3. El aditivo aromático: Elección del parámetro de suavizado

A partir de estos enunciados se trabajará con el conjunto de datos `alturas_n_490.csv` que se encuentran disponibles para su descarga en el campus.

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 1,5 y 12 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 1 y 10 con paso 0,5 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 **??**. 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.