Tarea 2

Fecha de entrega: 2 de junio 2023

Profesor: Felipe Tobar
Auxiliares: Catherine Benavides, Camila Bergasa, Víctor Caro,
Camilo Carvajal Reyes, Diego Cortez M., Stefano Schiappacasse

Formato de entrega: Reporte en formato PDF con extensión máxima de 5 páginas, presentando y analizando sus resultados, y detallando la metodología utilizada. Recomendación: Incorpore gráficos para apoyar su análisis y utilice formato doble columna si le falta espacio. Adicionalmente, debe entregar el código generado con la resolución de la tarea. Fije una semilla con su rut sin el dígito verificador para poder replicar sus resultados.

Pregunta teórica - 30 %

(a) (1,5 pts.) Dado un conjunto de datos linealmente separable, SVM busca encontrar un hiperplano que separe las dos clases. Formule el problema de máximo margen y resuélvalo, es decir, encuentre la solución de SVM. Extienda su análisis en el caso que el problema de máximo margen admite datos mal clasificadas, es decir, datos que están en el lado incorrecto del clasificador lineal. Explique cada uno de sus pasos y mencione las diferencias entre ambas soluciones.

(b) (1,5 pts.) SVM: Supongamos que tenemos puntos en $\mathbb{R}^2$ pertenecientes a dos clases dispuestos en un patrón de círculo, donde la clase positiva se encuentra dentro de un círculo y la clase negativa forma un anillo alrededor de la clase positiva. Estas dos clases no son linealmente separables en el espacio bidimensional original. Demuestre que en este caso no linealmente separable, existe un kernel que permite formular el problema como uno linealmente separable.

(c) (1,5 pts.) El kernel polinomial se define de la siguiente manera:

$$K(x_i, x_j) = (\langle x_i, x_j \rangle + c)^d. \tag{1}$$

Demuestre lo anterior usando el kernel polinomial, es decir, demuestre cómo el kernel polinomial puede transformar un conjunto de datos no linealmente separable en un espacio donde las clases si son linealmente separables. Para simplificar su demostración, puede considerar $c = 1$ y $d = 2$.

(d) (1,5 pts.) Muestre que lo anterior se cumple para cualquier $d \in \mathbb{N}$.

Pregunta práctica - 70 %

El objetivo de esta pregunta es implementar diferentes modelos de clasificación vistos en el curso con `Sklearn`, para luego elegir el más adecuado a través de métodos de selección de modelos. Para ello, utilizaremos el dataset `CancerData`, que posee diferentes variables asociadas a tumores. Con estos datos se considerará el problema de clasificación binaria discriminado entre tumores benignos y malignos.

Antes de comenzar a resolver la pregunta cargue los datos y sepárelos de forma aleatoria en subconjuntos de entrenamiento y testeo, usando su rut sin verificador como semilla aleatoria.

1. Árboles de decisión y *Random Forest*

Para esta sección, utilizaremos métodos de árboles para resolver el problema de clasificación de tumores. Para la pregunta b) y c) utilice las variables `perimeter_worst` y `perimeter_mean`, y en la pregunta d), utilícelas todas

(a) (1 pt.) Explique cómo se poda un árbol según el enfoque de `minimal cost complexity`, según lo revisado en clases, y comente en qué consiste el parámetro de complejidad α.

(b) (1 pt.) Cree un modelo de árbol con los parámetros por defecto de `Sklearn`, entrénelo y calcule el *accuracy* sobre el conjunto de entrenamiento y de prueba. Visualice este árbol y comente qué profundidad tiene. Por otro lado, grafique cómo cambia la profundidad (`max_depth`) del árbol y la cantidad de nodos de éste, en función de α. Comente sus resultados.

(c) (1 pt.) Grafique cómo cambia el *accuracy* del modelo, tanto en el conjunto de entrenamiento como en el de prueba, en función del parámetro de complejidad y determine qué valor utilizar, argumentando su respuesta. Fijando este valor, vuelva a graficar el árbol ya podado y comente sus resultados comparándolos con los obtenidos en la pregunta anterior.

(d) (2 pt.) Implemente un *bagging* de árboles, evalué su desempeño respecto a la métrica *accuracy*, en el conjunto de entrenamiento y en el conjunto de prueba, y compárelo con los resultados obtenidos con un modelo *Random forest* (bosque aleatorio). Para esta implementación de *bagging* de árboles, genere tantos conjuntos de datos como árboles vaya a agregar, del tamaño del conjunto de entrenamiento, muestreando de este mismo (recuerde que debe ser con reemplazo). Comente sus resultados e indique cuales son las diferencias entre hacer un *bagging* de árboles y un bosque aleatorio.

(e) (1 pt.) Investigue cómo el método de *Random Forest* hace para estimar la importancia de las distintas variables predictoras sobre la variable de interés, y explique cómo se calcula.

2. Support vector machines

Para el desarrollo de las preguntas de la Sección 2 utilice las variables `perimeter_worst` y `perimeter_mean`.

(a) (1,5 pts.) Considere la formulación por margen suave de SVM lineal e implemente dicho modelo en función del hiperparámetro c. Luego, grafique $||w||$ en función de c_i, con $c_i \in \{10^{-7}, 10^{-6}..., 10^{-1}\}$. Explique porqué cambia $||w||$ al aumentar c_i y analice sus resultados. Luego, grafique la cantidad de vectores de soporte obtenidos en función de c_i y explique sus resultados.

(b) (1,5 pts.) Considere el kernel polinomial $K(x, y) = (1 + x \cdot y)^2$ y el kernel gaussiano $K(x, y) = \exp(-\gamma||x - y||^2)$. Implemente SVM de margen suave, con $c = 1$, para los kernel lineal, polinomial y gaussiano con las variables `perimeter_worst` y `perimeter_mean`. Muestre cómo cambia la frontera de decisión para cada uno de esos casos y comente cuál kernel le parece más adecuado.
Hint: Para visualizar la frontera cree una grilla para predecir cada combinación de la región de decisión.

3. Selección de modelos

(a) (1 pt.) Implemente un método que permita calcular la curva ROC (*Receiver operating characteristic curve*) y explique como interpretarla.

(b) (1 pts.) Grafique las curvas ROC para los modelos de SVM, *Decision tree* y *Random forest* en el conjunto de prueba y comente sus resultados.

(c) (1 pt.) Elija alguno de los modelos experimentados y argumente su decisión. Explique qué ventajas y desventajas observa de utilizar la curva ROC para hacer la elección, y proponga alguna otra métrica de evaluación justificando por qué cree que es una mejor opción que la anterior.