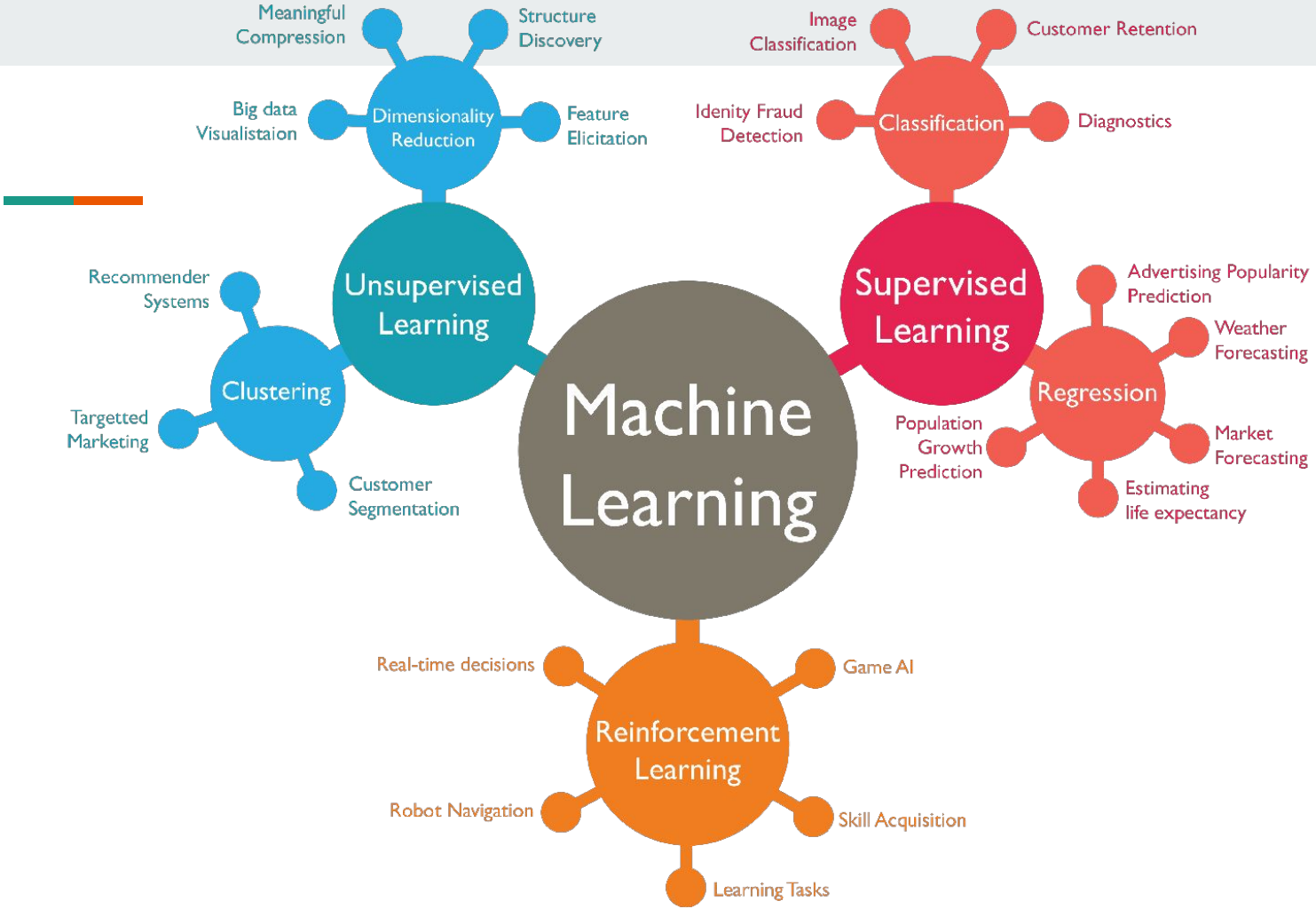
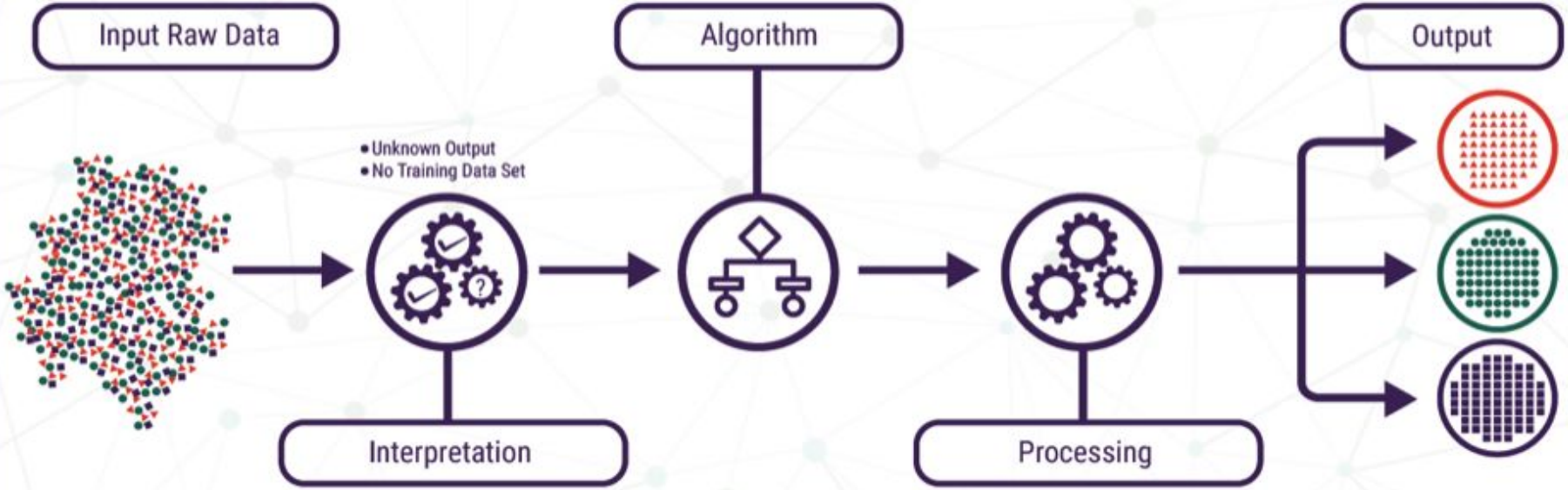
Clustering

