Guía: Modelos supervisados

En esta guía abordaremos problemas de los modelos supervisados.

1. Explique cuidadosamente las diferencias entre los metodos de clasificación y regresión utilizando KNN para ambos casos.

2. Asuma que tenemos un dataset con cinco predictores:

   - $X_1 = GPA$ (grade point average).
   - $X_2 = IQ$ (coeficiente intelectual).
   - $X_3 =$ Nivel (1 para College y 0 para High School).
   - $X_4 =$ Interacción entre GPA y IQ.
   - $X_5 =$ Interacción entre GPA y Nivel.

   La respuesta es salario inicial post graduación (en miles de dolares). Asuma que utilizamos mínimos cuadrados para entrenar el modelo, y obtenemos $\widehat{\beta}_0 = 50$, $\widehat{\beta}_1 = 20$, $\widehat{\beta}_2 = 0.07$, $\widehat{\beta}_3 = 35$, $\widehat{\beta}_4 = 0.01$, $\widehat{\beta}_5 = -10$.

   a) ¿Que respuesta es correcta, y por qué?
      i) Para un valor fijo de IQ y GPA, los estudiantes graduados de high school, ganan más en promedio, que los estudiantes graduados de college.
      ii) Para un valor fijo de IQ y GPA, los estudiantes graduados de college, ganan más en promedio, que los estudiantes graduados de high school.
      iii) Para un valor fijo de IQ y GPA, los estudiantes graduados de high school, ganan más en promedio, que los estudiantes graduados de college suponiendo que el GPA es lo suficientemente alto.
      iv) Para un valor fijo de IQ y GPA, los estudiantes graduados de college, ganan más en promedio, que los estudiantes graduados de high school suponiendo que el GPA es lo suficientemente alto.

   b) Calcule el valor del salario de un estudiante graduado de college con un IQ = 110 y GPA = 4.

   c) Diga si es verdadero o falso: Dado que el coeficiente resultante de la interacción GPA/IQ es un termino muy pequeño, hay muy poca evidencia de un efecto de interacción. Justifique su respuesta.

3. Responda a continuación las preguntas respecto a LDA y QDA.

   a) Si el límite de desición de Bayes es lineal, ¿podemos esperar que LDA o QDA tengan un mejor rendimiento en el conjunto de entrenamiento? ¿y en el conjunto de prueba?

b) Si el límite de desición de Bayes no es lineal, ¿podemos esperar que LDA o QDA tengan un mejor rendimiento en el conjunto de entrenamiento? ¿y en el conjunto de prueba?

c) En general, si la muestra de tamaño n aumenta, podemos esperar que la precisión de QDA relativa a LDA ¿mejore?, ¿disminuya? o ¿permanecerá igual? ¿Por qué?

d) Verdadero o falso: Incluso si el límite de decisión de Bayes es lineal para un problema dado, probablemente obtendremos un error de prueba superior usando QDA que si usaramos LDA, ya que QDA es lo suficientemente flexible para modelar un límite de desición linear. Justifique su respuesta.

4. Suponga que recolectamos data de estudiantes de estadística con las siguientes variables, X_1 = horas estudiadas, X_2 = GPA de pregrado, y Y = recibieron una A (calificación alta). Al entrenar una regresión logistica obtenemos los siguientes coeficientes, $\widehat{\beta}_0$ = -6, $\widehat{\beta}_1$ = 0.05, $\widehat{\beta}_2$ = 1.

   a) Estime la probabilidad de que un estudiante que estudia por 40 hrs y tiene un GPA de pregrado de 3.5 obtenga una A en la clase.

   b) ¿Cuantas horas necesitaria estudiar el estudiante de la pregunta anterior, para tener un 50 % de probabilidades para obtener una A?

5. Suponga que tenemos un dataset, y lo dividimos en partes iguales tener data de entrenamiento y prueba, de manera de probar dos procedimientos de clasificación.

6. Primero, utilizamos una regresión logistica, la cual nos entrega un error del 20 % en el conjunto de entrenamiento y de 30 % en el conjunto de prueba. Luego utilizamos KNN (con $n = 1$) y obtenemos un error promedio (promediado sobre el conjunto de entrenamiento y prueba) del 18 %. Basados en estos resultados, ¿cual método deberíamos usar para clasificar nuevas observaciones? ¿por qué?