Universidad de los Andes
Facultad de Ingeniería

Universidad de los Andes
MINERÍA DE DATOS (MIIA 4200)
Profesor: Rafael Amaya Gómez
Fecha: 11 de marzo 2024

RECUERDE:
El link para enviar los archivos de soporte en parejas estará habilitado hasta las 11:59p.m. del 22 de marzo, cualquier entrega posterior tendrá una calificación sobre 4.5
La entrega del taller deberá hacerse en un informe autocontenido incluyendo las gráficas, interpretación y conclusiones. Adicionalmente debe entregarse el código utilizado en un archivo R, de considerar necesario comentar todas las partes del código que crean necesarias para el entendimiento de este.

TAREA 2

PUNTO 1 (40%)

Se sabe que las máquinas de soporte vectorial (SVM) se pueden ajustar con un kernel no lineal para resolver un problema de clasificación con una decisión de frontera no lineal. Se quiere revisar qué pasaría si se aborda una regresión logística usando una transformación no lineal de las características. Para ello contemple lo siguiente:

a. Genere un conjunto de datos con $n = 500$ observaciones y $p = 2$ categorías con una frontera de decisión cuadrática usando lo siguiente

$$X_1 \sim Unif(-0.5, 0.5)$$
$$X_2 \sim Unif(-0.5, 0.5)$$
$$y = \begin{cases} 1, & X_1^2 - X_2^2 > 0 \\ 0, & X_1^2 - X_2^2 \leq 0 \end{cases}$$

b. Grafique las observaciones incluyendo colores que diferencien ambas categorías. Utilice como eje x el valor de X_1 y en el eje y el valor de X_2. ¿Qué puede decir sobre la región de clasificación?
c. Ajuste una regresión logística usando los datos generados y X_1 y X_2 como predictores. ¿Qué puede decir sobre la capacidad de predicción?
d. Aplique este modelo a los datos de entrenamiento para obtener una clase predicha para cada observación. Grafique las observaciones usando como colores las clases predichas. ¿Qué forma tiene esta frontera? Comente sus resultados.
e. Ahora ajuste un modelo de regresión logística a los datos usando funciones no lineales de X_1 y X_2 como predictores (por ejemplo, $X_1^2, X_1 * X_2,\ X_2^2$, entre otros) y relacione los resultados del modelo.
f. Aplique el modelo a los datos de entrenamiento para obtener datos predichos para cada observación. Grafique las observaciones usando los colores predichos. Compare la frontera de decisión con la obtenida anteriormente.

g. Ajuste un clasificador de vectores de soporte a los datos de X_1 y X_2 como predictores. Obtenga una predicción para cada observación y su correspondiente figura usando los colores predichos.
h. Ajuste ahora un SVM usando un kernel no lineal de tipo polinomial y radial. Obtenga una clase para cada observación. Grafique la frontera de decisión con base en las clases predichas.
i. Analice los resultados obtenidos en cada inciso. ¿Qué conclusiones obtiene frente a estas herramientas y este tipo de problemas de clasificación?

PUNTO 2 (30%)

Para este punto considere el conjunto de datos de Auto disponible en Bloque Neón.

a. Resuma Cree una variable binaria que tome valores de 1 para los carros que tengan un kilometraje por galón (**`mpg`**) por encima de la mediana y 0 en caso contrario.
b. Ajuste un clasificador de vectores de soporte a los datos con varios valores de costo. Reporte los errores por cros-validación obtenidos con los diferentes parámetros. Comente los resultados.
c. Repita el inciso anterior usando un SVM con un kernel radial y polinomial con diferentes valores de **`gamma`**, **`degree`** y **`cost`**. Comente los resultados obtenidos
d. Haga algunas figuras para soportar sus conclusiones en los incisos b) y c) usando la función **`plot`** y sabiendo que se pueden graficar por parejas de variables usando **`plot(svmfit, dat, x1~x4)`** reemplazando los nombres de x1 y x4 correspondientes.
e. Analice los resultados obtenidos.

PUNTO 3 (30%)

Para este punto considere el conjunto de datos de OJ disponible en Bloque Neón (Paquete ISLR).

a. Cree una muestra de entrenamiento con 800 muestras aleatorias y el resto como muestra de prueba
b. Ajuste un clasificador de vectores de soporte a los datos de entrenamiento usando un costo de 0.01 con $Purchase$ como la variable respuesta y las demás como predictores. Analice los resultados obtenidos
c. ¿Cuáles son las tasas de error de entrenamiento y prueba para este modelo?
d. Seleccione el parámetro óptimo de costo. Considere valores entre 0.01 y 10.
e. Calcule las tasas de error de entrenamiento y prueba usando este parámetro.
f. Repita los pasos anteriores usando un SVM con un kernel radial. Use el valor por defecto de gamma.
g. Repita los pasos anteriores usando un SVM con un kernel polinomial de grado 2
h. De manera general cuál aproximación obtiene mejores resultados. Analice su respuesta.