UNIVERSITAS Miguel Hernández

Tema 3. kNN. Ejemplos.

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.

1. Carga y Preparación de los Datos

- Descargar el archivo `clasificacion.xlsx` que contiene los datos sobre los que deseamos emplear la técnica de clasificación.
- Leer el archivo Excel en un `DataFrame` de `pandas`.
- Visualizar las primeras filas del conjunto de datos.
- Definir las variables independientes X_1, X_2, X_3 y la variable dependiente Y.

2. Obtención del k del k-NN.

- Divide los datos en un conjunto test del 30% y uno de entrenamiento del 70%
- Obtén a partir del conjunto de entrenamiento el k entre 1 y 31 que mejor funciona para el modelo $k - NN$ (puedes utilizar la función *GridSearchCV*.).
- ¿Qué k resultó ser el que maximizaba el área bajo la curva ROC (AUC)?
- Prueba distintas métricas de rendimiento (score) y compara los resultados.

3. Evaluación del modelo

- Obtén los pronósticos de la muestra de prueba con el mejor k que obtuviste.
- Obtén las probabilidades de clasificación de la muestra de prueba.
- Graba el dataframe de los datos de prueba (X's, Y's, Predicciones y Probabilidades). Visualízalo y razona cómo son obtenidas las probabilidades. Interprétalas.
- Obtén el informe de clasificación del modelo, utilizando la librearía sklearn.

4. 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).

5. 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).

5. 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?

6. kNN para regresión.

- Realiza una regresión sobre el fichero 'regresion.xlsx' mediante kNN.
- Utiliza el mejor valor de k entre 1 y 31.
- Obtén el error cuadrático medio y el R^2 del modelo.
- Obtén un gráfico de los valores observados frente a los valores predichos.
- Prueba a seleccionar prototipos a ver si mejoras las predicciones.


UNIVERSITAS
Miguel Hernández