Método basado en vecinos más cercanos

Rafael Jordá Muñoz

En este PDF se resumen las tres actividades que se van a realizar para comprender el algoritmo de ML basado en vecinos más cercanos.

1. Ejercicio 1

Dado el siguiente conjunto de datos de clasificación con 6 observaciones, 3 variables de entrada y una variable de salida:

Observación	**X1**	**X2**	**X3**	**Y**
1	0	3	2	1
2	3	0	3	0
3	0	3	-1	0
4	3	0	0	1
5	1	2	1	1
6	2	1	0	0

Tabla 1: Datos de entrenamiento

Suponiendo que se quiere hacer la predicción de la variable de salida para X1=0, X2=0, X3=0 mediante KNN:

a) Computar la distancia entre cada observación y el punto de test.

b) ¿Cuál es la predicción para K=1? ¿Por qué?

c) ¿Cuál es la predicción para K=3? ¿Por qué?

Este ejercicio debe hacerse SIN usar scikit-learn, para aplicar directamente los conceptos teóricos del método.

2. Ejercicio 2

Dado el problema de clasificación Blood Transfusion Service Center:

a) Analiza las características del conjunto de datos: número y tipo de variables de entrada y salida, número de instancias, número de clases y distribución de las mismas, correlación entre las variables, valores perdidos, etc.

b) Una de las clases que implementa el algoritmo KNN en la librería scikit-learn es sklearn.neighbors.KNeighborsClassifier. Revisa los parámetros y métodos que tiene. Dado que en determinadas operaciones (división del conjunto de datos en entrenamiento y test, validación cruzada, entrenamiento) el resultado obtenido depende de números aleatorios, y se quiere que el resultado sea repetible en otras máquinas, es necesario que utilicéis la función np.random.seed(1) al comienzo del fichero .ipynb.

c) Divide los datos en entrenamiento (80 %) y test (20 %).

d) Realiza la experimentación con KNN (KNeighborsClassifier) usando como hiper-parámetro el número de vecinos:

- Muestra la gráfica del error de entrenamiento con validación cruzada (5-CV) frente al valor del hiper-parámetro. ¿Cuál es el menor error de validación cruzada, su desviación estándar y el valor del hiper-parámetro para el que se consigue? ¿Cuál es el valor del hiper-parámetro si se aplicase la regla de una desviación estándar?
- Muestra la gráfica del error de test frente al valor del hiper-parámetro, y valora si la gráfica del error de entrenamiento con validación cruzada ha hecho una buena estimación del error de test. ¿Cuál es el menor error de test y el valor del hiper-parámetro para el que se consigue? ¿Cuál es el error de test para el valor del hiper-parámetro seleccionado por la validación cruzada? ¿Cuál es el error de test para el valor del hiper-parámetro seleccionado por la validación cruzada mediante la regla de una desviación estándar?

3. Ejercicio 3

Repite el ejercicio 2 pero para el problema de regresión Energy Efficiency con la variable de salida cooling load. Al ser un problema de regresión deberás utilizar KNeighborsRegressor, y como medida de error de entrenamiento y test el MSE.