Profesor: Alfredo Alegría J.
Ayudante: Daniela Díaz

Fecha de entrega: 15/10/21

Tarea 1: Aplicaciones de la Matemática en Ingeniería

1 de Octubre de 2021

1. En este problema se analiza la conexión entre la regresión logística y el análisis discriminante lineal. Considere un problema de clasificación binaria. El *log-odd* se define como

$$\log\left(\frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})}\right).$$

a. Calcule el *log-odd* para el modelo de regresión logística.
b. Muestre que en el caso del análisis discriminante lineal el *log-odd* se puede escribir en el formato $\alpha_0 + \boldsymbol{\alpha}^\top \boldsymbol{x}$, para algún $\alpha_0 \in \mathbb{R}$ y $\boldsymbol{\alpha} \in \mathbb{R}^d$. Determine explícitamente α_0 y $\boldsymbol{\alpha}$.
c. ¿Qué se puede concluír? ¿Cuáles son las diferencias fundamentales entre ambos métodos?

2. La siguiente tabla muestra un conjunto de datos que contiene 6 observaciones, 3 covariables y 1 variable respuesta cualitativa:

X_1	X_2	X_3	Y
0	3	0	Rojo
2	0	0	Rojo
0	1	3	Rojo
0	1	2	Verde
−1	0	1	Verde
1	1	1	Rojo

Suponga que queremos usar estos datos para predecir Y cuando $X_1 = X_2 = X_3 = 0$ usando el método de K vecinos cercanos.
a. Calcule la distancia Euclideana entre cada observación y el punto $X_1 = X_2 = X_3 = 0$.
b. ¿Cuál es la predicción con $K = 1$?
c. ¿Cuál es la predicción con $K = 3$?
d. Si en este problema la frontera de decisión del clasificador de Bayes es altamente no-lineal, ¿se espera que el *mejor* valor de K sea grande o pequeño?

3. Una estación de radio clasificará a sus auditores en jóvenes (0) o adultos (1) a partir de sus gustos musicales. Para llevar a cabo este proceso, se realizó una encuesta a 10 auditores, donde cada uno ha manifestado si le agradan o no ciertos grupos. Los datos se encuentran disponibles en el archivo `gustos_musicales.txt` en AULA.
a. Ajuste un modelo de Bayes ingenuo a este conjunto de datos. Reporte las distribuciones marginales estimadas para cada covariable. Comente los resultados.
b. Responda la encuesta con sus gustos musicales. De acuerdo al modelo ajustado en el inciso **a.**, ¿a cuál de las dos clases pertenece?

4. Considere el conjunto de datos de **factores de riesgo coronario** en tres regiones rurales de Sudáfrica (ver el archivo `datos_heart_disease.txt` disponible en AULA). Lleve a cabo un proceso de selección de variables siguiendo los siguientes pasos:
i. Ajuste un modelo de regresión logística con las 7 covariables que se señalan a continuación
† presión sanguínea sistólica (`sbp`)
† tabaco acumulado (`tobacco`)
† colesterol unido a lipoproteínas de baja densidad (`ldl`)

- † historial familiar de problemas cardiacos (`famhist`)
- † obesidad (`obesity`)
- † consumo actual de alcohol (`alcohol`)
- † edad (`age`)

ii. Elimine la covariable menos significativa basándose en el test de Wald.

iii. Vuelva a ajustar un modelo de regresión logística con las covariables restantes y nuevamente elimine la menos significativa.

iv. Repita este procedimiento hasta que no se puedan quitar más covariables del modelo.

Responda las siguientes preguntas:

a. ¿Cuales son las variables que sobrevivieron a este proceso?

b. ¿De qué otra manera se podría realizar un proceso de selección de variables?