Máquinas de soporte vectorial

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 16 observaciones, 2 variables de entrada y una variable de salida, mediante una SVM lineal con C=1 se han obtenido los coeficientes α_i indicados en la última columna:

Observación	**X1**	**X2**	**Y**	α_i
0	2	6	1	0
1	4	3	1	1
2	4	4	1	0.3333
3	4	6	1	0
4	6	3	1	1
5	7	7	1	0.1667
6	8	4	1	1
7	9	8	1	1
8	2	1	-1	1
9	6	2	-1	0.5
10	7	4	-1	1
11	8	8	-1	1
12	9	1	-1	0
13	10	3	-1	0
14	10	6	-1	1
15	12	4	-1	0

Tabla 1: Datos de entrenamiento

Indica:

- Cuáles son los vectores de soporte y cuáles de ellos están sobre el margen.
- Cuáles son los coeficientes del hiperplano (β y β_0) y el valor de M.
- Los valores de ϵ_i y las observaciones incorrectamente clasificadas.

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) La clase que implementa las SVM en problemas de clasificación en scikit-learn es sklearn.svm.SVC (existen otras dos clases, pero nos centraremos en ésta). 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.

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

c) Realiza la experimentación con SVC usando los valores por defecto de los parámetros, excepto para kernel en donde deberás probar el 'linear', 'poly' (con gamma=1) y 'rbf'. Además, utiliza como hiper-parámetro la variable C (en todos los kernels), degree (grado del polinomio, debe ser mayor que 1) en el caso del kernel polinomial, y gamma en el caso del kernel rbf.

   - Muestra la gráfica del error de entrenamiento con validación cruzada (5-CV) frente al valor del hiper-parámetro (en el caso del kernel rbfmuestra la gráfica frente a C para algunos valores de gamma –los que consideres más representativos-; de forma equivalente para degreecon el kernel polinomial – el grado debe ser mayor que 1, si no sería lineal-), y justifica la elección del valor más apropiado. Para cada tipo de kernel, ¿cuál es el menor error de validación cruzada, su desviación estándar y el valor de los hiper-parámetros para el que se consigue?
   - 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. Para cada tipo de kernel, ¿cuál es el menor error de test y el valor de los hiper-parámetros para el que se consigue? ¿Cuál es el error de test para el valor de los hiper-parámetros seleccionados por la validación cruzada?

3. Ejercicio 3

Repite el ejercicio 2 pero para el problema de clasificación Pima Indians Diabetes.