

# UD3-APRENDIZAJE NO SUPERVISADO

*Curso de Especialización en Inteligencia Artificial y Big Data*

*MP Sistemas de Aprendizaje Automático*

## Contenidos de la UD3

1. Clustering K-means
2. Clustering Jerárquico
3. Principal Component Analysis (PCA)

## APRENDIZAJE NO SUPERVISADO

El aprendizaje supervisado usa algoritmos de machine learning para analizar y agrupar conjuntos de datos no etiquetados.

Estos algoritmos son capaces de descubrir patrones ocultos en la información sin necesidad de intervención humana.

Su capacidad para descubrir similitudes y diferencias en la información los hace ideales para minería de datos, venta cruzada, segmentación de clientes, reconocimiento de imágenes, detección de anomalías, etc.

El aprendizaje no supervisado trata de resolver dos tipos principales de problemas:

- El análisis de grupos o clustering
- La reducción de la dimensionalidad

## CLUSTERING K-MEANS

# AN INTUITION TO K-Means Clustering

STARTING WITH UNSUPERVISED LEARNING

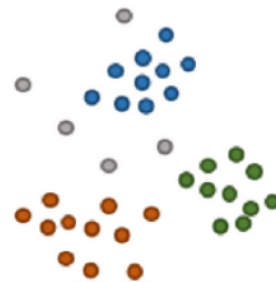


### 1.) WHAT IS UNSUPERVISED LEARNING

Unsupervised learning allows us to approach problems with little or no idea what our results should look like. Unsupervised algorithms find patterns based only on input data. This technique is useful when we're not quite sure what to look for.

### 2.) CLUSTERING ALGORITHMS

Clustering Algorithms do the task of dividing the population or data points into a variety of groups such that data points within the same cluster are similar to other data points within the same cluster than those in other groups. Basically, the aim is to separate groups with similar traits and assign them into clusters.



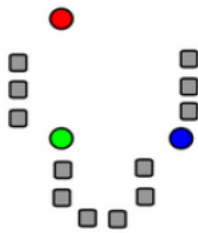
### 3.) K MEANS CLUSTERING



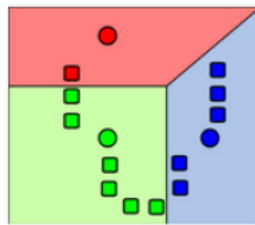
In this algorithm, we group the items into  $k$  clusters such that all items in the same cluster are as similar to each other as possible. And items not in the same cluster are as different as possible.

Distance measures (like Euclidean distance) are used to calculate similarity and dissimilarity between the data points. Each cluster has a centroid. Centroid can be thought as the point that is most representative of the cluster.

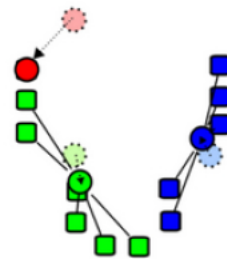
### 4.) HOW K-MEANS CLUSTERING WORKS



1.  $k$  initial "means" (in this case  $k=3$ ) are randomly generated within the data domain.



2.  $k$  clusters are created by associating every observation with the nearest mean.



3. The centroid of each of the  $k$  clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

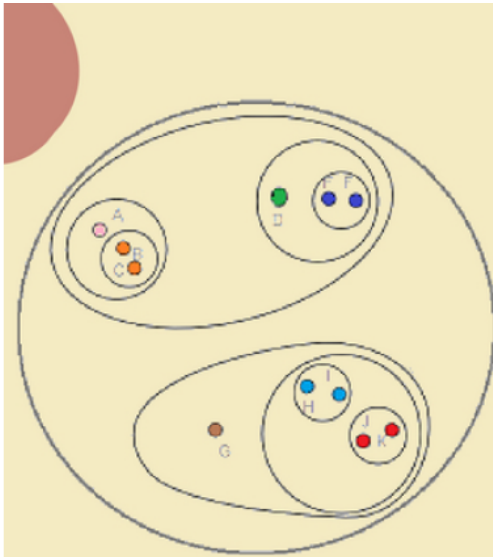
The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

