UNIVERSITAS Miguel Hernández

Tema 1. Regresión Lineal. Práctica.

José L. Sainz-Pardo Auñón

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones.

Introducción a la Regresión Logística

- La regresión logística es un modelo estadístico usado para predecir la probabilidad de una clase binaria.
- Fórmula básica:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

- Ejemplos de aplicación:
  - Predicción de enfermedades (enfermo/no enfermo).
  - Clasificación de correos electrónicos (spam/no spam).
  - Diagnóstico de cáncer (benigno/maligno).

Descarga y Carga de los Datos

- Descarga el dataset **Breast Cancer Wisconsin (Diagnostic)** desde el UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`.
- Carga los datos en tu entorno de programación utilizando Python. Asegúrate de que todas las variables se carguen correctamente.
- Revisa las primeras filas del dataset y obtén un resumen estadístico básico para entender la distribución de las características.

Explora los Datos

- Analiza la variable objetivo (`Diagnosis`) y asegúrate de identificar la proporción de casos benignos y malignos.
- Visualiza las distribuciones de las variables más importantes como `Radius Mean`, `Texture Mean`, y otras características clave.
- Convierte la variable `Diagnosis` a numérica: asigna 0 a los valores "B" (benigno) y 1 a los valores "M" (maligno).
- Dibuja un mapa de correlaciones entre las variables para identificar relaciones importantes.

Preprocesa los Datos

- Elimina la columna `ID`, ya que no aporta información relevante al modelo.
- Divide el dataset en conjunto de entrenamiento (70%) y de prueba (30%) utilizando la función `train_test_split`.

Ajusta el Modelo de Regresión Logística

- Entrena un modelo de regresión logística utilizando los datos del conjunto de entrenamiento.
- Analiza los coeficientes obtenidos.

Evalúa el Modelo

- Genera predicciones tanto para el conjunto de prueba como para el de entrenamiento utilizando el modelo ajustado.
- Visualiza la matriz de confusión (tanto en el conjunto de prueba como en el de entrenamiento) para analizar el rendimiento del modelo en términos de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.
- En ambos conjuntos, calcula las métricas de rendimiento como la precisión (accuracy), sensibilidad (recall) y precisión (precision).
- Traza la curva ROC y calcula el área bajo la curva (AUC) para evaluar la capacidad del modelo para distinguir entre clases.
- Prueba a realizar predicciones con algún umbral distinto a 0.5.

Interpreta los Resultados

- Calcula la exactitud del modelo y analiza los resultados obtenidos.
- Interpreta la matriz de confusión y el reporte de clasificación para identificar qué tan bien el modelo predice los tumores benignos y malignos.
- Revisa el valor del AUC y discute la capacidad del modelo para distinguir entre casos positivos y negativos.
- Reflexiona sobre posibles mejoras al modelo, como ajustar hiperparámetros o probar otros algoritmos de clasificación.
- Si normalizas los datos numéricos, ¿obtienes las mismas estimaciones?

Practica con Otros Datasets

- Repite la práctica con otros datasets disponibles en el UCI Machine Learning Repository para reforzar los conceptos de regresión logística.
- Algunas sugerencias:
  - **Pima Indians Diabetes Database:** `https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes`
  - **Heart Disease Dataset:** `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`
  - **Bank Marketing Dataset:** `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`
  - **Adult (Census Income) Dataset:** `https://archive.ics.uci.edu/ml/datasets/Adult`
- Sigue el mismo flujo de trabajo: carga los datos, explora las variables, preprocesa el dataset, ajusta un modelo de regresión logística y evalúa los resultados. Si hubiera variables **independientes categóricas**, descártalas del modelo.