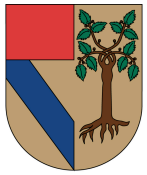


UNIVERSIDAD
PANAMERICANA

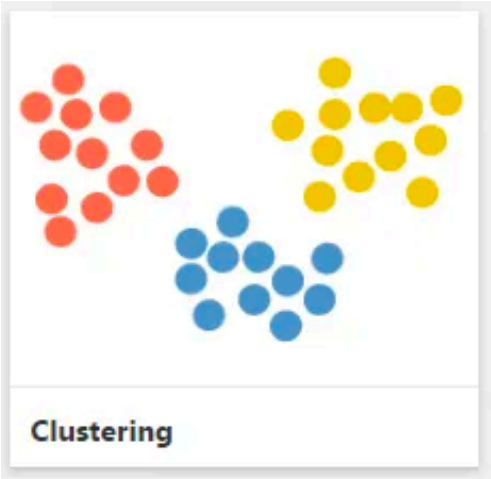
Aprendizaje No Supervisado

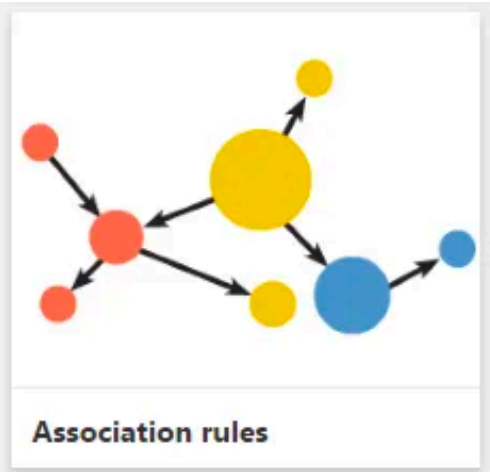
Aprendizaje no supervisado

Los métodos de **aprendizaje no supervisado** permiten encontrar patrones u organización de datos, a partir de las características de los mismos y sin emplear alguna salida etiquetada.

El aprendizaje no supervisado resuelve dos tipos de problemas:

- **Problema de agrupación:** grupos
- **Problema de asociación:** relaciones o dependencias





Dagdoo.org: Clustering Visual Power BI, 2020.

Aprendizaje no supervisado

El **problema de agrupación** se refiere a encontrar k grupos de un conjunto de datos dado, donde los elementos de dichos grupos compartan una similitud.

Dos tipos de métodos de agrupación:

- **Aprendizaje no paramétrico:** grupos
- **Aprendizaje paramétrico:** estimación de la distribución basada en parámetros fijos

Otra clasificación de los métodos de agrupación:

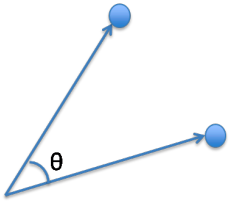
- **Por partición**
- **Superpuestos**
- **Probabilísticos**
- **Jerárquicos**

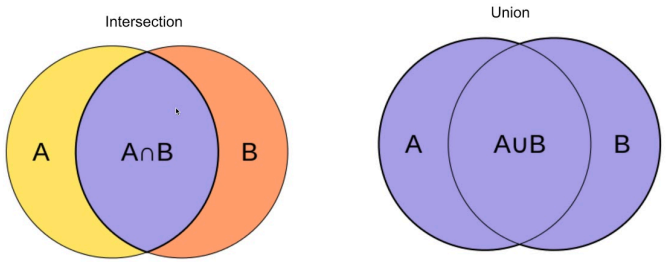
El **problema de agrupación** se refiere a encontrar *k* grupos de un conjunto de datos dado, donde los elementos de dichos grupos compartan una similitud.

