

Boletín 5: Máquinas de Soporte Vectorial

Para la realización de las prácticas correspondientes a este boletín se utilizará [scikit-learn](https://scikit-learn.org/). Utilizaremos un **SEED_VALUE=1**.

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
1	2	6	1	0
2	4	3	1	1
3	4	4	1	0,3333
4	4	6	1	0
5	6	3	1	1
6	7	7	1	0,1667
7	8	4	1	1
8	9	8	1	1
9	2	1	-1	1
10	6	2	-1	0,5
11	7	4	-1	1
12	8	8	-1	1
13	9	1	-1	0
14	10	3	-1	0
15	10	6	-1	1
16	12	4	-1	0

Indica:

- Cuáles son los vectores de soporte y cuáles de ellos están en el límite del 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 ésta). 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 los siguientes hiper-parámetros:
 - i. *kernel* en donde deberás probar el '*linear*', '*poly*' (con $\gamma=1$) y '*rbf*'.
 - ii. *degree*: grado del polinomio en el *kernel* polinómico. Debe ser mayor que 1, si no sería lineal.
 - iii. *gamma* en el caso del *kernel rbf*. Prueba potencias enteras de 10 (...; 0,01; 0,1; 1; 10; 100; ...). Para el *kernel* polinómico utiliza *gamma*=1.
 - iv. *C*, parámetro de regularización (para todos los *kernels*). Prueba potencias enteras de 10 (...; 0,01; 0,1; 1; 10; 100; ...).

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. 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. Repite el ejercicio 2 pero para el problema de clasificación [Pima Indians Diabetes](#).

Entregable

Se debe entregar un único fichero comprimido con el nombre *PrimerApellido_SegundoApellido.zip* (también son válidos los formatos .rar y .7z), que contenga dos archivos:

- El primer archivo debe ser de tipo pdf, y contendrá exclusivamente las respuestas a los ejercicios (incluyendo las gráficas necesarias para justificar dichas respuestas). No se incluirá en este archivo ningún otro tipo de texto.
- El segundo archivo será de tipo ipynb, y permitirá reproducir toda la experimentación realizada en el boletín.