Práctica: Clustering

1. Implemente una funcion *Kmedias* que dado una matriz X de datos (por fila) y un valor k, efectúe un clustering sobre los datos y devuelva un vector y con el número de clase a la que pertenece.

2. (Simulación "buena")

   (a) Genere 50 datos de la clase A como sigue: $X_i \sim U(-2,-1)$, $Y_i \sim U(1,2)$.

   (b) Genere 50 datos de la clase B como sigue: $X_i \sim N(\mu = 0, \sigma^2 = 1)$, $Y_i \sim N(\mu = 0, \sigma^2 = 1)$.

   (c) Genere 50 datos de la clase C como sigue: $X_i \sim U(1,2)$, $Y_i \sim U(-1,2)$.

   (d) Grafique los tres conjuntos de datos con distintos colores.

   (e) Efectúe un clustering K-medias de tres grupos y grafique los resultados del clustering con distintos colores.

3. (Simulación "mala")

   (a) Genere 100 datos de la clase A como sigue: $\theta_i \sim U(0, 2\pi)$, $\rho_i \sim U(0,1)$. En base a eso, construya (X_i, Y_i) como sigue: $(X_i, Y_i) = (\rho_i cos(\theta_i), \rho_i sen(\theta_i)$.

   (b) Genere 100 datos de la clase B como sigue: $\theta_i \sim U(0, 2\pi)$, $\rho_i \sim U(2, 2.5)$. En base a eso, construya (X_i, Y_i) como sigue: $(X_i, Y_i) = (\rho_i cos(\theta_i), \rho_i sen(\theta_i)$.

   (c) Grafique los dos conjuntos de datos con distintos colores.

   (d) Efectúe un clustering K-medias de dos grupos. Qué se concluye?

4. Considere los datos **productos.txt**. Efectúe un clustering K-medias con dos clusters considerando Precio y Marketing. Considere la conveniencia de hacer un paso previo de estandarización para homogeneizar los datos. Haga gráficos con los clusters obtenidos y otro con los verdaderos grupos.

5. Considere los datos **iris.data**. Efectúe un clustering K-medias con tres clusters para los datos ignorando el tipo de especie. Hay alguna semejanza entre el clustering obtenido y las distintos tipos de flores?

6. (DBSCAN) Implemente una función **dbscan** que dado una matri de datos X, un $\epsilon > 0$ y un $minPts$ entero mayor que cero efectúe un clustering DBSCAN. Debe devolver un vector de índices con la asignación de cada fila de X.

7. Aplique DBSCAN a la simulación del ejercicio 3, para tratar de "reparar" el problema que se tenía con K-medias. Qué pasa al probar el método con la simulación del ejercicio 2?