Boletín 5: Máquinas de Soporte Vectorial

Para la realización de las prácticas correspondientes a este boletín se utilizará scikit-learn.

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	X_1	X_2	Y	α_i
0	2	6	1	0
1	4	3	1	1
2	4	4	1	0,33
3	4	6	1	0
4	6	3	1	1
5	7	7	1	0,17
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

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.

Nota: este ejercicio debe hacerse sin utilizar ninguna función de scikit-learn.

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 esta). Revisa los parámetros y métodos que tiene.

   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) en el caso del *kernel* polinomial, y *gamma* en el caso del *kernel rbf*.

      i. Muestra la gráfica del error de entrenamiento con validación cruzada (5-CV) frente al valor del hiper-parámetro (en el caso del kernel *rbf* muestra la gráfica frente a *C* para algunos valores de *gamma* –los que consideres más representativos-; de forma equivalente para *degree* con el *kernel* polinomial), y justifica la elección del valor más apropiado.

      ii. 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.

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