Profesor: Dr. Oldemar Rodríguez Rojas
Análisis de Datos 1
Fecha de Entrega: Viernes 14 de junio - 8 a.m.
Instrucciones:

- Las tareas deben ser subida la Aula Virtual antes de las 8:00 a.m. Luego de esta hora pierde 10 puntos y cada día de retraso adicional perderá 10 puntos más.
- Las tareas son estrictamente individuales.
- Tareas idénticas se les asignará cero puntos.
- Todas las tareas tienen el mismo valor en la nota final del curso.
- Cada día de entrega tardía tendrá un rebajo de 10 puntos.

Tarea Número 11

- **Pregunta 1:** [25 puntos] Complete las demostraciones de los Teoremas 2 y 4 de la presentación de la clase.
- **Pregunta 2:** [25 puntos] Diseñe un algoritmo en pseudocódigo para el Método del Análisis Discriminante Lineal según la teoría vista en clase.
- **Pregunta 3:** [25 puntos] Programa en **R** el algoritmo diseñando en el ejercicio anterior, incluya el gráfico del plano principal y el círculo de correlaciones. Compare los resultados con respecto a la función `train.lda(...)` el paquete `traineR`, para esto use los archivos de datos `Ejemplo_AD.csv` y `tumores.csv` ¿Se obtienen los mismos resultados?
- **Pregunta 4:** [25 puntos] En este ejercicio se generalizan los conceptos de Inercia Total, Inercia Inter-Clases e Inercia Intra-Clases presentados en el curso al caso matricial (en el curso se presentan para el caso de un vector).

  Se consideran p variables continuas (variables explicativas) $\mathbf{x}^1, \ldots, \mathbf{x}^p$ observadas en una muestra Ω de n individuos. Cada individuo $i \in E$ se identifica con su vector (fila) de mediciones en $\mathbb{R}^p, \mathbf{x}_i^t = (x_{i1}, \ldots, x_{ip})$ y cada variable $\mathbf{x}^j$ con su vector (columna) de valores asumidos $\mathbf{x}^j = (x_{1j}, x_{2j}, \ldots, x_{nj})^t$. La variable cualitativa $\mathbf{y}$ (a explicar) determina una partición $P = \{C_1, \ldots, C_r\}$, del conjunto de individuos Ω en r grupos.

  Se denota como:

  - $\mathbf{X}$ la matriz de tamaño $n \times p$ la cual se supone centrada en sus columnas. Como es usual sus columnas son las variables explicativas $\mathbf{x}^j$ (previamente centradas) y los individuos $\mathbf{x}_i^t$ son sus filas.
  - $\mathbf{D} = \text{diag}\,(p_i)$ es la matriz de pesos del conjunto de individuos Ω.
  - A cada clase C_s se le asigna el peso q_s y centro de gravedad $\mathbf{g}_s$ para $s = 1, \ldots, r$ donde:

$$q_s = \sum_{i \in C_s} p_i \quad \text{y} \quad \mathbf{g}_s = \frac{1}{q_s} \sum_{i \in C_s} p_i \mathbf{x}_i.$$

  Se escribe $\mathbf{D}_q = \text{diag}\,(q_j)$ la matriz diagonal de los pesos de las r clases.

- Se denota como $\mathbf{C}_g$ la matriz cuyas filas son los centros de gravedad $\mathbf{g}_s^t$.

Como se supone que las variables son centradas entonces el centro de gravedad del conjunto de todos los individuos Ω es $\mathbf{g} = \mathbf{0}$ y la matriz de covarianza (total) $\mathbf{V}$, de las p variables calculadas sobre Ω es:

$$\mathbf{V} = \mathbf{X}^t\mathbf{D}\mathbf{X} = \sum_{i=1}^{n} p_i \mathbf{x}_i \mathbf{x}_i^t = \sum_{s=1}^{r} \sum_{i \in C_s} p_i \mathbf{x}_i \mathbf{x}_i^t$$

Sea $\mathbf{V}_s$ la matriz de covarianza de las p variables, calculada sobre los individuos de la s-ésima clase:

$$\mathbf{V}_s = \frac{1}{q_s} \sum_{i \in C_s} p_i \left(\mathbf{x}_i - \mathbf{g}_s\right) \left(\mathbf{x}_i - \mathbf{g}_s\right)^t .$$

El promedio de estas matrices se define como la matriz de covarianza de todas las clases y se denomina matriz de covarianza intra-clase y se denota como $\mathbf{V}_W$:

$$\mathbf{V}_W = \sum_{s=1}^{r} q_s \mathbf{V}_s = \sum_{s=1}^{r} \sum_{i \in C_s} p_i \left(\mathbf{x}_i - \mathbf{g}_s\right) \left(\mathbf{x}_i - \mathbf{g}_s\right)^t .$$

Finalmente la matriz $\mathbf{V}_B$ de covarianza correspondiente a las p variables calculadas sobre los centros de gravedad, se denomina matriz de covarianza inter-clase, la cual es igual a:

$$\mathbf{V}_B = \sum_{s=1}^{r} q_s \mathbf{g}_s \mathbf{g}_s^t = \mathbf{C}_g^t \mathbf{D}_q \mathbf{C}_g.$$

Con las definiciones anteriores pruebe lo siguiente: Si $\mathbf{V}, \mathbf{V}_B, \mathbf{V}_W$ son las matrices de covarianza total, inter-clase intra-clase, respectivamente, entonces:

1. $\mathbf{V} = \mathbf{V}_B + \mathbf{V}_W$
2. $\sum_{s=1}^{r} q_s \mathbf{g}_s = 0$. Es decir, rang $(\mathbf{C_g}) \leq r - 1$
3. rang $(\mathbf{C}_{\mathrm{g}})$ = rang $(\mathbf{V}_B)$

Además, para la tabla de datos `Ejemplo_AD.csv` calcule: $\mathbf{g}_A$, $\mathbf{g}_B$, $\mathbf{g}_C$, $\mathbf{V}$, $\mathbf{V}_B$, $\mathbf{V}_W$ y verifique que $\mathbf{V} = \mathbf{V}_B + \mathbf{V}_W$.