Regresión Logística

Tutoría

{desafío} latam_

Supongamos que se busca predecir si un estudiante aprobará o reprobará un examen en función de su tiempo de estudio u otros hábitos.

¿Cómo podríamos abordar este problema? ¿Qué tipo de modelo sería adecuado?

Autoaprendizaje

Recursos asincrónicos

- ¿Revisaste los recursos de la semana 7 (Guía y desafío)?
- ¿Tienes dudas sobre alguno de ellos?

{desafío} latam_

Ideas fuerza

*La **regresión logística** nos permite realizar **clasificaciones** en base a variables independientes y una **variable de clasificación** (generalmente binaria).*

*Utiliza una **función de probabilidad** para realizar la clasificación, lo que permite **ajustar el modelo** a partir de **métricas***

*Podemos **evaluar un modelo** de regresión logística por medio de **métricas e indicadores,** que comparan las relaciones entre preicciones y realidad.*

/* Regresión Logística*/

Regresión logística

¿En qué consiste?

La regresión logística es un modelo estadístico ampliamente utilizado en el campo de la ciencia de datos y el aprendizaje automático para abordar problemas de clasificación.

Aunque su nombre incluye la palabra "regresión", en realidad se emplea para determinar la probabilidad de que un evento pertenezca a una de dos o más categorías, y asignarlo a una de ellas según el resultado obtenido.

Regresión logística

Características e importancia

Simplicidad

Es un modelo menos complejo que otros métodos de ML. Por lo tanto, pueden implementarse incluso sin mayor experiencia en ML

Velocidad

Permite procesar grandes volúmenes de datos a alta velocidad ya que requieren menos capacidad computacional en cuanto a memoria y potencia de procesamiento.

Flexibilidad

Permite responder preguntas que tienen dos o más resultados finitos e incluso preprocesar datos.

Visibilidad

Permite una mayor visibilidad de los procesos de software internos que otras técnicas de análisis de datos

Regresión logística

Funcionamiento

De manera similar a la utilizada para la regresión lineal, buscamos relacionar un conjunto de variables independientes con una dependiente, lo que hacemos utilizando coeficientes que determinan un modelo lineal.

$$x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + + \beta_n x_n$$

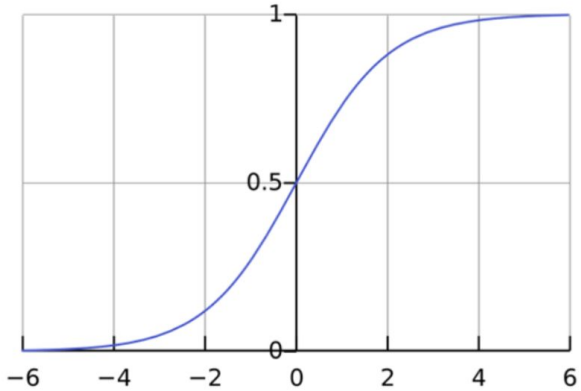
A través de una función logarítmica o logit, transforma la salida en una escala log-odds, lo que permite interpretar y tomar decisiones basadas en la probabilidad.

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

Regresión logística

Función logística

La función logística o logit devuelve solo valores entre 0 y 1 para la variable dependiente. Se establecen así zonas que permiten asignar, probabilísticamente, si un valor pertenece a una u otra categoría.



Regresión logística

Umbral de decisión

Al realizar una clasificación binaria, definiremos una categoría como “positiva” y la otra como “negativa”. Para afirmar que un elemento pertenece a la clase positiva, asignamos un **umbral de decisión**, es decir, un umbral por sobre el cual debe encontrarse la probabilidad para afirmar, con certeza, que el elemento pertenece a dicha categoría.