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

Labels in the diagram:

- number of clusters:  $k$
- number of cases:  $n$
- case  $i$ :  $x_i^{(j)}$
- centroid for cluster  $j$ :  $c_j$
- Distance function:  $\|x_i^{(j)} - c_j\|^2$
- objective function:  $J$

## CLUSTERING JERÁRQUICO



### HIERARCHICAL CLUSTERING

Hierarchical clustering, as the name suggests is an algorithm that builds a hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. There are two types of hierarchical clustering, Divisive and Agglomerative.

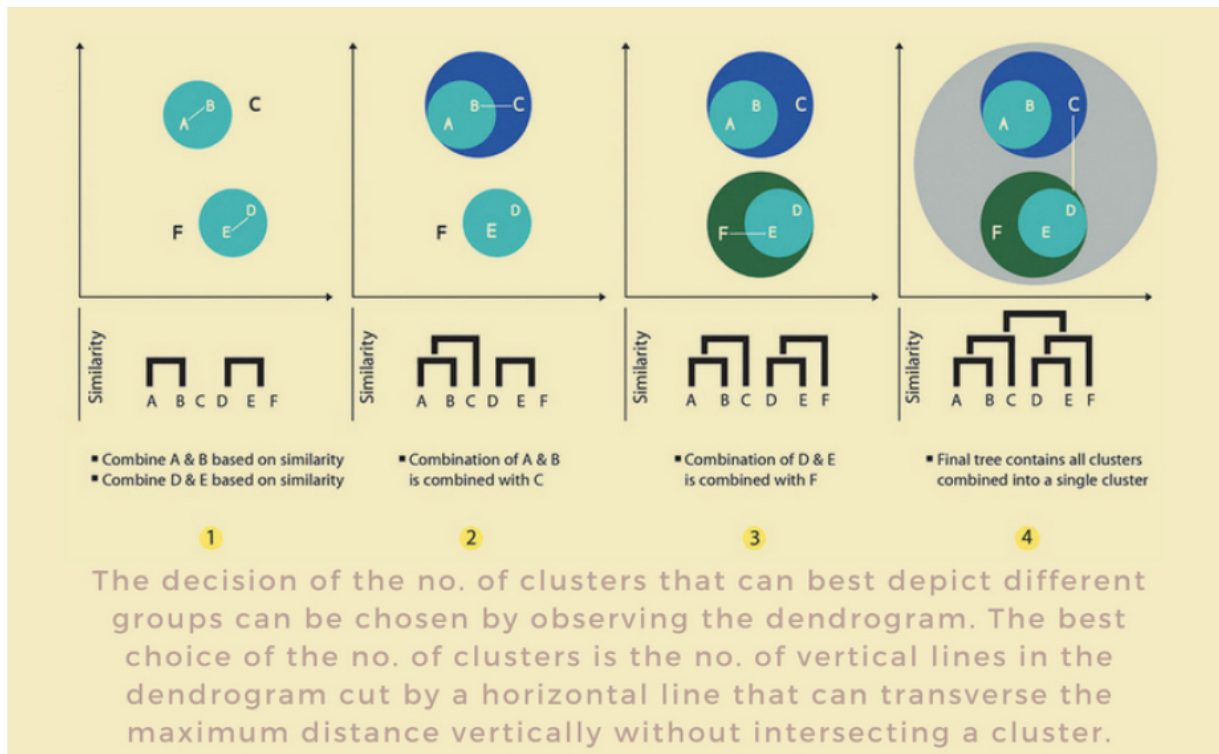


### AGGLOMERATIVE HIERARCHICAL CLUSTERING

Here, each observation is initially considered as a cluster of its own (leaf). Then, the most similar clusters are successively merged until there is just one single big cluster (root). This hierarchy of clusters is represented as a tree (or dendrogram).

### DENDROGRAM

The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.



## WEBGRAFÍA

- Wikipedia
- [100 Days of ML Code](#)