



UNIVERSITAS
Miguel Hernández

Tema 5. SVM. 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.

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+\(Diagnostic\)](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 SVM

- Obtén a partir del conjunto de entrenamiento el C que mejor funciona para el modelo SVM . Prueba, por ejemplo, con el siguiente vector de búsqueda(0.001, 0.01, 0.1, 1, 10, 100 y 1000).
- Obtén los pronósticos de la muestra de prueba con el mejor C que obtuviste.
- Obtén el informe de clasificación del modelo, utilizando la librería `sklearn`.
- Realiza distintas pruebas con distintos scores (accuracy, precision, ...)

Validación cruzada

- Realiza una validación cruzada utilizando 4 pliegues (4-fold, o 25% de ítems en la muestra de reserva) para evaluar el modelo con el conjunto de datos de entrenamiento.
- Calcula y muestra la media de las puntuaciones AUC obtenidas durante la validación cruzada.
- Obtén las predicciones sobre el conjunto de prueba.
- Obtén el informe de clasificación del conjunto de prueba, utilizando la librería sklearn.
- Obtén la curva ROC y el área bajo la misma (AUC).

Practica con Otros Datasets

- Repite la práctica con otros datasets disponibles en el UCI Machine Learning Repository y ya conocidos tanto de Regresión como de Clasificación. Por ejemplo:
 - ▶ **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>
 - ▶ **Auto MPG Dataset:** <https://archive.ics.uci.edu/dataset/9/auto+mpg>
 - ▶ **Wine Quality Dataset:**
<https://archive.ics.uci.edu/dataset/186/wine+quality>
 - ▶ **Energy Efficiency Dataset:**
<https://archive.ics.uci.edu/dataset/242/energy+efficiency>
 - ▶ **Concrete Compressive Strength:**
<https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength>
- Sigue el mismo flujo de trabajo: carga los datos, explora las variables, preprocesa el dataset, ajusta el modelo y evalúa los resultados sin y con validación cruzada. Prueba también a seleccionar prototipos.
- Compara los resultados con aquellos obtenidos con Regresión Lineal, Logística y/o kNN.