PhD.(c) Junior Fabian Arteaga

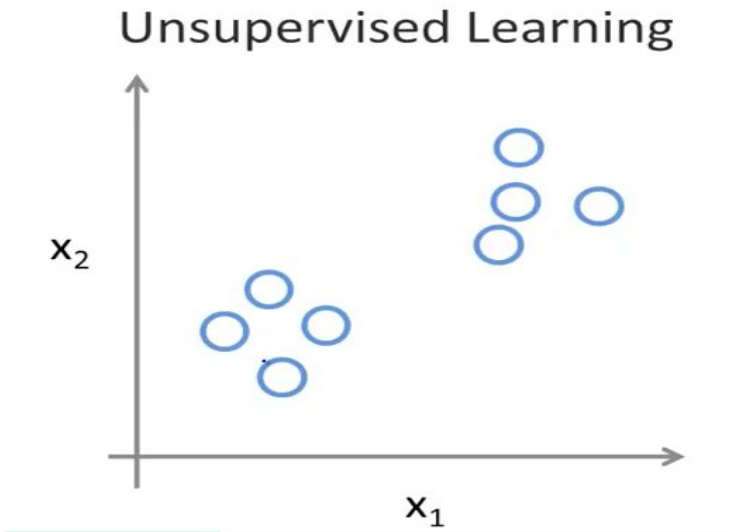

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Estimating life expectancy
Population Growth Prediction
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

