

Clustering



Universidad
de Alcalá



Non supervised learning

Clustering is a non supervised learning techniques, so the **classes** of training samples is **not known**.

Clustering means created groups of sample that will define a class.

Clustering techniques steps:

1. Metric definition to compare samples
2. Criterion for grouping samples
3. Definition of the function to minimize (maximize).

“Flat Clustering”

There is no hierarchy

K-Means
ISODATA

Hierarchy Algorithms

Dividing
Aglomerative

Metric

A valid metric to compare vectors x and y must satisfy:

$$d(x, y) \geq 0$$

$$d(x, y) = 0 \text{ Si } x = y$$

$$d(x, y) = d(y, x)$$

$$d(x, y) \leq d(x, z) + d(z, y)$$

L_2 is the common used metric:

$$d(x, y) = \sqrt{\sum_{i=1}^l (x_i - y_i)^2}$$



Criterion and minimization function

The common function to minimize is the distance from to the centroids:

$$J_{MSE} = \frac{1}{l} \sum_{k=1}^C \sum_{h=1}^l \|x_h - \mu_k\|$$

Let μ_k the centroid of the cluster k , defined as:

$$\mu_k = \frac{1}{l} \sum_{x \in C_k} x_h$$

So, the samples are assigned to the cluster whose centroid is the nearest.

Algorithm K-Means

One of the most popular clustering algorithms, K-means.

Initially the number of clusters is fixed.

K-means steps:

- Set initial centroids (random, uniform, ...)
- Associate each samples to one cluster (minimum distance to the centroid).
- Obtain the new centroid for the previous assignation
- Repeat from step 2 until there is no change.

Hierarchical Clustering

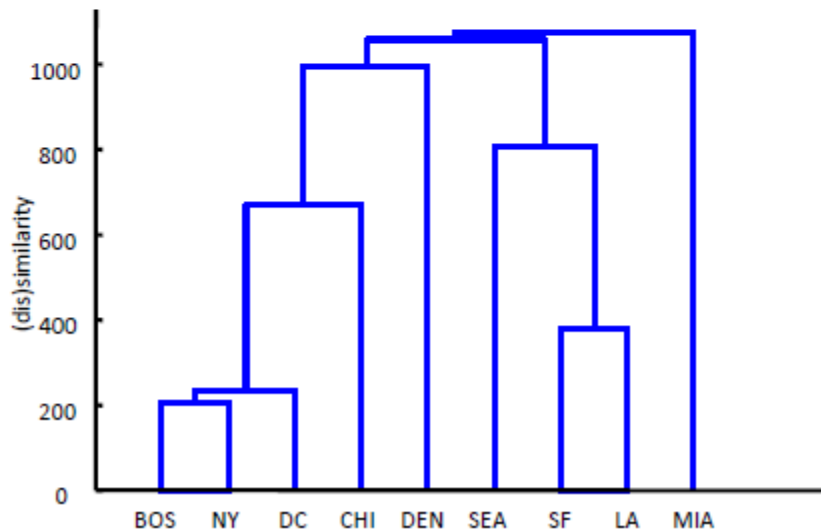
1. Inicialmente tenemos tantos centroides como muestras.
2. Buscamos la pareja de centroides con menor distancia entre sí.
3. Calculamos el centroide de la pareja.
4. Continuamos el algoritmo hasta que solo quede un centroide.
5. Poda: Buscamos inconsistencias entre las distancias de un nodo y la media de las distancias de sus descendientes. Si existe consistencia ese nodo forma centroide. Si no, dividimos.

Clustering Jerárquico

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	200	420	1504	903	2070	3005	2070	1040
NY	200	0	233	1308	802	2815	2034	2780	1771
DC	420	233	0	1075	071	2084	2700	2031	1010
MIA	1504	1308	1075	0	1320	3273	3053	2087	2037
CHI	903	802	071	1320	0	2013	2142	2054	900
SEA	2070	2815	2084	3273	2013	0	808	1131	1307
SF	3005	2034	2700	3053	2142	808	0	370	1235
LA	2070	2780	2031	2087	2054	1131	370	0	1050
DEN	1040	1771	1010	2037	900	1307	1235	1050	0



Single-linkage



Complete-linkage