/*Evaluación del modelo*/

Evaluación del modelo

Métricas de evaluación

	Indicador	Descripción	Ecuación
1	**Exactitud (Accuracy)**	Proporción de predicciones correctas sobre el total de predicciones. Es útil cuando las clases están balanceadas.	(VP + VN)/ (VP+VN+FP+FN)
2	**Precisión (Precision)**	Proporción de predicciones positivas que fueron correctas. Es útil cuando se busca minimizar los falsos positivos.	VP/(VP+FP)
3	**Sensibilidad (Recall)**	Proporción de verdaderos positivos que se identificaron correctamente. Es útil cuando se busca minimizar los falsos negativos.	VP/(VP+FN)
4	**Puntuación F1 (F1-Score)**	Combina la precisión y el recall en una sola métrica.	2*(Precision*Recall)/ (Precision+Recall)

{desafío} latam_

VP: Verdadero positivos
FP: Falsos positivos

VN: Verdaderos negativos
FN: Falsos negativos

Evaluación del modelo

Herramientas de evaluación

01	**Matriz de confusión**	Es una tabla que muestra el número de predicciones correctas e incorrectas para cada clase. Es útil para visualizar el rendimiento del modelo en detalle.
02	**Área bajo la curva (AUC - ROC)**	La curva ROC (Receiver Operating Characteristic) es una representación gráfica del rendimiento del modelo en función del umbral de decisión.
03	**Área bajo la curva PR (AUC - PR)**	La curva PR (Precision-Recall) es una representación gráfica del rendimiento del modelo en términos de precisión y recall.

Evaluación del modelo

AUC - ROC

Luego, calculamos el área bajo esta curva ROC. Cuanto mayor sea el AUC-ROC, mejor será el rendimiento del modelo. Así:

- El AUC-ROC varía entre 0 y 1.
- Un AUC-ROC de 0.5 indica un rendimiento similar al azar, mientras que un AUC-ROC de 1 indica un rendimiento perfecto. En general, cuanto mayor sea el AUC-ROC, mejor será el modelo en la clasificación binaria.
- La curva ROC muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos a diferentes umbrales de clasificación. Un modelo con un AUC-ROC alto tiene una mayor capacidad para discriminar entre clases.

Evaluación del modelo

AUC - PR

- El AUC-PR varía entre 0 y 1.
- Al igual que con el AUC-ROC, un AUC-PR de 0.5 indica un rendimiento similar al azar, mientras que un AUC-PR de 1 indica un rendimiento perfecto.
- El AUC-PR se centra más en la precisión y la recuperación en problemas de clasificación desequilibrados, donde una clase es mucho más común que la otra.
- Un AUC-PR alto indica que el modelo tiene un buen equilibrio entre precisión y recuperación en la clasificación.

Evaluación del modelo

AUC -ROC vs AUC - PR

En resumen, el AUC-ROC mide la capacidad del modelo para discriminar entre clases en función de la tasa de falsos positivos y verdaderos positivos, mientras que el AUC-PR se centra en la precisión y la recuperación en problemas de clasificación desequilibrados. Ambas métricas son útiles para evaluar el rendimiento de un modelo de clasificación binaria, y la elección de cuál utilizar depende del contexto y la distribución de las clases en tus datos.

/*Desbalanceo de clases*/

Desbalanceo de clases

¿Qué problemas trae?

Si una clase tiene muchos más elementos que la otra, se pueden ocasionar problemas al generar el modelo ya que se podría favorecer en demasía a favorecer la clase mayoritaria, 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.

Desbalanceo de clases

Solución

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.

¡Manos a la obra!
Regresión logística y cáncer

{desafío}
latam_

Manos a la Obra

Regresión logística y cáncer

A continuación, veremos cómo implementar un modelo de regresión logística a partir de una base de datos que relaciona diferentes características con un diagnóstico positivo o negativo de cáncer.

Para esto, abre tu archivo de Jupyter Notebook utilizado durante la clase.

Puedes descargarlo desde la sesión de la clase

Prueba
“Análisis estadístico con Python”

{desafío}
latam_

Prueba

"Análisis estadístico con Python"

- ¿Leíste la prueba? ¿Comprendes bien lo que se solicita en cada caso?
- ¿Hay contenidos que necesitas repasar antes de comenzar este desafío?
- ¿Necesitas algún ejemplo o indicación para alguna pregunta o requerimiento específico?

{desafío} latam_

{desafío}
latam_

Academia de
talentos digitales