UNSUPERVISED LEARNING



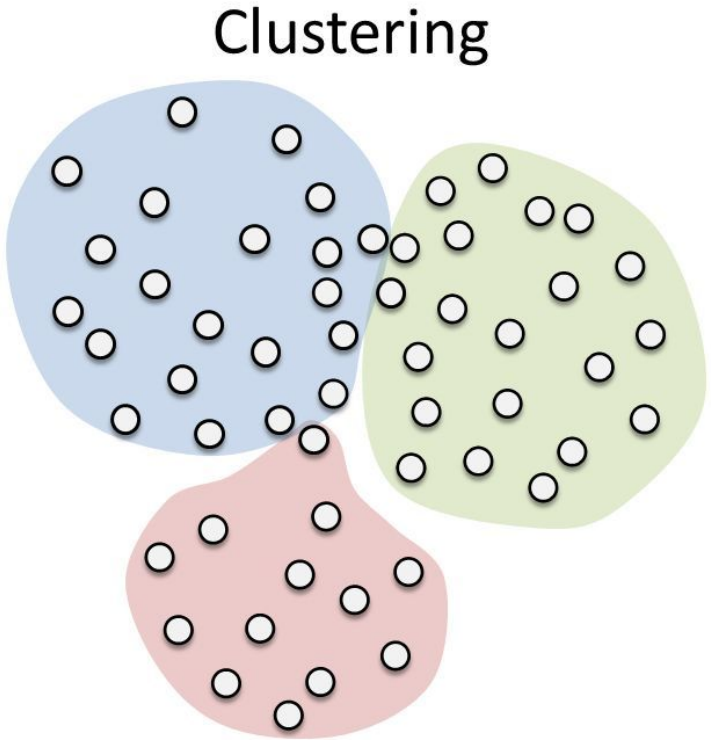
Aprendizaje No Supervisado

Nos permite abordar los problemas con poca o ninguna idea de cómo deben ser nuestros resultados. Podemos derivar la estructura de los datos donde no necesariamente sabemos el efecto de las variables.



Clustering

Clustering es la clasificación de objetos en diferentes grupos, o más precisamente, la *partición de un dataset* en subconjuntos (clusters), de modo que los datos de cada subconjunto "idealmente" *comparten algún característica en común*. Esto en base a una medida de distancia definida.

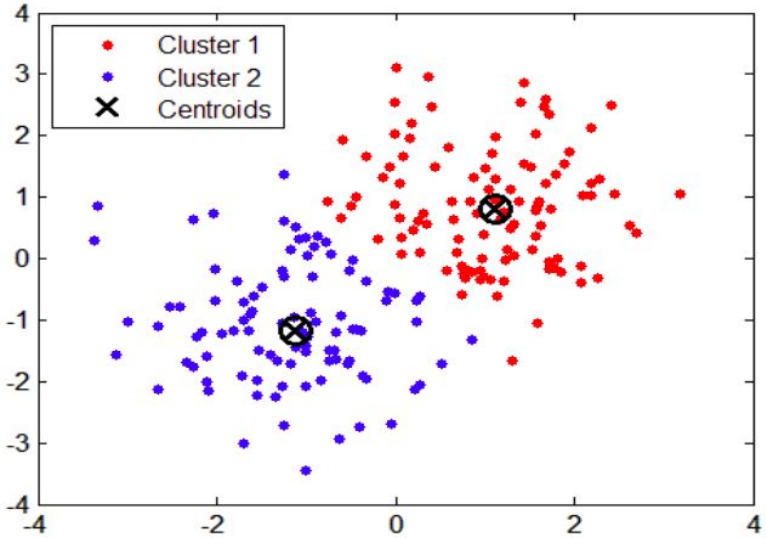


Clustering es un método de clasificación no supervisada: No existen clases predefinidas

Partitional Clustering

Los algoritmos particionales determinan todos los clusters a la vez.

- *K-means y sus variaciones*
- Fuzzy c-means clustering
- QT clustering algorithm



Clustering

- Un buen método de clustering producirá clusters de alta calidad con:
  - *Alta similaridad intra-class*
  - *Baja similaridad inter-class*

- La calidad del resultado de clustering depende de 2 factores importantes: La *medida de similitud* empleada y por la *implementación*.

- La calidad de un método de clustering es medida por su capacidad para descubrir algunos o todos los patrones ocultos en el dataset.

Clustering (*Métricas de Distancias*)

- Las medidas de distancia determinará cómo es calculada la similaridad de dos elementos, además esta medida influye en la forma de los clusters.

- *Distancia Euclidiana* (*2-norm distance*) está dada por:

$$d_{\mathbf{x},\mathbf{y}} = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2}$$

- *Distancia de Manhattan* (1-norm distance) está dada por:

$$\sum_{i=1}^{k} \left| x_i - y_i \right|$$

K-means Clustering

- El algoritmo *K-means* es un algoritmo para agrupar *"n"* objetos en *"k"* particiones, en base en sus atributos. *(k<n)*

- Este algoritmo es el método de partición más simple para el análisis de clusters y es ampliamente utilizado en aplicaciones de data mining.

