**Introducción a la Estadística y Ciencia de Datos - Segundo cuatrimestre 2022**

**Trabajo Práctico - Clasificación**

---

**Este trabajo práctico es de carácter obligatorio, y la nota formará parte de la calificación final de la materia. Se debe entregar un informe en formato pdf con la resolución y resultados del ejercicio, incluyendo todos los gráficos que crean pertinentes, y el archivo .R donde se realizaron los cálculos y se programaron las funciones que se piden. El trabajo se puede realizar en grupos de hasta 3 integrantes.**

# Teórico

Consideremos un vector aleatorio $(\mathbf{X}, Y)$, donde $\mathbf{X} \in \mathcal{X} \in \mathbb{R}^p$ 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 $g : \mathcal{X} \to \mathcal{Y}$. Cuando observamos un nuevo $\mathbf{x}$ predecimos la clase como $g(\mathbf{x})$.

Consideramos el **Error de Clasificación Medio** del clasificador $g$ definido como

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

1. Encontrar la regla de clasificación óptima $g^{op}$ que minimiza el Error de Clasificación Medio.

2. Respecto de las distribuciones condicionales, supongamos que $\mathbf{X}|Y = 1$ tiene densidad $f_1$ y $\mathbf{X}|Y = 0$ tiene densidad $f_0$. Respecto de las marginales, sean $\pi_1 = \mathbb{P}(Y = 1)$ y $\pi_0 = \mathbb{P}(Y = 0)$. Hallar una expresión para $g^{op}$ en términos de las densidades condicionales $f_0$ y $f_1$ y de $\pi_1$ y $\pi_0$.

3. 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

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

   donde $r_i(x) = (\mathbf{x} - \boldsymbol{\mu}_i)^t\, \boldsymbol{\Sigma}_i^{-1}\, (\mathbf{x} - \boldsymbol{\mu}_i)$ para $i = 0, 1$.

# Práctico

En el archivo *lluviaAus.csv* se encuentran datos acerca de numerosos factores climáticos en Australia, a partir de diversas estaciones de clima, entre otras cosas. El objetivo es decidir si puede crearse algún modelo de clasificación con estas variables para poder predecir si va a llover al día siguiente.

La variable objetivo **RainTomorrow** significa: ¿llovió al día siguiente? Y clasifica como Sí o No.

1. Cargar los datos de `lluviaAus.csv`. En una primera etapa exploratoria de los datos usar la función ggpairs para inspeccionar la relación entre las variables. ¿Qué sugiere este gráfico? Verificar que las variables tengan el tipo correcto (de haber variables categóricas, transformarlas en factor).

2. Realizar un plot de las variables **Sunshine** (nivel de radiación solar por sobre un valor límite) vs. **Humidity3pm** (humedad a las 3 pm) pintando de diferente color según la variable **RainTomorrow**. ¿Qué se observa en este gráfico? ¿Da la impresión de que las dos variables tienen la misma capacidad predictiva?

3. Realizar boxplots paralelos para la variable **Sunshine** clasificando por la variable **RainTomorrow**. ¿Qué se observa? Repetir con la variable **Humidity3pm**.

4. Implementar una función `clasificador.movil(datos,etiquetas,h,x0)` que clasifique a un punto `x0` de acuerdo a la regla de clasificación de promedios móviles teniendo en cuenta a la variable clasificadora `datos`= $(x_1, \ldots, x_n)$, a las `etiquetas`=$(y_1, \ldots, y_n)$ y al valor de la ventana `h`.

5. Determinar $h_{vc}$ la ventana óptima para el clasificador basado en promedios móviles que brinda el método de validación cruzada.

6. Utilizando el set de datos, calcular el error de clasificación empírico mediante validación cruzada del clasificador, si se desea clasificar solamente usando la variable **Sunshine**.