Requieren métricas de similitud (o proximidad) para la generación de grupos. *Ejemplos*:

- **Similitud coseno:** en vectores
- **Distancia Jaccard:** en conjuntos
- **Distancia Euclidiana:** en puntos

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$





$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

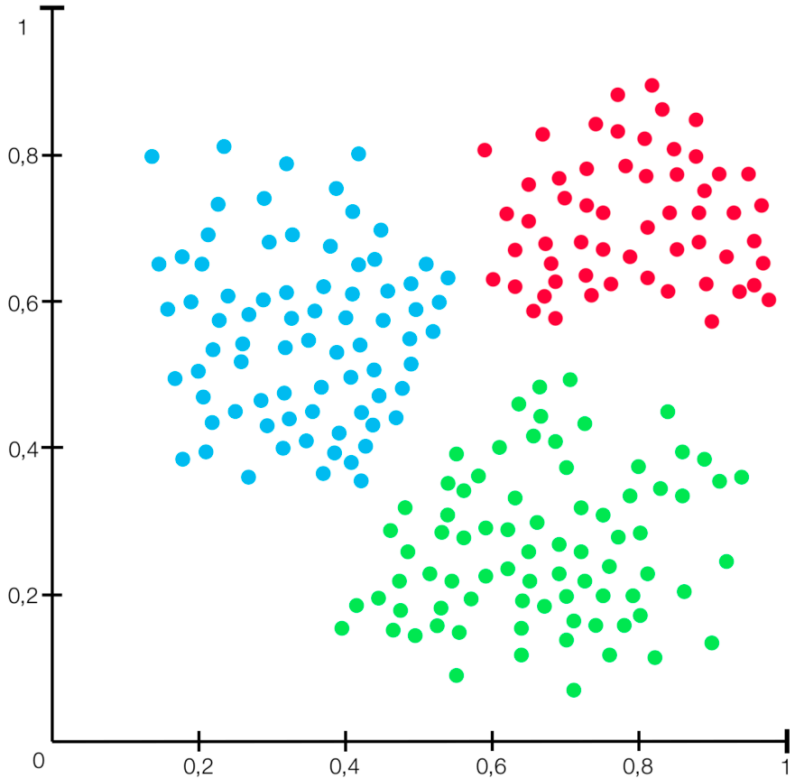
Pragmatic Data Science, 2020.

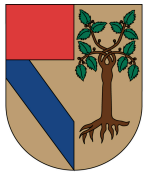
Aprendizaje no supervisado

El **problema de agrupación** se refiere a encontrar *k* grupos de un conjunto de datos dado, donde los elementos de dichos grupos compartan una similitud.

Condiciones para realizar una agrupación:

- Los elementos de un grupo son similares
- Los grupos son disimilares





UNIVERSIDAD
PANAMERICANA

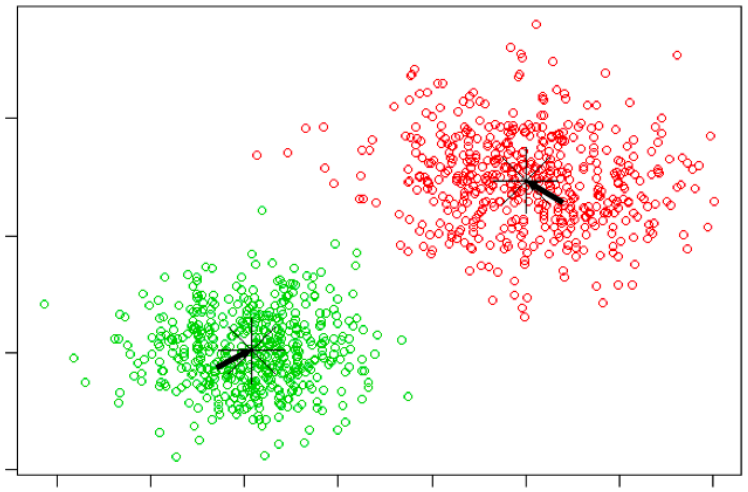
k-medias

El **método *k*-medias (o algoritmo de Lloyd)** es un algoritmo de partición que resuelve el problema de agrupación mediante *k* centroides que definen los grupos en los cuales un conjunto de datos dado es dividido.

