Universidad de la República **Estadística Multivariada Computacional**
Facultad de Ingeniería - IMERL **Primer Semestre 2020**

Práctico 5: Clustering

1. Pruebe que minimizar $\sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} \|\mathbf{x} - \overline{\mathbf{x}}_k\|^2$ equivale a minimizar $\sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{\mathbf{x}_i, \mathbf{x}_{i'} \in C_k} \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2$.

2. Vamos a trabajar sobre los datos MNIST (disponibles en http://yann.lecun.com/exdb/mnist/). Esta base de datos contiene un conjunto de 60000 imágenes (train) y un segundo conjunto de 10000 imágenes (test). Cada imagen tiene un tamaño de 28x28 y representa un número de manuscrito. Nuestro objetivo será lograr la agrupación en la base de datos train.

   a) Descargue los datos de MNIST de http://yann.lecun.com/exdb/mnist/, así como el archivo para manipular los datos mnist.R .

   b) Descomprima los archivos .gz y, utilizando el archivo mnist.r, cargue los datos en R y visualice una de las imágenes.

   c) ¿Cómo se almacenan las imágenes bajo R?

   d) Comience extrayendo al azar una muestra de 1000 imágenes.

   e) Agrupe esta muestra usando un método jerárquico y k-means. Use una técnica para elegir la cantidad de clusters.

   f) Interprete su clustering representando el promedio de imágenes de cada grupo.

   g) Repita el procedimento anterior con la base de datos completa de 60000 imágenes y con 10 clusters.

   h) Valide los métodos de clustering utilizados con la verdadera etiqueta de cada observación con los indices de Rand, Rand Ajustado, Jaccard y MCE

3. Se consideran los puntos $x_1 = 1, x_2 = 2, x_3 = 9, x_4 = 12$ y $x_5 = 20$.

   a) Aplique el método de k-means y calcule la inercia intra-clusters en los casos siguientes:

      1) $k = 2, \mu_1 = 1, \mu_2 = 20$
      2) $k = 3, \mu_1 = 1, \mu_2 = 12, \mu_3 = 20$
      3) $k = 4, \mu_1 = 1, \mu_2 = 9, \mu_3 = 12, \mu_4 = 20$

   b) Misma pregunta usando un método jerárquico ascendente con la distancia del vecino más cercano y dibujar el dendograma.

4. En este problema, aplique k-means con $k = 2$ sobre las 6 observaciones siguientes: $(1, 4), (1, 3), (0, 4), (5, 1), (6, 2)$ y $(4, 0)$.

   a) Ubique las observaciones en el plano

   b) Asigne aleatoriamente una etiqueta de grupo a cada observación. Puede usar el comando sample () en R para hacer esto. Reportar las etiquetas del grupo para cada observación.

   c) Calcule el centroide para cada grupo.

   d) Asigne cada observación al centroide al que esté más cerca, en términos de distancia euclidiana. Reportar las etiquetas del grupo para cada observación.

   e) Repita (c) y (d) hasta que las respuestas obtenidas dejen de cambiar.

   f) En el plot hecho en (a), coloree las observaciones de acuerdo con las etiquetas de grupo obtenidas.

5. *a*) Retomar el ejemplo pginas 64 y 65 de la presentación, y a partir del mismo, identificar y comentar cada paso del algoritmo Spectral Clustering.

   b) Armar un conjunto de datos como lo de los tres grupos de circunferencias concentrícas de la presentación en la parte de Spectral Clustering. Aplicar k-means y Spectral Clustering y evaluar los resultados obtenidos.

6. Se obtuvieron las particiones siguientes

$$\mathcal{C} = \{1, 1, 2, 2, 2, 1\} \quad y \quad \mathcal{C}' = \{1, 2, 1, 2, 1, 1\}$$

   Calcule el índice de Rand, Rand Ajustado, Jaccard y el MCE entre estas dos particiones.