Análisis de datos con Python
Ejercicios Parte 2 - Clase 3 - UNSAM 2018
Training y Testing - K-folding y Cross Validation

1. Genere muestras:
   (a) Considere una variable $x \sim U(0, 10)$ y genere $n = 100$ muestras de x.
   (b) Considere $y = 3x + 2 + \varepsilon$ donde $\varepsilon \sim N(0, 1)$ y genere 100 muestras a partir de las anteriores.
2. Evalúe
   (a) Seleccione el 80% de los pares (x, y) en un conjunto $TRAIN$ y el 20% restante en $TEST$.
   (b) Use los datos en $TRAIN$ para ajustar regresiones lineales de grado $0, 1, \ldots, 25$ y el otro conjunto para evaluar el error cuadrático medio.
   (c) Qué modelo da mejor?
   (d) Repita 100 veces la generacion de muestras y selección de modelo: Da siempre el mismo resultado?
3. Parta el conjunto original de 100 pares en $k \in \mathbf{N}$ *folds*.
   (a) Seleccione un fold F y use los demás para ajustar una regresión y el fold F para evaluar el error cometido E_F.
   (b) Repita para regresiones lineales de diferente grado, como antes.
   (c) Calcule el error medio $\sum_F \frac{E_F}{k}$ cuando F recorre todos los folds.
   (d) Para $k = 5$, ¿qué modelo da mejor?
   (e) Repita 100 veces la generacion de muestras y selección de modelo: Da siempre el mismo resultado?
4. En el caso $k = n$ cada fold tiene un solo elemento.
   (a) Reescriba el código de forma de optimizarlo para este caso particular.
   (b) Repita el ejercicio anterior para $k = n$.