Se cumple que cada dato pertenece solo a un grupo, por lo que la intersección de los grupos es vacía:

$$S_i = \{x_p : \|x_p - \mu_i\| \leq \|x_p - \mu_j\|, \forall 1 \leq j \leq k\}$$

Centroides



***k*-medias –** *algoritmo estándar*

1. Inicializar *k* centroides.

2. Mientras no se cumpla una condición de paro:

i. Asignar cada ejemplo del conjunto de datos a un grupo usando una métrica.

ii. Actualizar los *k* centroides:

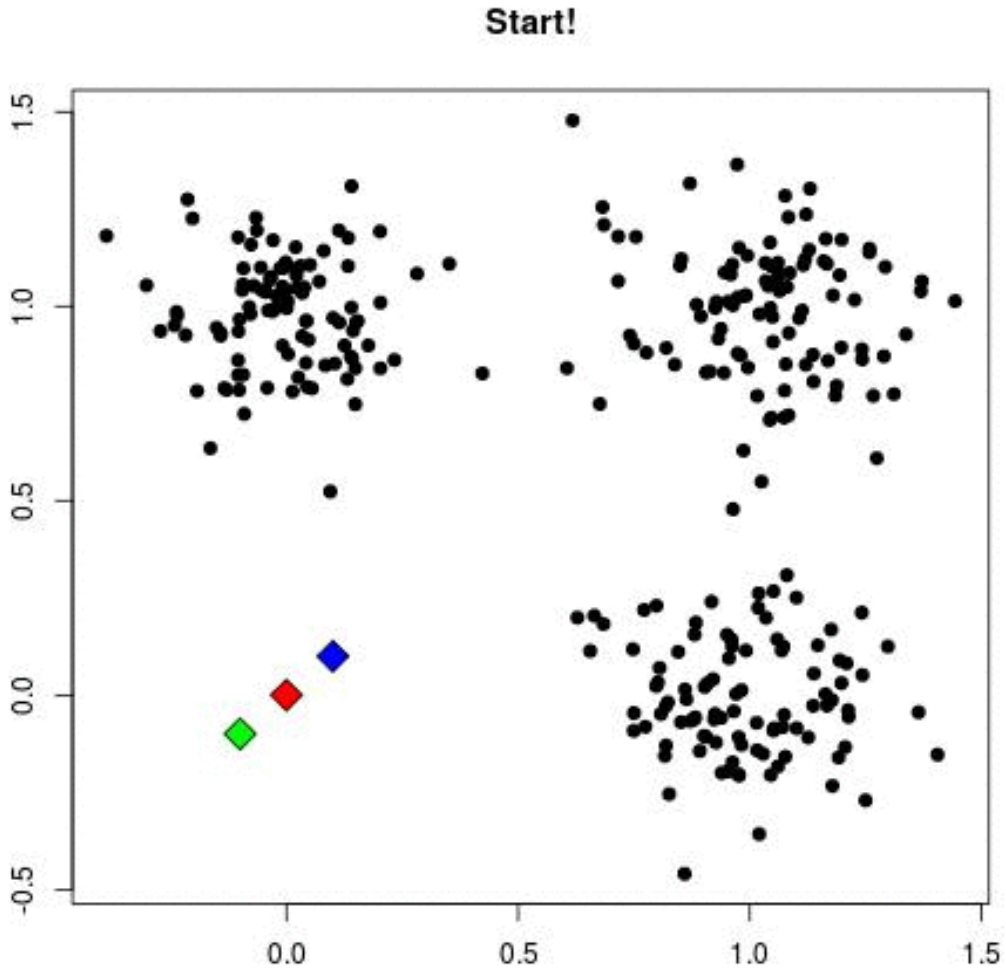
$$\mu_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j$$

iii. Evaluar la función objetivo:

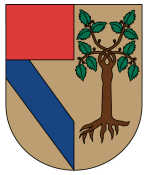
$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \|x_j^{(i)} - \mu_i\|^2$$

3. Regresar los *k* centroides.

***k*-medias –** *algoritmo estándar*



(blank slide)