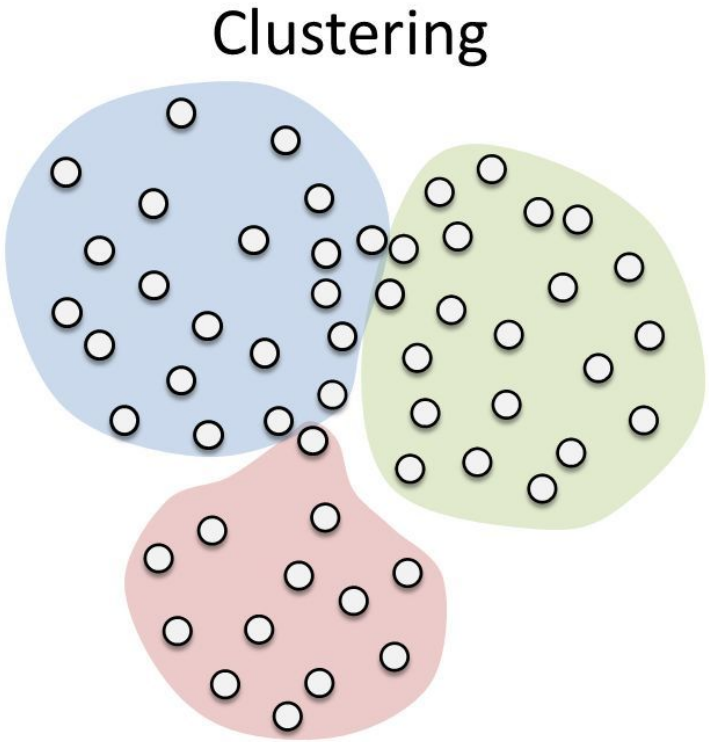
- Cada cluster está representado por el centro del cluster (*centroide*). El algoritmo converge en centroides estables para cada cluster.

K-means Algorithm

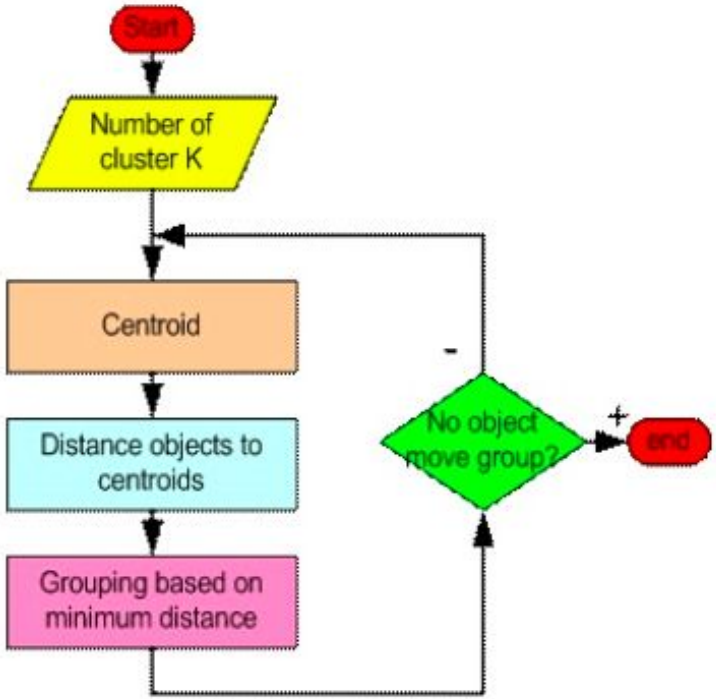
Input: *K*, data

Inicialización: Establecer **k** centroides (randomly)

