UFRO
UNIVERSIDAD DE LA FRONTERA
FACULTAD DE INGENIERÍA y CIENCIAS
DEPTO. INGENIERÍA MATEMÁTICA

Instrucciones Tarea Clustering
Módulo Análisis de Datos, EMI2016-1
1^{er} Semestre 2019

1. Para la tarea se ha subido a la plataforma el archivo Wine_Transactions que tiene dos pestañas:

   OfferInformation Describe un total de 32 ofertas realizadas en el año destacando características de las mismas (procedencia de vino, descuento, etc.)

   Transactions Lista 324 transacciones indicando el cliente que realizó la compra y la oferta asociada a la misma.

2. El objetivo de la tarea es agrupar los clientes sobre la base de las compras realizadas para buscar características comunes y preparar ofertas más específicas para estos grupos.

3. Los pasos a seguir son los siguientes:

   i Construir una matriz de incidencia entre clientes y ofertas. Los datos serían los clientes que se caracterizan por un vector de dimensión 32 y componentes binarias.

   ii Correr el algoritmo kmeans para encontrar los clusters. Para ello se puede utilizar en Python el objeto KMeans de sklearn.clusters.

   iii Caracterizar los clusters de clientes resultantes en base a las características de las ofertas más utilizadas en cada cluster.

   iv Identificar el número más adecuado de clusters utilizando el método del codo, y el de silhouette (objeto silhouette_score de sklearn.metrics).

4. Para realizar el paso ii, también se puede resolver directamente el problema de optimización subyacente al método kmeans. A saber (para K clusters): Dados $x^1, \ldots, x^N \in R^n$ calcular los centros $\mu^1, \ldots, \mu^K \in \mathbb{R}^n$ como solución del siguiente problema:

$$\min_{\mu} \sum_{j=1}^{N} \min_{k=1,\ldots K} \|x^j - \mu^k\|_2^2$$

   donde $\|.\|_2^2$ se refiere al cuadrado de la norma euclideana

$$\|x^j - \mu^k\|_2^2 = \sum_{i=1} n(x_i^j - \mu_i^k)^2$$