Guión de prácticas
P2 – Máquinas de Vectores de Soporte

En esta práctica comprenderemos mejor las máquinas de vectores de soporte (*SVMs*, del inglés *Support Vector Machines*). En particular, los objetivos son:

- Profundizar en el funcionamiento de las SVMs, entendiendo los conceptos de hiperplano, márgenes dura y blanda (*hard* and *soft margins*), vectores de soporte, *kernel*, etc.
- Analizar el impacto del parámetro C en el entrenamiento de una SVM.
- Conocer los *kernels* más relevantes (lineal, polinomial y de base radial Gaussiana) así como sus particularidades y diferencias.

Dividiremos la práctica en dos partes. En la primera (estudio guiado) trabajaremos con conjuntos de datos "de juguete" sintéticos que permiten asentar los fundamentos teóricos de las SVMs. En la segunda (caso práctico) abordaremos un problema de mayor envergadura y el objetivo será alcanzar el mejor rendimiento posible poniendo en práctica lo aprendido previamente en el estudio guiado. A lo largo de toda la práctica utilizaremos la implementación de SVM proporcionada por Scikit Learn:

sklearn.svm.SVC()

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

1. Estudio guiado

1.1. Problema separable linealmente

Comenzaremos con el caso más sencillo: un problema de clasificación binario con 2 variables predictoras donde los datos de cada clase son linealmente separables. Seguid los siguientes pasos:

a. Cargad en memoria el Dataset 1 ("dataset_1.csv") e inspeccionad brevemente sus propiedades.
b. Dibujad la nube de puntos en un gráfico de dispersión coloreando cada ejemplo de datos en función de su clase.
c. Tratad de dibujar sobre el gráfico anterior algún hiperplano que separe los datos en función de su clase con una tasa de acierto del 100%.
d. Entrenad una SVM con todo el conjunto de datos. Utilizad para ello la función svm.SVC() con kernel lineal. Usad una C=100.0 y dejad los demás hiperparámetros con sus valores por defecto. Comprobad que la SVM tiene un 100% de precisión sobre los datos de entrenamiento.
e. Dibujad de nuevo la nube de puntos, pero esta vez añadid el hiperplano aprendido por la SVM y sus márgenes. Si queréis, también podéis pintar cada lado del hiperplano con un color distinto para que quede bien claro cuáles son las fronteras de decisión. Finalmente, marcad de algún modo cuáles son los vectores de soporte.
f. Repetid los pasos *d. - e.,* pero esta vez con C=1.0. Analizad las diferencias y explicad el impacto del hiperparámetro C en el proceso de aprendizaje.

1.2. Problema cuasi-separable linealmente

Ahora vamos a trabajar con el Dataset 2 ("dataset_2.csv"). Se trata de una nueva versión del conjunto de datos previo al que hemos añadido algo de ruido. Los pasos a seguir son:

g. Cargad el Dataset 2 e inspeccionadlo.

Guión de prácticas
P2 – Máquinas de Vectores de Soporte

h. Entrenad una SVM con kernel lineal. Estimad el valor óptimo de C mediante alguna técnica de validación cruzada.
i. Representad gráficamente el mejor ajuste alcanzado.

1.3. Problema no separable linealmente (I)

A continuación trabajaremos con un conjunto de datos de características similares a los dos anteriores, pero que no es separable linealmente. Haced lo siguiente:

j. Cargad el Dataset 3 ("dataset_3.csv") y visualizad su nube de puntos.
k. Tratad de ajustar una SVM con kernel lineal.
l. Probad ahora con un kernel polinómico. Emplead un polinomio de grado 3 y un término independiente mayor que 0. Podéis fijar C=100.0.
m. Visualizad y comparad los ajustes de los modelos construidos en los pasos *k.* y *l*. Analizad el funcionamiento del kernel polinómico y su impacto en los resultados.

1.4. Problema no separable linealmente (II)

Finalmente vamos a trabajar con otro conjunto de datos algo más grande que tampoco es separable linealmente:

n. Cargad el Dataset 4 ("dataset_4.csv") y visualizad su nube de puntos.
o. Tratad de ajustar una SVM con kernel lineal o polinómico.
p. Probad ahora con un kernel de base radial. Podéis emplear C=100.0 y dejar los demás hiperparámetros con sus valores por defecto.
q. Analizad el funcionamiento del kernel de base radial y su impacto en los resultados.

2. Caso práctico

Vamos a trabajar con los conjuntos de datos **Breast Cancer** y **Adult** (los mismos de la Práctica P1). El objetivo es entrenar una SVM que consiga el mejor rendimiento posible en validación cruzada 5CV para cada conjunto de datos. Podéis ajustar los hiperparámetros que consideréis oportunos, siguiendo los métodos que prefiráis de entre los vistos en la primera parte de la práctica. Podéis reutilizar el pre-procesado de datos hecho en P1. Comparad los resultados obtenidos con SVM respecto a los conseguidos con los mejores modelos entrenados en P1 con **sklearn.tree.DecisionTreeClassifier** y **sklearn.neighbors.KNeighborsClassifier**.

Guión de prácticas
P2 – Máquinas de Vectores de Soporte

Realización, entrega y evaluación de la práctica

- Realizaréis el trabajo en el mismo **grupo de personas** que el resto de prácticas, según las pautas siguientes:
  - La entrega será única, siendo responsables todas las personas integrantes del grupo.
  - Debéis explicar y justificar suficientemente las diferentes tareas realizadas en la práctica, demostrando dominio de los contenidos.
- Dedicaremos a la práctica dos sesiones interactivas.
- La fecha límite de entrega será la indicada en el aula virtual.
- La entrega consistirá en un único fichero comprimido en **formato .zip** que contenga una memoria (en PDF) describiendo el trabajo realizado, justificando cada decisión tomada, y todos los ficheros con la resolución de los ejercicios indicados en el guion:
  - El fichero fuente y el HTML del notebook Python con la realización de la práctica, y todo el código debidamente documentado.
  - Todos los ficheros auxiliares necesarios.
- La entrega se calificará sobre un máximo de 10 puntos. En la evaluación tendremos en cuenta:
  - La claridad y calidad de las descripciones, explicaciones y justificaciones en el informe.
  - Calidad de los modelos construidos y claridad del código desarrollado.
  - La calificación de la entrega tiene dos componentes:
    - **C1 (70%):** Una componente común idéntica para todos los miembros del equipo, que se corresponde con la parte conjunta del trabajo, la memoria escrita y el código desarrollado. En esta parte se valorarán por separado el proceso seguido y los resultados obtenidos en cada parte de la práctica. El primer ejercicio pesará un 6/10, mientras que el segundo (‘Caso Práctico’) supondrá el 4/10 restante.
    - **C2 (30%):** Una componente individual, que se corresponde con el resultado de las respuestas al test de evaluación de la práctica 2 (que incluirá cuestiones asociadas a los ejercicios realizados en las 2 sesiones interactivas correspondientes a la P2).