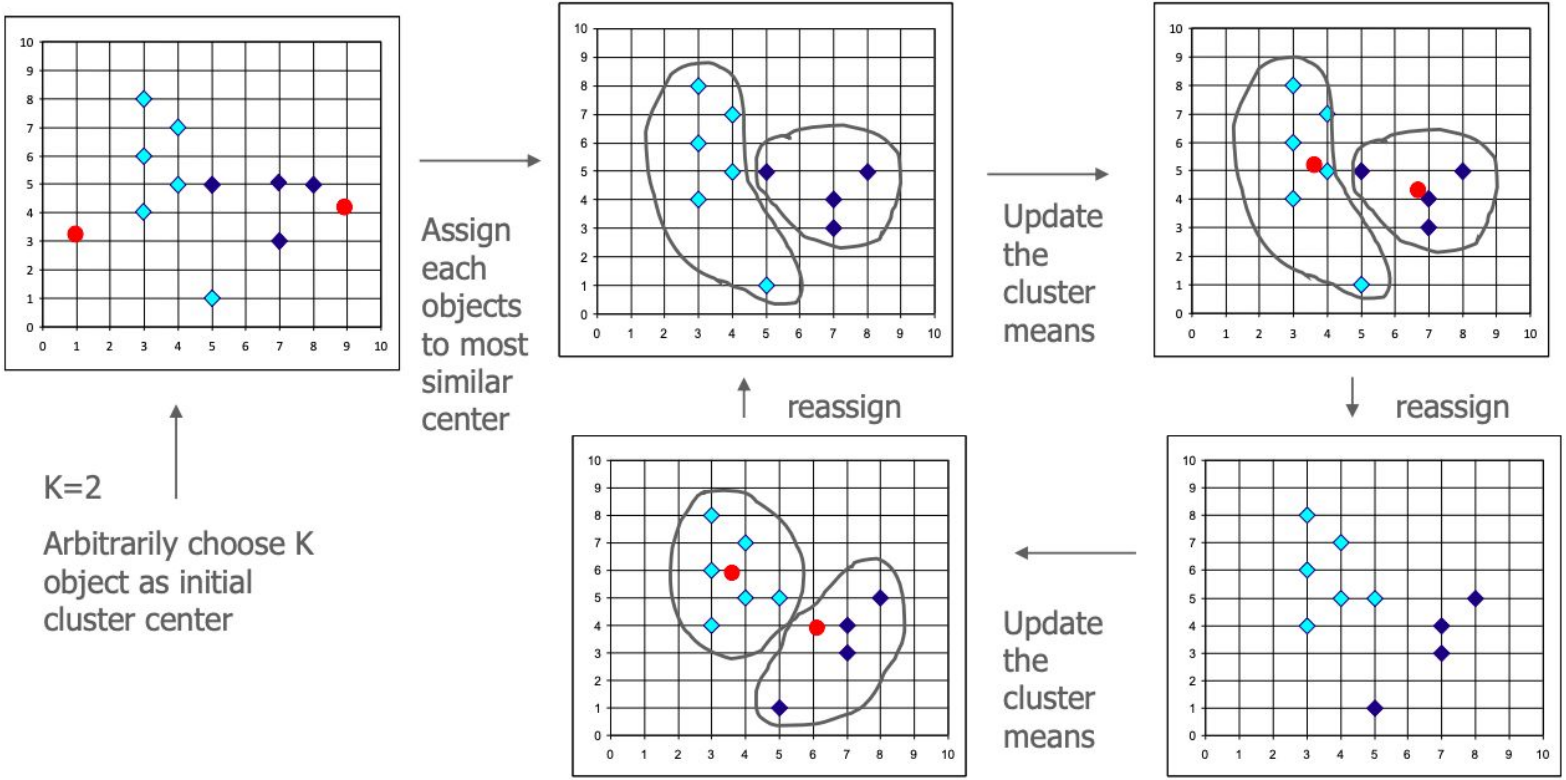
1.) Asignar cada dato al cluster más cercano, según alguna métrica de distancia.
2.) Calcular nuevos centroides de cada cluster. Centroide es el centro del cluster (*mean point*).
3.) Ir al paso 1, detenerse cuando no hay más asignaciones.



