Ejercicio 1 Clustering

En este ejercicio trabajaremos con datos de ingresos y gastos de clientes de un shopping. El conjunto de datos se encuentra en el archivo `Mall_Customers.csv` y contiene 200 observaciones de las siguientes 5 variables: ID que es un simple identificador del cliente, Sexo, Edad, Ingreso anual y Puntaje de gastos.

1. Realizar k-medias para agrupar en 5 clusters ($K = 5$), usando solamente las variables `Annual Income (k$)` y `Spending Score (1-100)`. Inicializar el algoritmo 25 veces.

   Con los resultados obtenidos generar un gráfico en donde cada grupo tenga un color distintivo y en el cuál se distingan claramente los centros de los diferentes grupos.

2. Recordar que la variación total dentro de cada cluster la definimos de la siguiente manera:

   $$W = \sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

   en donde x_i denota una observación perteneciente al cluster C_k, μ_k es el promedio de los puntos asignados al cluster C_k y K es el número de clusters. Por ejemplo, en R esta cantidad puede obtenerse con la función `kmeans()` accediendo mediante `$tot.withinss`.

   Hacer un gráfico de W en función de K.

   Para este gráfico usar siempre 25 inicializaciones diferentes.

   En base al gráfico, discutir brevemente si la elección de $K = 5$ clusters es correcta.

3. Agrupar en 5 grupos utlizando k-medias difusas y graficar.

4. Agrupar en 5 grupos nuevamente, pero asumiendo que los datos provienen de una mezcla de gausianas y graficar nuevamente.

5. Comparar los resultados obtenidos en las partes anteriores. Discutir similitudes y diferencias.

Ejercicio 2 Regresión

En este ejercicio analizaremos la base de datos `Credit` que contiene las siguientes variables:

`balance`: Promedio de deuda de las tarjetas de crédito

`age`: Edad

`cards`: Número de tarjetas de crédito

`education`: Años de educación

`income`: Ingresos en miles de dolares

`limit`: Límite de crédito

`rating`: Rango de crédito

`gender`: género

`student`: estado como estudiante

`status`: estado marital

`ethnicity`: grupo étinico: caucásico, afroamericano o asiático

1. Realizar un scatterplot de todos los pares de variables. Justificar a partir de la gráfica por qué podemos eliminar la variable `rating` del juego de datos. Crear un nuevo juego de datos, sin esta variable.
2. Investigar si hay diferencias en el límite de la tarjeta de crédito por grupo étnico, luego hacerlo por género y luego para la interacción género-grupo étinico. Escribir los tres modelos resultantes e interpretar los coeficientes.
3. Dividir el conjunto de datos en un conjunto de entrenamiento y otro de test (80 % y 20 %).
4. Ajustar un modelo lineal en el conjunto de entrenamiento para predecir el `balance` de las tarjetas de crédtios según `ingresos` y si `estudiante` o no. Investigar si es necesario agregar un término de interacción o no. Interpretar el modelo.
5. Graficar `balance` en función de `ingresos` e incluir la o las rectas resultantes del ajuste. Diferenciar los puntos del conjunto de entrenamiento de los puntos del conjunto de test.
6. Calcular el porcentaje de varianza explicada por el modelo.
7. Calcular el Error Cuadrático Medio en el conjunto de test.
8. Proponer un modelo para predecir balance, teniendo en cuenta todas las variables que le parezcan necesarias. Justificar.
9. Calcular el Error Cuadrático Medio en el conjunto de test y comparar con el error obtenido en la parte 7.