Aprendizaje No Supervisado

Maestría en Ciencia de Datos

Lucas Fernández Piana

Primavera 2022

Universidad de San Andrés

k-means

K-means

Sean $A = \{x_1, \ldots, x_n\}$ un conjunto de puntos de $\mathbb{R}^p$ y sea $\mathcal{C} = \{C_1, \ldots, C_K\}$ una partición de A.

Llamaremos dispersión intra grupo a

$$W(C_k) = \frac{1}{2N_k} \sum_{\substack{i \in C_k, \\ j \in C_k}} ||x_i - x_j||_2^2,$$

y dispersión total intra grupos a

$$W(\mathcal{C}) = \sum_{k=1}^{K} W(C_k).$$

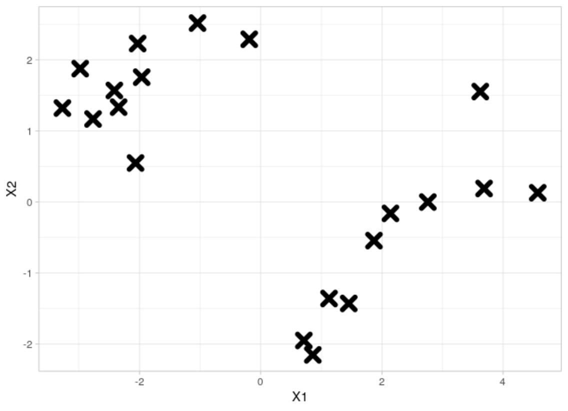


Figura 1: Mal elegida la partición aumenta W

Proposición: sean $\{x_1, \ldots, x_n\}$ un conjunto de puntos que viene en $\mathbb{R}^p$ y llamemos $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Entonces,

$$\frac{1}{2n}\sum_{j=1}^n\sum_{i=1}^n ||x_i - x_j||_2^2 = \sum_{i=1}^n ||x_i - \overline{x}||_2^2.$$

El criterio en el que se basa K-means para definir los clusters es particionar A en $\{C_1, \ldots, C_K\}$ para que sea mínima la dispersión total intra grupos.

Es decir, encontrar

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \bar{x}_k||_2^2.$$

K-means

Calcular todas las dispersión intra grupos para todas las particiones es computacionalmente muy costoso!

K-means

Calcular todas las dispersión intra grupos para todas las particiones es computacionalmente muy costoso!

Pensemos desde un ángulo diferente,

$$\min_{\mathcal{C},m} G(\mathcal{C}, m), \text{ donde}$$

$$G(\mathcal{C}, m) = G(\mathcal{C}, m_1, \ldots, m_k) = \sum_{k=1}^{K} \sum_{C(i)=k} ||x_i - m_k||_2^2.$$

Existen técnicas de optimización que permiten encontrar una solución (no siempre óptima) para este tipo de problemas de optimización.

Proposición: sea $S = \{x_1, \ldots, x_N\}$ un conjunto de puntos en $\mathbb{R}^p$, entonces

$$\bar{x} = \arg \min_{m \in \mathbb{R}^p} \sum_{i=1}^{N} ||x_i - m||_2^2 \tag{1}$$

Observación: en palabras simples dicimos que $\bar{x}$ es el punto que está a menor distancia del resto.

Fijar la cantidad de grupos de antemano, K. Comenzar con una partición inicial, puede ser tomada al azar.

Pasos

1. Dada una partición $\mathcal{C}$ calcular $\overline{x}_i$ para todo $1 \leq i \leq K$.
2. Crear una nueva partición, $\mathcal{C}'$, asignando cada observación a la media más cercana. Es decir,

$$i \in C_l' \text{ si } ||x_i - \overline{x}_l||_2^2 = \min_{1 \leq k \leq K} ||x_i - \overline{x}_k||_2^2.$$

3. Repetir (1) y (2) hasta que $W(C)$ no varie de paso a paso.

Observemos que el algoritmo es descendiente, en este sentido,

Comienzo con una partición inicial $\mathcal{C}^0$, aplico el paso (1), tomo $m^1 = (\bar{x}_1, \ldots, \bar{x}_K)$:

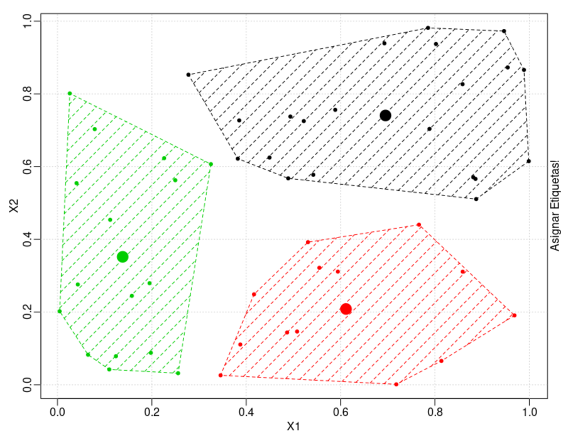
$$G(\mathcal{C}^0, m) \geq G(\mathcal{C}^0, m^1).$$

Aplico el paso (2), es decir, construyo la nueva partición $\mathcal{C}^1$:

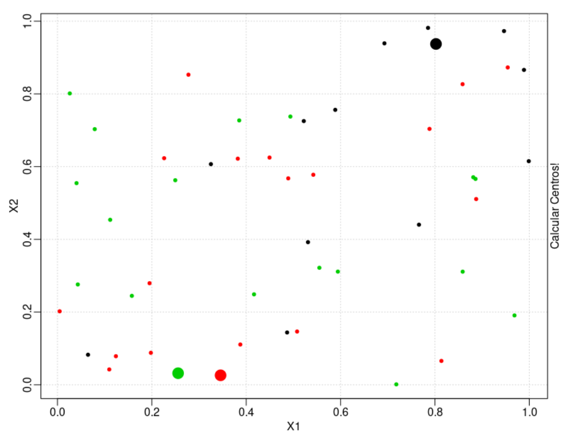
$$G(\mathcal{C}^0, m^1) \geq G(\mathcal{C}^1, m^1).$$

Vuelvo a aplicar el paso (1), ...