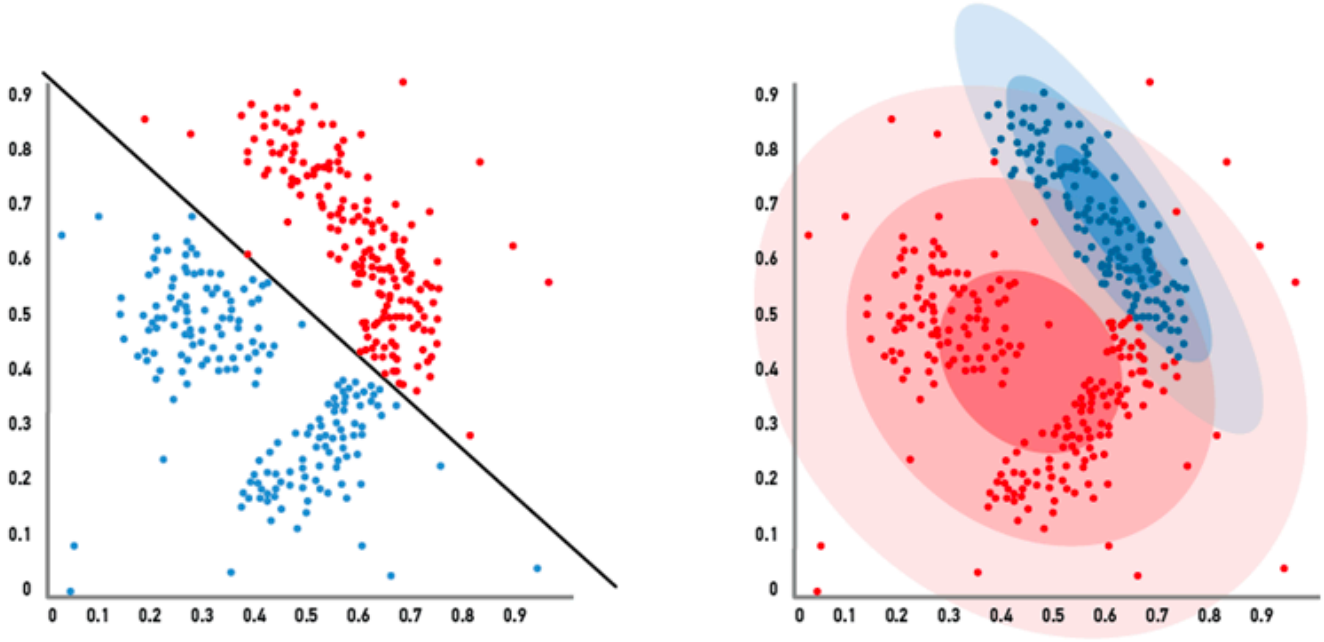


UNIVERSIDAD
PANAMERICANA

Agrupamiento difuso (*fuzzy* c-*means*)

Agrupamiento difuso

El método de **agrupamiento difuso de medias (*fuzzy* c-*means*)** es un algoritmo de superposición que resuelve el problema de agrupación mediante *k* centroides que definen grupos difusos y así designar la pertenencia de cada elemento del conjunto de datos a los *k* grupos creados.



Worldquant, Fuzzy approaches in financial modeling, 2020.

agrupamiento difuso de medias – *algoritmo simple*

1. Inicializar la matriz difusa *U*.

2. Mientras no se cumpla una condición de paro:

i. Calcular los *k* centroides:

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

ii. Evaluar la función objetivo:

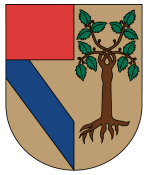
$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} (u_{ij})^m \|x_j - v_i\|^2$$

iii. Actualizar la matriz difusa *U*:

$$u_{ij} = \sum_{c=1}^{k} \left(\frac{\|x_i - v_j\|}{\|x_i - v_c\|} \right)^{-2/(m-1)}$$

3. Regresar los *k* centroides.

(blank slide)



UNIVERSIDAD
PANAMERICANA

Mezclas Gaussianas

Mezclas Gaussianas

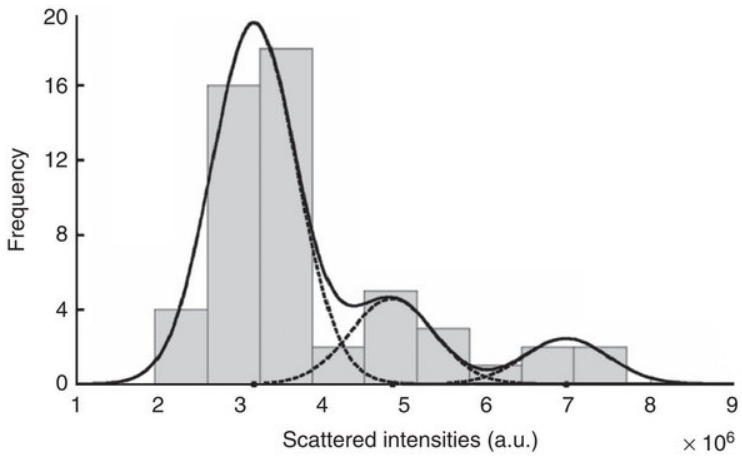
El **modelo de mezclas Gaussianas** es un método probabilístico que resuelve el problema de agrupación mediante la aproximación de la distribución de un conjunto de datos utilizando *k* grupos representados por funciones Gaussianas.