K-means Algorithm



K-means Method



K-means Clustering

- Un algoritmo para particionar (o agrupar) ***N*** data points en ***K*** subconjuntos disjuntos S_j, debe minimizar el criterio de la suma de los cuadrados:

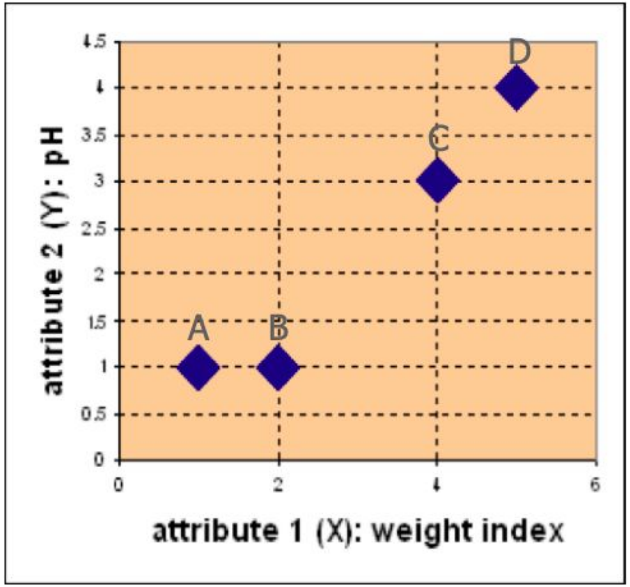
$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \left| x_n - \mu_j \right|^2,$$

- Donde X_n es un vector representando el *n*-ésimo data point y u_j es el centroide geométrico de los data points en S_j.

K-means *Clustering (Ejemplo)*

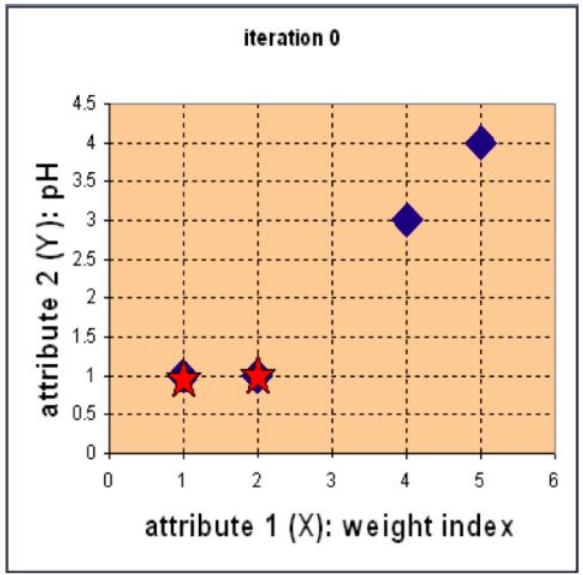
Supongamos que tenemos 4 tipos de Medicina (A, B, C, D) y cada uno tiene 2 atributos (pH, Peso). Nuestro objetivo es agrupar estos objetos en K=2 grupos de medicina.

Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4



K-means Clustering (Ejemplo)

Paso 1: Establecer los seed points (centroides) iniciales



$$c_1 = A, c_2 = B$$

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

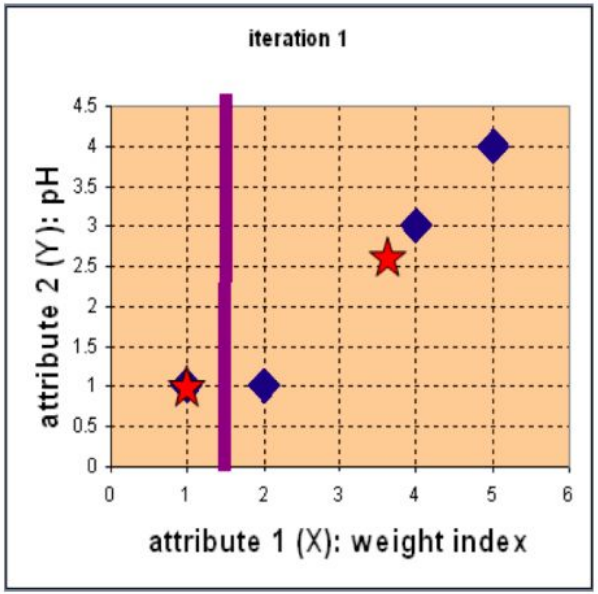
$$d(D, c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$d(D, c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

Asignar cada objeto al cluster con el seed point (centroide) más cercano

K-means Clustering (Ejemplo)

