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:

$$
\begin{aligned}
&d(x,y) \geq d_0 \\
&d(x,y) = d_0 \text{Si } x = y \\
&d(x,y) = d(y,x) \\
&d(x,y) \leq d(x,z) + d(z,y)
\end{aligned}
$$

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 commond 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:

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

So, the samples are assigned to the cluster whos 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 (ramdon, uniform, ...)
- Asociate 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	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0