El principal objetivo de esta guía es introducir los primeros contenidos de clasificación e integrarlos con conocimientos y trabajos prácticos previos: ***vamos a mezclar todo en la coctelera!****. Para eso, vamos a volver los datos de alturas. En este caso la idea es predecir a partir del dato de altura la de un individuo su género implementando distintos plug–in en la regla óptima de Bayes.*

1. La base

1. Descargar del campus virtual el conjunto de datos, compuesto por $n = 500$ observaciones y que está disponible en el archivo `alturas_n_500.csv`. Estos datos van a jugar el rol de *muestra de entrenamiento* y solo usaremos las variables `altura` y `genero` (codificada como F o M).

   Leer los datos en `R` y realizar un gráfico que permita visualizar la relación entre estas dos variables.

2. El cuerpo: Métodos de Clasificación

2.1. Discriminativos

2. Con la *regla de la mayoría* vamos a *aprender* a clasificar el género de un individuo como femenino (1) o masculino (0) cuando su altura es $x = 165$.

   a) En primer lugar, mediante el método de **vecinos más cercanos**. Para ello, considerar los $k = 10$ vecinos más cercanos y calcular la proporción de 1's. Según este método, ¿cómo se clasificaría al género de un nuevo individuo con altura igual a 165 cm, F o M?

   b) En segundo lugar, mediante el método de **proporciones locales**. Para ello, considerar una ventana de tamaño $h = 1{,}5$ alrededor del punto de interés y calcular la proporción de 1's. Según este método, ¿cómo se clasificaría al género de un nuevo individuo con altura igual a 165 cm, F o M?

   (Hint: Se puede resolver "a mano" o bien utilizar las funciones programadas en la guía de Actividades de Clase - Unidad 5.)

3. Repetir el punto anterior con $x = 175$.

2.2. Generativo

Recordemos que cuando la covariable X es continua, la regla óptima de Bayes puede escribirse como:

$$g^{opt}(x) = \begin{cases} 1 & \text{si } f_1(x)\mathbb{P}(Y=1) > f_0(x)\mathbb{P}(Y=0) \\ 0 & \text{si } c.c. \end{cases}$$

donde f_0 es la densidad de $X|Y=0$ y f_1 la de $X|Y=1$. Como hemos mencionado en clase, en los métodos generativos la regla de Bayes se estima reemplazando las densidades f_0 y f_1 y la probabilidad $\mathbb{P}(Y=1)$ por sus estimadores.

4. Realizar un histograma de las alturas para cada género y superponer la curva de la densidad estimada usando el núcleo normal y la ventana óptima de CV.
¿Qué relación guardan estas curvas con las densidades f_0 y f_1?

5. Estimar $\mathbb{P}(Y=0)$ y $\mathbb{P}(Y=1)$ a partir de tus datos. Recordar que definimos Y como 0 si el individuo es masculino y 1 si es femenino.

6. Haciendo un plug–in en $g^{opt}(x)$ con las estimaciones de $f_1(x)$, $f_0(x)$ calculadas en el ítem 4 y las estimaciones de $\mathbb{P}(Y=0)$ y $\mathbb{P}(Y=1)$ calculadas en 5, clasificar el género de un individuo como femenino (1) o masculino (0) cuando su altura es $x=165$.

7. Repetir el punto anterior para $x=175$.

3. El aditivo aromático: Elección del parámetro de suavizado

8. Para el método de **vecinos más cercanos**, hallar el k óptimo por el método de Validación Cruzada Leave One Out (LOOCV) realizando la búsqueda en una grilla de valores entre 3 y 20. Comparar la pérdida de LOOCV evaluada en el k óptimo hallado con la evaluada en $k=10$ y realizar un gráfico de k vs. $CV(k)$.

9. Para el método de **proporciones locales**, hallar el h óptimo por el método LOOCV realizando la búsqueda en una grilla de valores entre 3 y 5 con paso 0,05. Comparar la pérdida de LOOCV evaluada en el h óptimo hallado con la evaluada en $h=1{,}5$ y realizar un gráfico de h vs. $CV(h)$.

10. Para el método **generativo**, hallar los h_0 y h_1 óptimos por el método LOOCV realizando la búsqueda en una grilla de valores entre 0,01 y 2 con paso 0,05 para h_0 y h_1.

4. A batir!

11. Ahora vamos a testear las reglas y compararlas entre sí:

    En el archivo alturas_testeo.csv se encuentran 34 datos de altura que separamos para testear cómo funcionan los tres métodos implementados. Para ello, aplicar a este conjunto de datos cada una de las tres reglas implementadas en base a la muestra de entrenamiento usando en cada caso el/los parámetro/s de suavizado óptimo/s, calculados en la sección 3. Obtener el Error de Clasificación de testeo de cada clasificador (es decir la proporción de observaciones mal clasificadas) sobre estos datos. ¿Cuál de ellas clasifica mejor?

5. Bonus Track

Leyendo el fondo de la copa...

12. Graficar un vector xNuevo (en el eje de abscisas) tomando valores entre 160 y 170 con un paso de 0.01 y en el eje de ordenadas el valor con el que clasifica a cada valor de xNuevo el método de vecinos más cercanos con el k óptimo hallado en 8. Interpretar el criterio con el que clasifica esta regla.

    (*Hint: usar, para clasificar, los datos de la muestra de entrenamiento*)

13. Repetir el ítem anterior con los métodos de promedios locales y generativo y superponer con otro color al gráfico anterior. Interpretar y comparar el criterio con el que clasifica cada regla.

6. Bonus del bonus

14. Sean Y una v.a. dicotómica que toma los valores 0 y 1 y X una variable aleatoria discreta con rango R_X. Dado un clasificador $g : R_X \to \{0, 1\}$, probar que

$$\mathbb{P}(g(X) \neq Y) = \mathbb{E}\{Y - g(X)\}^2$$

    Es decir que el error de clasificación coincide con el error cuadrático medio.