

# PRÁCTICA 4

**Pregunta 1 – Abre el script y explica su contenido. Podrás ver que se utiliza el primer dataset de ejemplo y se realiza la representación gráfica del SVM. Comenta que tipo de kernel se está utilizando y cuáles son los parámetros de entrenamiento.**

En primer lugar, se hace una lectura de los datos a utilizar, cargando el dataset1.

En segundo lugar, se entrena el modelo SVM usando un kernel lineal y  $C=100$  lo que indica que el modelo es bastante estricto en el proceso de clasificación, pudiendo llevar a un sobreajuste del modelo.

Por último, tras entrenar el modelo con los datos del dataset y los parámetros indicados, se hace una representación gráfica del mismo, mostrando los puntos y el hiperplano separador de las dos clases.

**Pregunta 2 – Intuitivamente, ¿qué hiperplano crees que incurrirá en un menor error de test en la tarea de separar las dos clases de puntos?**

Aparentemente, lo mejor es un hiperplano lineal que separe claramente a las dos clases representadas.

**Pregunta 3 – Modifica el script probando varios valores de  $C$ , en concreto,  $C \in \{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4\}$ . Observa qué sucede, explica porqué y escoge el valor más adecuado.**

A medida que aumenta el valor del parámetro  $C$ , el modelo SVM se vuelve más estricto. Visualizando las gráficas resultantes con los distintos valores de  $C$ , podemos observar que para los valores más grandes de  $C$  obtenemos unos mejores resultados de clasificación. A partir de  $C=1$  todos los patrones se han clasificado correctamente por lo que el valor más adecuado de  $C$  que debemos coger sería  $C=10$  o  $C=100$  para que no haya sobre – entrenamiento.

**Pregunta 4 – Prueba a lanzar una SVM lineal con los valores para  $C$  que se utilizaron en la pregunta anterior. ¿Consigues algún resultado satisfactorio en el sentido de que no haya errores en el conjunto de entrenamiento? ¿Por qué?**

Utilizando los valores del parámetro  $C$  del apartador anterior no se ha obtenido ningún resultado satisfactorio, ya que se detectan en las gráficas muchísimos errores. Al usar un kernel lineal en el modelo SVM no se pueden separar las dos clases que se representan.

**Pregunta 5 – Propón una configuración de SVM no lineal (utilizando el kernel tipo RBF o Gaussiano) que resuelve el problema. El resultado debería ser similar al de la Figura 3. Muévete en potencias de 2 utilizando un rango del tipo  $\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$ . ¿Qué valores has considerado para  $C$  y para  $\gamma$ ? Además, incluye un ejemplo de una configuración de parámetros que produzca sobre - entrenamiento y otra que produzca infra - entrenamiento.**

**Pregunta 6 – En este caso, ¿es el dataset linealmente separable? A primera vista, ¿detectas puntos que presumiblemente sean outliers? ¿por qué?**

En este caso, el dataset no es linealmente separable ya que se ven los puntos de ambas clases mezclados en todo el espacio. Por otro lado, tampoco se detectan puntos outliers, es decir puntos que claramente están fuera de su clase.

**Pregunta 7 – Lanza una SVM para clasificar los datos, con objeto de obtener un resultado lo más parecido al de la Figura 5. Muévete en potencias de 2 utilizando un rango del tipo  $\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$ . Establece el valor de los parámetros óptimos. Además, incluye un ejemplo de una configuración de parámetros que produzca sobre – entrenamiento y otra que produzca infra – entrenamiento.**

El valor de los parámetros óptimos son  $C=10000$  y  $g=0.2$

Por otro lado, con los siguientes valores de parámetros podemos obtener las siguientes configuraciones:

- Infra-entrenamiento:  $C=1$  y  $g=0.2$
- Sobre-entrenamiento:  $C=10$  y  $g=200$