos que el modelo clasificó como positivos que realmente son positivos. Mide la exactitud de las predicciones positivas.

Es importante cuando los falsos positivos son costosos o problemáticos en un problema dado, como en la detección de fraudes.

Recall (Recuperación o Sensibilidad)

Se calcula como VP / (VP + FN), y mide la capacidad del modelo para capturar todos los ejemplos positivos. Se centra en la proporción de ejemplos positivos que fueron identificados correctamente.

Es importante cuando los falsos negativos son costosos o críticos en un problema, como en pruebas médicas.

F1-Score (Puntuación F1)

Se calcula como 2 * (Precision * Recall) / (Precision + Recall), y combina la precisión y el recall en un solo número. Es útil cuando se busca un equilibrio entre la precisión y la capacidad de recuperación, en situaciones donde los falsos positivos y los falsos negativos tienen consecuencias significativas.

Herramientas de evaluación

Matriz de confusión

Simplemente es una organización en una tabla de los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, para poder observarlos más facilmente y calcular las métricas anteriores.

Área bajo la curva ROC

El área bajo la curva ROC (AUC-ROC) es una métrica importante para evaluar el rendimiento de un modelo de clasificación, como la regresión logística, especialmente cuando se trata de problemas de clasificación binaria.

La Curva ROC es una representación gráfica que muestra cómo varía la tasa de verdaderos positivos (VP) con respecto a la tasa de falsos positivos (FP) a medida que se ajusta el umbral de decisión del modelo. En el eje X de la Curva ROC se registra la tasa de falsos positivos, que mide cuántos falsos positivos se cometen en relación con el total de negativos reales. En el eje Y se registra la tasa de VP (Recall o Sensibilidad), que mide cuántos verdaderos positivos se identifican correctamente en relación con el total de positivos reales.

El AUC-ROC es una métrica numérica que resume el rendimiento del modelo en un solo número. Representa el área bajo la Curva ROC.

- El AUC-ROC varía entre 0 y 1, donde un valor más cercano a 1 indica un mejor rendimiento. Un AUC-ROC de 0.5 indica que el modelo tiene un rendimiento similar al azar.
- Un AUC-ROC de 1.0 significa que el modelo puede separar perfectamente las clases positivas y negativas sin cometer errores.
- Un AUC-ROC de 0.5 indica que el modelo no es mejor que una predicción al azar.
- Un AUC-ROC entre 0.5 y 1.0 indica un rendimiento útil del modelo, donde un valor más alto representa un mejor rendimiento.

El AUC-ROC es especialmente útil para comparar múltiples modelos de clasificación y seleccionar el mejor entre ellos. También es valioso para ajustar el umbral de decisión del modelo en función de los requisitos específicos del problema. Por ejemplo, permite aumentar la sensibilidad a costa de la especificidad o viceversa.
En resumen, el AUC-ROC es una métrica esencial para medir la capacidad de un modelo de regresión logística (u otros modelos de clasificación) para discriminar entre las clases positivas y negativas. Proporciona una evaluación cuantitativa del rendimiento y es una herramienta valiosa en la evaluación de modelos en problemas de clasificación binaria.

Área bajo la curva PR

El AUC-PR es otra métrica importante para evaluar el rendimiento de modelos de clasificación binaria, como la regresión logística.

La Curva PR es una representación gráfica que muestra cómo varía la precisión con respecto al recall a medida que se ajusta el umbral de decisión del modelo. A medida que se aumenta el umbral de decisión, la precisión generalmente aumenta mientras que el recall disminuye, y viceversa.

El AUC-PR es una métrica numérica que resume el rendimiento del modelo en un solo número asociándolo al área bajo la curva PR.

- El AUC-PR varía entre 0 y 1, donde un valor más cercano a 1 indica un mejor rendimiento en términos de precisión y recall.
- Un AUC-PR de 1.0 significa que el modelo puede lograr una precisión perfecta y un recall perfecto, clasificando todos los ejemplos positivos correctamente sin falsos positivos ni falsos negativos.
- Un AUC-PR de 0.0 indica que el modelo tiene un rendimiento muy pobre y no es capaz de clasificar correctamente los ejemplos positivos.

El AUC-PR es particularmente útil cuando estás trabajando con conjuntos de datos desequilibrados, donde una clase es mucho más grande que la otra. Ayuda a evaluar la capacidad del modelo para recuperar ejemplos positivos y evitar falsos positivos, lo cual es crucial en problemas donde los falsos positivos son costosos o críticos.

En resumen, el AUC-PR es una métrica esencial para medir la capacidad de un modelo de regresión logística (u otros modelos de clasificación) para clasificar con precisión ejemplos positivos y negativos, especialmente en situaciones de desequilibrio de clases.

El problema del desbalanceo de clases

El desbalanceo de clases se refiere a una situación en la que las clases en un conjunto de datos no están representadas de manera equitativa, es decir,, algunas clases tienen muchas más muestras que otras lo que puede causar problemas con el modelo. Esto puede ser problemático especialmente al utilizar algoritmos de aprendizaje automático, ya que estos tienden a favorecer la clase mayoritaria debido a su mayor presencia en el conjunto de datos, lo que puede llevar a modelos sesgados y no generalizables.

