Elements of Machine Learning 2025

Tarea 01

22.enero.2025

En esta tarea vamos a ajustar distribuciones a un conjunto de datos, como vimos en clase. También vamos a explorar dos conjuntos de datos: (1) el conjunto de datos *Palmer Penguins*, que ya iniciamos en el aula, y (2) un conjunto de datos *Tips* de propinas en restaurantes.

El conjunto de datos *Palmer Penguins*, se puede cargar desde Python como
`df = sns.load_dataset('penguins')`
o también está disponible en el archivo **penguins.csv**.

El conjunto de datos *Tips*, se puede cargar desde Python como
`df = sns.load_dataset('tips')`
o también está disponible en el archivo **tips.csv**.

1. Utilice el conjunto de datos **data_aula04.csv** que se compartió en la clase anterior. Elija 3 de las variables incluidas en los datos, y para cada una de ellas hacer lo siguiente:

   - Proponer 4 ó 5 distribuciones de probabilidad que traten de modelar los datos de esa variable, y para cada una de esas distribuciones, graficar los gráficos de probabilidad QQ-plot y PP-plot, el contraste entre las densidades teórica y empírica, y el contraste entre las funciones de distribución, para determinar cuál de las distribuciones se acopla mejor a los datos.
   - Realizar una prueba de hipótesis de Kolmogorov-Smirnov para constrastar cada distribución contra los datos empíricos. Ordenar de menor a mayor p-value, y seleccionar la de mayor p-value como el modelo que mejor se ajusta.

   Puede utilizar los archivos **qq-plots.ipynb** y **aula04.ipynb** para inspirarse y tomar ideas de cómo calcular estos gráficos y pruebas de hipótesis.
   (No se vale utilizar el mismo código que les estoy compartiendo).

2. Completar el análisis exploratorio del conjunto de datos *Penguins*. Elaborar gráficas que analicen algunas relaciones entre las variables (gráficas que no hayamos realizado en clase). A partir de sus gráficas, establecer algunas conclusiones sobre el comportamiento de los datos.

3. Realizar un análisis exploratorio para el conjunto de datos *Tips*.

   Mencionar los siguientes aspectos:

   - el tamaño del conjunto de datos,
   - las variables que se incluyen y de qué tipo son,
   - indicar si hay datos faltantes, y hacer algún tratamiento con ellos (*e.g.* removerlos),
   - hacer gráficos o tablas de cada variable categórica para indicar su distribución,
   - hacer histogramas o funciones de densidad de cada variable numérica,
   - cuando considere conveniente, separar estos histogramas por categoría (`hue`),
   - hacer *scatterplots*, *pairplpots* o mapas de calor en 2D, para identificar relaciones entre variables,
   - calcular algunos estadísticos descriptivos.

Pueden usar como material de apoyo https://seaborn.pydata.org/tutorial/distributions.html

Redactar un pequeño informe con sus análisis de datos, destacando sus conclusiones o *insights* más importantes.