Profesor: Dr. Oldemar Rodríguez Rojas
Análisis de Datos 2
Fecha de Entrega: Domingo 15 de octubre a las 12 media noche
Instrucciones:

- Las tareas deben ser subida la Aula Virtual antes de las 6:00pm. Luego de esta hora pierde 20 puntos y cada día de retraso adicional perderá 20 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 20 puntos.

# Tarea Número 8

- **Pregunta 1:** [10 puntos] Complete las demostraciones de los Teoremas 2 y 4 de la presentación de la clase.
- **Pregunta 2:** [10 puntos] Replique en Python la presentación desarrolla en R en el archivo `AnalisisDiscriminate_2022.html`.
- **Pregunta 3:** [20 puntos] Diseñe un algoritmo en pseudocódigo para el Método del Análisis Discriminante Lineal según la teoría vista en clase. Luego agregue a la clase `Analisis_Predictivo`, desarrollada en Python, métodos para el algoritmo diseñado anteriormente, también incluya métodos para el gráfico del plano principal y del círculo de correlaciones. Compare los resultados con respecto a usar `modelo = lda` en Python, para esto use el archivo de datos `Ejemplo_AD.csv`.
- **Pregunta 4:** [20 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 $\Omega$ 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 $\Omega$ 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.
  - $\mathbf{D} = \operatorname{diag}(p_i)$ es la matriz de pesos del conjunto de individuos $\Omega$.
  - $A$ 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 $\mathbf{D}_q = \operatorname{diag}(q_j)$ 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 $\Omega$ es $\mathbf{g} = \mathbf{0}$ y la matriz de covarianza (total) $\mathbf{V}$, de las $p$ variables calculadas sobre $\Omega$ 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. $\displaystyle\sum_{s=1}^{r} q_s \mathbf{g}_s = 0$. Es decir, rang $\left(\mathbf{C}_{\mathbf{g}}\right) \leq r - 1$
3. rang $\left(\mathbf{C}_{\mathrm{g}}\right) =$ rang $\left(\mathbf{V}_B\right)$

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 $\mathbf{V} = \mathbf{V}_B + \mathbf{V}_W$.

- **Pregunta 5:** [20 puntos] La tabla de datos `novatosNBA.csv` contiene diferentes métricas de desempeño de novatos de la NBA en su primera temporada. Para esta tabla, las 21 primeras columnas corresponden a las variables predictoras y la variable Permanencia es la variable a predecir, la cual indica si el jugador permanece en la NBA luego de 5 años. La tabla contiene 1340 filas (individuos) y 21 columnas (variables), con la tabla realice lo siguiente:

  1. Use `LDA` y `QDA` en Python para generar un modelo predictivo para la tabla `novatosNBA.csv` usando el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing. Obtenga los índices de precisión e interprete los resultados.

2. Construya un DataFrame que compare los modelos generado en el ítem anterior respecto los modelos de la tareas anteriores para la tabla `novatosNBA.csv`. Para esto en cada una de las filas debe aparecer un modelo predictivo y en las columnas aparezcan los índices *Precisión Global*, *Error Global*, *Precisión Positiva (PP)* y *Precisión Negativa (PN)*. ¿Cuál de los modelos es mejor para estos datos?

- **Pregunta 6:** [20 puntos] Este conjunto de datos es originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. El objetivo del conjunto de datos es predecir de forma diagnóstica si un paciente tiene diabetes o no, basándose en determinadas medidas de diagnóstico incluidas en el conjunto de datos. El conjunto de datos tiene 390 filas y 16 columnas:

  - **X:** Id del paciente.
  - **colesterol:** Colesterol en mg/dL.
  - **glucosa:** Glucosa en mg/dL.
  - **hdl_col:** Lipoproteínas (colesterol bueno).
  - **prop_col_hdl:** Proporción del colesterol entre el hdl.
  - **edad:** Edad del paciente.
  - **genero:** Género del paciente.
  - **altura:** Altura en pulgadas del paciente.
  - **peso:** Peso en libras del paciente.
  - **IMC:** índice de masa corporal.
  - **ps_sistolica:** Presión arterial sistólica.
  - **ps_diastolica:** Presión arterial diastólica.
  - **cintura:** Longitud de la cintura en pulgadas.
  - **cadera:** Longitud de la cadera en pulgadas.
  - **prop_cin_cad:** Proporción de la longitud de la cintura entre la longitud de la cadera.
  - **diabetes:** Diagnóstico de la diabetes.

  Realice lo siguiente:

  1. Cargue en Python la tabla de datos `diabetes.csv`.
  2. Use `LDA` y `QDA` en `Python` para generar un modelo predictivo para la tabla `diabetes.csv` usando el 75 % de los datos para la tabla aprendizaje y un 25 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías. ¿Son buenos los resultados? Explique.
  3. Construya un DataFrame que compare los modelos generado en el ítem anterior respecto a los modelos generados en las tareas anteriores para la tabla `diabetes.csv`. Para esto en cada una de las filas debe aparecer un modelo predictivo y en las columnas aparezcan los índices *Precisión Global*, *Error Global*, *Precisión Positiva (PP)* y *Precisión Negativa (PN)*. ¿Cuál de los modelos es mejor para estos datos?
  4. Repita el ítem 2, pero esta vez seleccione 6 variables predictoras ¿Mejora la predicción?