El problema del desbalanceo de clases es especialmente común en aplicaciones del mundo real, como la detección de fraudes, la detección de enfermedades raras, la clasificación de eventos raros en seguridad, entre otros. En estas situaciones, la clase minoritaria suele ser la más interesante y relevante, pero también es la más difícil de modelar debido a su falta de representación.

Para abordar el problema del desbalanceo de clases, se pueden aplicar diversas técnicas:

- **Sobremuestreo de la clase minoritaria:** Se generan nuevas instancias de la clase minoritaria, ya sea replicando muestras existentes o generando muestras sintéticas mediante técnicas como el método SMOTE (Synthetic Minority Over-sampling Technique).

- **Submuestreo de la clase mayoritaria:** Se reducen las muestras de la clase mayoritaria seleccionando aleatoriamente un subconjunto de ellas. Sin embargo, esto puede llevar a la pérdida de información importante.

- **Peso de clase:** Algunos algoritmos de machine learning permiten asignar pesos diferentes a las clases para que el modelo preste más atención a la clase minoritaria durante el entrenamiento.

- **Anomalía detección:** En algunos casos, se puede abordar el problema como un problema de detección de anomalías, donde se considera la clase minoritaria como la clase "anómala" y se aplican técnicas específicas de detección de anomalías.

- **Generación de características:** Se pueden crear características adicionales que capturen información importante de la clase minoritaria y ayuden al modelo a distinguirla mejor.

Smote

La librería SMOTE en Python se utiliza para abordar el problema del desbalanceo de clases generando muestras sintéticas de la clase minoritaria en un conjunto de datos desbalanceado. SMOTE significa "Synthetic Minority Over-sampling Technique" y es una técnica ampliamente utilizada para equilibrar las clases en conjuntos de datos desbalanceados. La idea detrás de SMOTE es aumentar el número de muestras en la clase minoritaria creando muestras artificiales que son combinaciones ponderadas de muestras existentes.

Regresión lineal y Python

A continuación, veremos cómo construir un modelo de regresión lineal con Python y evaluarlo, considerando un archivo de datos ficticio correspondiente a un archivo **datos.csv**. De él, seleccionaremos sus columnas “A” y “B” como variables independientes, y “C” como dependiente.

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, roc_auc_score, average_precision_score

# Cargar el conjunto de datos desde el archivo "datos.csv"
data = pd.read_csv("datos.csv")

# Variables independientes: "A" y "B"
X = data[["A", "B"]].values

# Variable dependiente: "c"
y = data["c"].values

# Estandarizar las características (opcional pero recomendado)
scaler = StandardScaler()
X = scaler.fit_transform(X)

# Crear el modelo de Regresión Logística
logistic_regression_model = LogisticRegression(random_state=42)

# Ajustar el modelo a los datos
logistic_regression_model.fit(X, y)

# Realizar predicciones en el conjunto de datos
y_pred = logistic_regression_model.predict(X)

# Calcular métricas de desempeño
accuracy = accuracy_score(y, y_pred)
precision = precision_score(y, y_pred)
recall = recall_score(y, y_pred)
f1 = f1_score(y, y_pred)
roc_auc = roc_auc_score(y, y_pred)
average_precision = average_precision_score(y, y_pred)

# Imprimir las métricas de desempeño
print("Métricas de Desempeño:")
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
print("AUC-ROC:", roc_auc)
print("AUC-PR:", average_precision)
```

Si queremos aplicar SMOPE, deberemos variar un poco el código, como se muestra

```
#agregamos la librería (las anteriores las supondremos cargadas
from imblearn.over_sampling import SMOTE

# Cargar el conjunto de datos desde el archivo "datos.csv"
data = pd.read_csv("datos.csv")

# Variables independientes: "A" y "B"
X = data[["A", "B"]].values

# Variable dependiente: "c"
y = data["c"].values

# Aplicar SMOTE para balancear las clases
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y)

# Estandarizar las características (opcional pero recomendado)
scaler = StandardScaler()
X_resampled = scaler.fit_transform(X_resampled)

# Crear el modelo de Regresión Logística
logistic_regression_model = LogisticRegression(random_state=42)

# Ajustar el modelo a los datos con SMOTE
logistic_regression_model.fit(X_resampled, y_resampled)

# Realizar predicciones en el conjunto de datos con SMOTE
y_pred = logistic_regression_model.predict(X_resampled)

# Calcular métricas de desempeño con SMOTE
accuracy = accuracy_score(y_resampled, y_pred)
precision = precision_score(y_resampled, y_pred)
recall = recall_score(y_resampled, y_pred)
f1 = f1_score(y_resampled, y_pred)
roc_auc = roc_auc_score(y_resampled, y_pred)
average_precision = average_precision_score(y_resampled, y_pred)

# Imprimir las métricas de desempeño con SMOTE
print("Métricas de Desempeño con SMOTE:")
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
```

```
print("AUC-ROC:", roc_auc)
print("AUC-PR:", average_precision)
```

Preguntas de cierre

- ¿Qué es la regresión logística y en qué tipos de problemas se utiliza comúnmente?
- ¿Cuál es la diferencia entre regresión lineal y regresión logística?
- Explique el término "logit" y su relación con la probabilidad.
- ¿Cómo se interpretan los coeficientes en la regresión logística?
- ¿Cuáles son las métricas más comunes utilizadas para evaluar la calidad de un modelo de regresión logística?