Cada grupo está definido por la media y la desviación estándar (covarianza), también conocidas como variables ocultas, que permiten una interpretación estadística.

El modelo de mezclas Gaussianas se representa como:

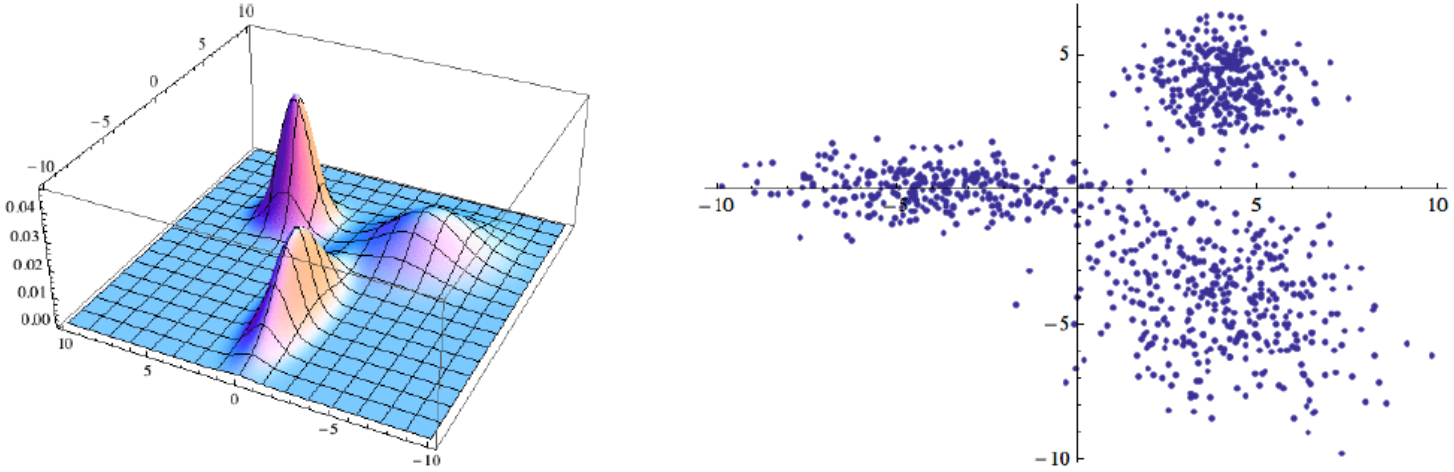
$$p(x_1, ..., x_n|w, \mu, \sigma) = \sum_{i=1}^{M} w_i g(x_1, ..., x_n|\mu_i, \sigma_i)$$

$$g(x|\mu_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp - \frac{(x - \mu_i)^2}{2\sigma_i^2}$$



Mezclas Gaussianas

El **modelo de mezclas Gaussianas** es un método probabilístico que resuelve el problema de agrupación mediante la aproximación de la distribución de un conjunto de datos utilizando *k* grupos representados por funciones Gaussianas.



Fall for Data, Soft Clustering with Gaussian Mixture Models (GMM), 2019.

Mezclas Gaussianas

El **modelo de mezclas Gaussianas** es un método probabilístico que resuelve el problema de agrupación mediante la aproximación de la distribución de un conjunto de datos utilizando *k* grupos representados por funciones Gaussianas.

Se emplea el método de **maximización de la esperanza** para realizar el entrenamiento no supervisado de este modelo, así como de muchos otros y se basa en encontrar los valores de los parámetros internos a partir de un algoritmo iterativo:

- **Asignación (expectativa) –** obtiene las medias de los parámetros ocultos del modelo
- **Maximización –** utiliza los valores estimados para optimizar la función objetivo

(blank slide)