Paso 2: Calcular los nuevos centroides en base a las nuevas particiones



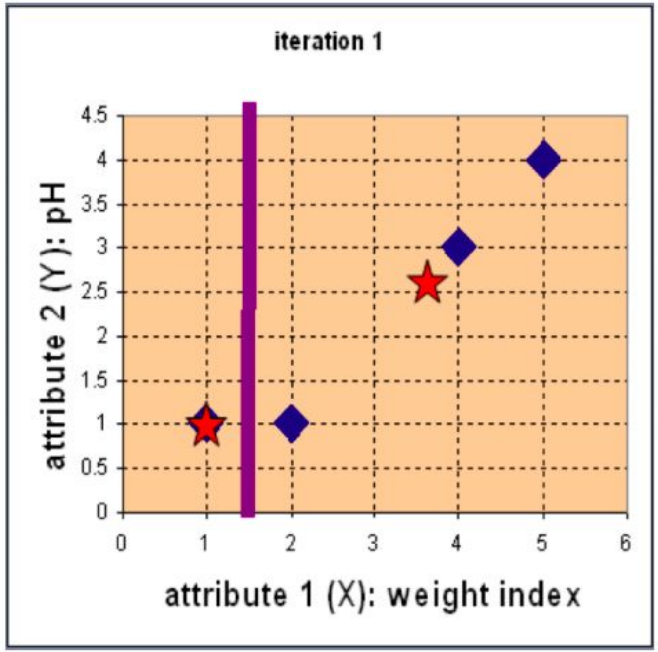
$$c_1 = (1,\ 1)$$

$$c_2 = \left(\frac{2+4+5}{3},\ \frac{1+3+4}{3}\right)$$

$$= \left(\frac{11}{3},\ \frac{8}{3}\right)$$

K-means *Clustering (Ejemplo)*

Paso 2: Renovar las particiones en base a los nuevos centroides



$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

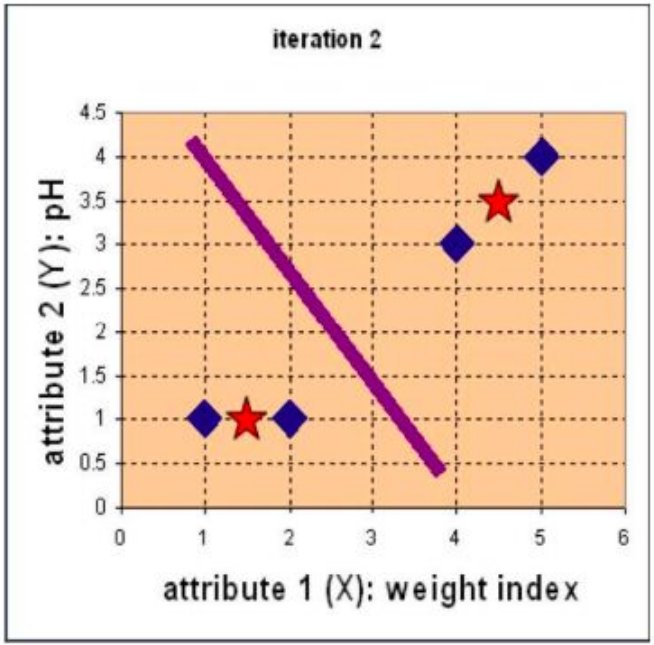
$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

Calcular las distancias a los nuevos centroides

K-means *Clustering (Ejemplo)*

Paso 3: Repetir los dos primeros pasos hasta la convergencia



$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} c_1 = (1\frac{1}{2}, 1) & group-1 \\ c_2 = (4\frac{1}{2}, 3\frac{1}{2}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

Calcular las distancias a los nuevos centroides

K-means Clustering (Problemas)

- Puede converger a un óptimo local
- Complejidad O(tKn): n = número de objetos, K = número de clusters, t = número de iteraciones.
- Se necesita especificar el K a priori.
- Sensible a outliers (*K-medoids algorithm*)
- Sensible a la forma de la distribución de la data
- ¿Cómo medir la performance del algoritmo?

K-means Clustering (Performance)

Elbow Method:

- Calcular la suma de los errores cuadrados (SSE) para algunos valores de k (2,4,6,8..).

- SSE es definida por

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)^2$$

K-means Clustering (Performance)

Elbow Method: