UNIVERSITAS Miguel Hernández

Tema 3. kNN. 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)`.
- 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 kNN

- Obtén a partir del conjunto de entrenamiento el k entre 1 y 31 con el que que mayor AUC se obtiene para el modelo k-NN (puedes utilizar la función GridSearchCV).
- Obtén los pronósticos de la muestra de prueba con el mejor k que obtuviste.
- Obtén el informe de clasificaciónn 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 librearía sklearn.
- Obtén la curva ROC y el área bajo la misma (AUC).

Selección de prototipos mediante CNN.

- Utiliza Condensed Nearest Neigbor (CNN) para seleccionar prototipos de la muestra de entrenamiento.
- Graba en un fichero Excel el conjunto de entrenamiento y en otro fichero Excel el de prototipos obtenido. Observa ambos. ¿En qué porcentaje se realizó la reducción?
- Obtén las predicciones y el informe de clasificación utilizando el conjunto de prototipos. ¿Mejoró la proporción de acierto utilizando prototipos? Compara ambos informes de clasificación (con y sin prototipos).

Selección de prototipos mediante ENN.

- Utiliza Edited Nearest Neigbor (ENN) para seleccionar prototipos de la muestra de entrenamiento.
- Graba en un fichero Excel el conjunto de entrenamiento y en otro fichero Excel el de prototipos obtenido. Observa ambos. ¿En qué porcentaje se realizó la reducción?
- Obtén las predicciones y el informe de clasificación utilizando el conjunto de prototipos. Compara todos los informes de clasificación (sin prototipos, con prototipos CNN y con prototipos ENN). ¿En cuál de todos la proporción de acierto resultó mejor?

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 o Logística.

Reconocimiento de género musical

Roni Bandini creó un algoritmo para reconocer el género musical que su vecino escuchaba y cuando la clasificaba como 'Reaggaetone' interfería en su emisión:
`https://es.euronews.com/next/2024/04/09/`

`reggaetonbe-gone-esta-maquina-casera-silencia-la-musica-de-los-vec`

El Spotify Tracks Dataset (spotify.csv) contiene un conjunto de canciones con sus características de audio, como energía, tempo, etc. recopilado de Spotify mediante su API. Tiene 114.000 registros de canciones y su género. Son muchos registros para aplicar kNN, no obstante puedes practicar, por ejemplo, con 1000 registros tomados al azar. Por cierto, en Kaggle existe un hilo sobre este Dataset:

`https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset?resource=download`
en el que encontrarás la descripción de cada variable y código sobre métodos para tratarlo.

Propuesta de Trabajo Fin de Máster (TFM)

Como has podido ver, el Spotify Dataset Tracks contiene cientos de miles de canciones con diversas características acústicas y no es posible la aplicación de técnicas de clasificación sobre toda la base de datos.

El objetivo de esta propuesta de TFM es desarrollar un modelo de clasificación eficiente en el que se desarrollen técnicas avanzadas de selección de prototipos para reducir el tamaño del conjunto de entrenamiento sin sacrificar la precisión, mejorando así el rendimiento del modelo de clasificación.

Aquellas personas interesadas en realizar este TFM, enviad email a:
`jlsainz@umh.es`