

Se debe entregar un informe .Rmd en formato pdf con la resolución y resultados del ejercicio, incluyendo todos los gráficos que crean pertinentes y el archivo .Rmd donde se realizaron los cálculos y se programó la implementación del análisis pedido. El trabajo se debe realizar en grupos de 2 integrantes. En todos los archivos que se entreguen, el nombre del archivo debe incluir los apellidos de los integrantes del equipo y el número de grupo. Notar que en el campus está disponible el .tex de la parte teórica, por si les facilita tener el código fuente para escribir la resolución. La entrega estará habilitada hasta las 23:59 hs del 7 de diciembre de 2025.

1. Teórico

Consideremos un vector aleatorio (\mathbf{x}, Y) , donde $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$ es el vector de covariables e Y es la clase, la cual toma valores en $\mathcal{Y} = \{0, 1\}$. Un clasificador g es una función $g : \mathcal{X} \rightarrow \mathcal{Y}$. Cuando observamos un nuevo \mathbf{x} predecimos la clase como $g(\mathbf{x})$.

- (a) Consideramos el **Error de Clasificación Medio** del clasificador g definido como

$$L(g) = \mathbb{P}(g(\mathbf{x}) \neq Y).$$

Probar que dada la naturaleza binaria de Y , el **Error de Clasificación Medio** coincide con el Error Cuadrático Medio habitual, es decir

$$L(g) = \mathbb{E}((Y - g(\mathbf{x}))^2).$$

- (b) Supongamos que las distribuciones condicionales son normales multivariadas, es decir, $\mathbf{x}|Y=1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ y $\mathbf{x}|Y=0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Probar que la regla óptima resulta

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & \text{si } r_1(\mathbf{x}) \leq r_0(\mathbf{x}) + 2 \log \frac{\pi_1}{\pi_0} + \log \frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} \\ 0 & \text{en c. c.} \end{cases},$$

donde para $i = 0, 1$ se tiene que $\pi_i = P(Y=i)$, $r_i(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$ y $|\boldsymbol{\Sigma}_i|$ es la notación para indicar el determinante de la matriz $\boldsymbol{\Sigma}_i$.

- (c) Probar que si $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$, entonces la regla del ítem anterior resulta

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & \text{si } D_1(\mathbf{x}) - 2 \log(\pi_1) \leq D_0(\mathbf{x}) - 2 \log(\pi_0) \\ 0 & \text{en c. c.} \end{cases}$$

donde $D_i(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$.

(d) Comprobar que si $\pi_0 = \pi_1$, entonces la regla del ítem anterior resulta

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & \text{si } D_1(\mathbf{x}) \leq D_0(\mathbf{x}) \\ 0 & \text{en c. c.} \end{cases}$$

es decir: se clasifica a \mathbf{x} en la población cuya media está más cerca en distancia de Mahalanobis. Si $\Sigma = \sigma^2 \mathbb{I}_p$ (siendo \mathbb{I}_p la identidad de $p \times p$), ¿a qué equivaldría esta regla?

(e) Mostrar que si Σ_0 y Σ_1 no coinciden, entonces la regla óptima puede escribirse como

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{x}^t \mathbf{A} \mathbf{x} + \mathbf{b}^t \mathbf{x} + c \leq 0 \\ 0 & \text{en c. c.} \end{cases}$$

y hallar las expresiones de \mathbf{A} , \mathbf{b} y c .

2. Práctico

Parte A

En esta actividad vamos a trabajar con los datos disponibles en `data_wdbc_X.csv`, que contiene 569 observaciones que corresponden a variables relacionadas con imágenes de lesiones de pacientes con diferentes pronósticos de cáncer de mama (más información acá).

Consigna 1

Hacer un análisis de componentes principales a partir de los datos estandarizados. Analizar la proporción de varianza explicada por las primeras componentes. Hacer un scree-plot. Luego, graficar las observaciones: en el espacio de las primeras dos componentes principales, y en el espacio de las primeras tres.

Consigna 2

Aplicar un clustering por k -means con $k = 2$ clusters usando:

- (a) las primeras dos componentes;
- (b) las primeras tres componentes.

Visualizar los resultados del clustering en:

- el gráfico de las primeras dos componentes coloreando los puntos según el cluster asignado por k -means del inciso (a).
- el gráfico de las primeras tres componentes coloreando los puntos según el cluster asignado por k -means del inciso (b).

Evaluar si $k = 2$ es una cantidad razonable de grupos en todos los casos.

Parte B

Para lo que sigue, vamos a trabajar con los datos disponibles en `data_wdbc_y.csv` que contiene el diagnóstico de cada una de las observaciones del dataset `data_wdbc_X.csv`. Este diagnóstico es “B” o “M”, según si el tumor fue diagnosticado como benigno o maligno, respectivamente.

Consigna 3

Para el clustering hecho sobre las dos primeras componentes:

- (a) Representar en gráficos separados las observaciones coloreadas según `diagnosis` (B/M) y cluster (1/2). En particular, observar qué tipo de frontera parece separar a los datos según `diagnosis`.
- (b) Construir una tabla de contingencia entre `diagnosis` (B/M) y cluster (1/2).
- (c) Suponiendo que cada cluster representa la clase mayoritaria dentro de él, calcular el porcentaje de observaciones correctamente “clasificadas” por el clustering¹. ¿Qué tan bien el “clustering ciego” separa benignos de malignos?
- (d) Repetir para el clustering hecho sobre las tres primeras componentes y comparar los resultados.

Consigna 4

En lo que sigue, se busca construir modelos supervisados que sí usen el diagnóstico. Para ello, fijar una semilla y dividir los datos en 80%-20% para los grupos de entrenamiento y testeo.

- (a) Hacer PCA sobre el conjunto de entrenamiento y clasificar por LDA usando **solo** las primeras 2 componentes principales. Calcular el *accuracy* en el conjunto de testeo.
Importante: al calcular los scores en el conjunto de testeo, tener en cuenta que deben usarse las componentes principales calculadas a partir del conjunto de entrenamiento. ¿Por qué? Explicar brevemente.
- (b) Repetir lo anterior usando:
 - **solo** las primeras 5 componentes principales;
 - **solo** las primeras 10 componentes principales.
- (c) Comparar los tres modelos (2, 5 y 10 componentes) en términos de precisión y complejidad.
- (d) Variar el número de componentes principales entre las primeras 2 a las primeras 30. Graficar la precisión del modelo en función de las componentes. ¿Cuántas componentes elegiría para un modelo final que balancee simplicidad y rendimiento? (*Puede ser útil evaluar cuán sensible es este análisis según la partición train/test que se tenga. Se pueden probar otras semillas y ver cómo cambia este gráfico.*)

¹Notar que las comillas aluden a que, en realidad, se agrupó sin tener en cuenta las verdaderas clases. Es decir, sin tener información de la variable `diagnosis`.