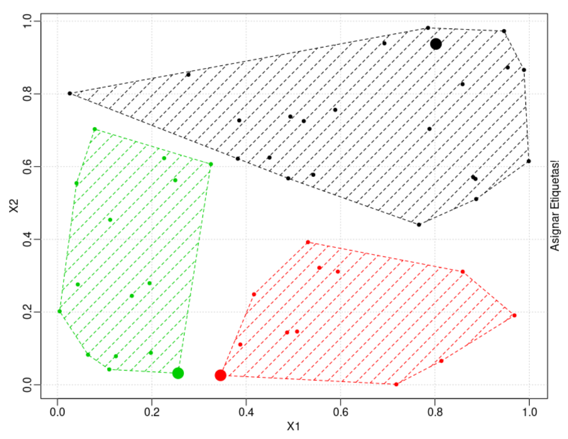
K-means Algoritmo



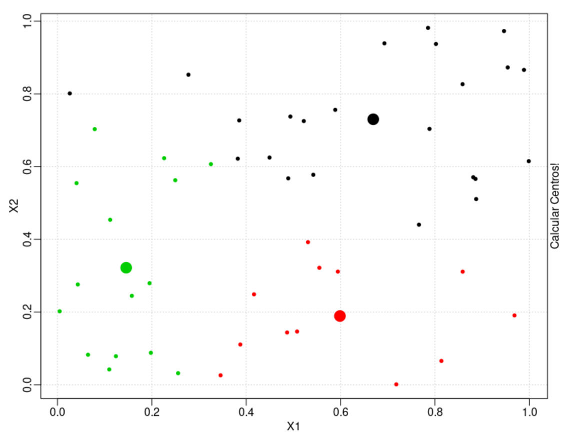
K-means Algoritmo



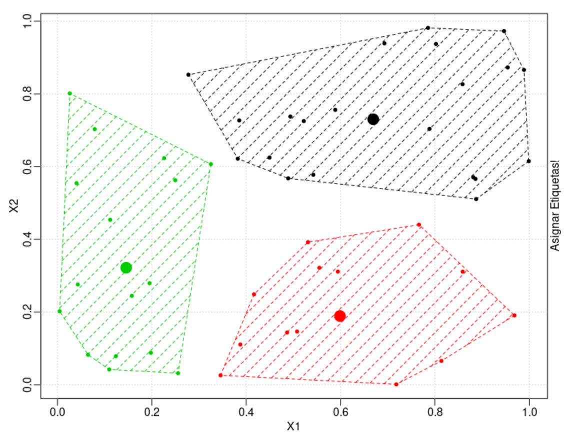
K-means Algoritmo



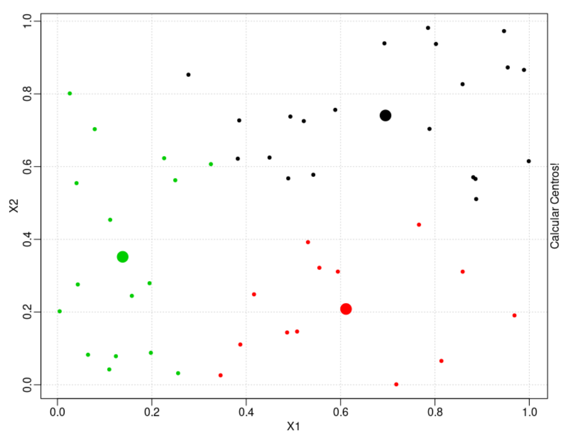
K-means Algoritmo



K-means Algoritmo



K-means Algoritmo



K-means

K-means es uno de los métodos de clustering más utilizados.

Debemos tener en cuenta que el método supone lo siguiente:

- Las observaciones provienen de variables cuantitativas.
- El número de clusters o particiones es asignado de antemano (K).
- La métrica que consideramos es la distancia Euclidea al cuadrado, es decir,

$$d_2^2(x, y) = ||x - y||_2^2 = \sum_{j=1}^{p} (x_j - y_j)^2$$

K-medoids

Dado un conjunto de puntos $S = \{x_1, \ldots, x_n\}$ en $\mathbb{R}^p$ y una métrica d, decimos que x_0 es un centroide de S si cumple,

$$x_0 = \arg\min_{x \in S} \sum_{i=1}^{n} d(x_i, x).$$

¿Quién es el centroide de S si d es la métrica euclídea?

K-medoids es una generalización de K-means, lo único que se modifica es el cálculo del centroide.

Escribamos el método:

1-

2-

3-

K-medoids

Pasos

1. Dada una partición $\mathcal{C}$ calcular m_i el centroide de cada grupo para todo $1 \leq i \leq K$.
2. Crear una nueva partición, $\mathcal{C}'$, asignando cada observación al cenotride más cercano. Es decir,

$$i \in C_l' \ \text{ si } \ d(x_i, m) = \min_{1 \leq k \leq K} d(x_i, m_k).$$

3. Repetir (1) y (2) hasta que $W(C)$ no varie de paso a paso, donde $W(C_k) = \sum_{i=1}^{n} d(x_i, m_k)$.

COFFEE BREAK!