Aprenentatge i Reconeixement de Formes
Pattern Recognition and Machine Learning:

5. Unsupervised Pattern Recognition and Clustering

Francesc J. Ferri

Dept. d'Informàtica. Universitat de València

Febrer 2010

Unsupervised Classification Methods

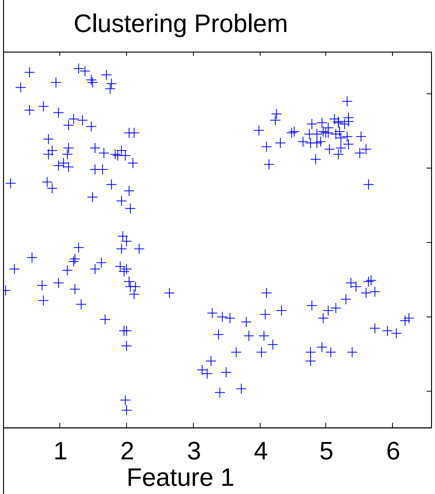
When data is given in a particular representation space and **no labels** are available, we talk about unsupervised methods.

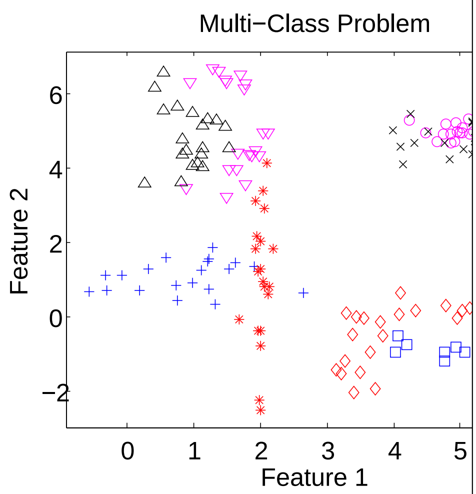
Clustering

Aims at finding groups in a given data in such a way that patterns grouped together are more similar or closer than patterns in other groups.

Clustering is not a well defined task because its goal is partially subjective.

Examples

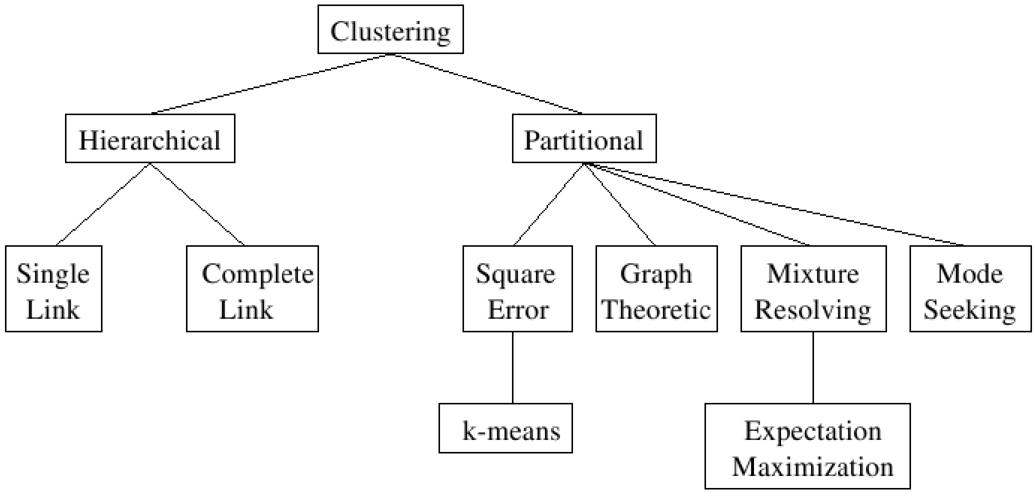




Cluster analysis

1. Representation (vectors, strings, graphs, trees, ...)
2. Similarity (different inter- intra-cluster distances)
3. Grouping method (algorithm to effectively partitioning the data)
4. Data abstraction (ways of representing each cluster)
5. Assessment (how good a clustering result is)

Taxonomy



Hierarchical Clustering

Hierarchical means that we obtain a sequence of **nested** partitions of the data instead of a **single** clustering result.

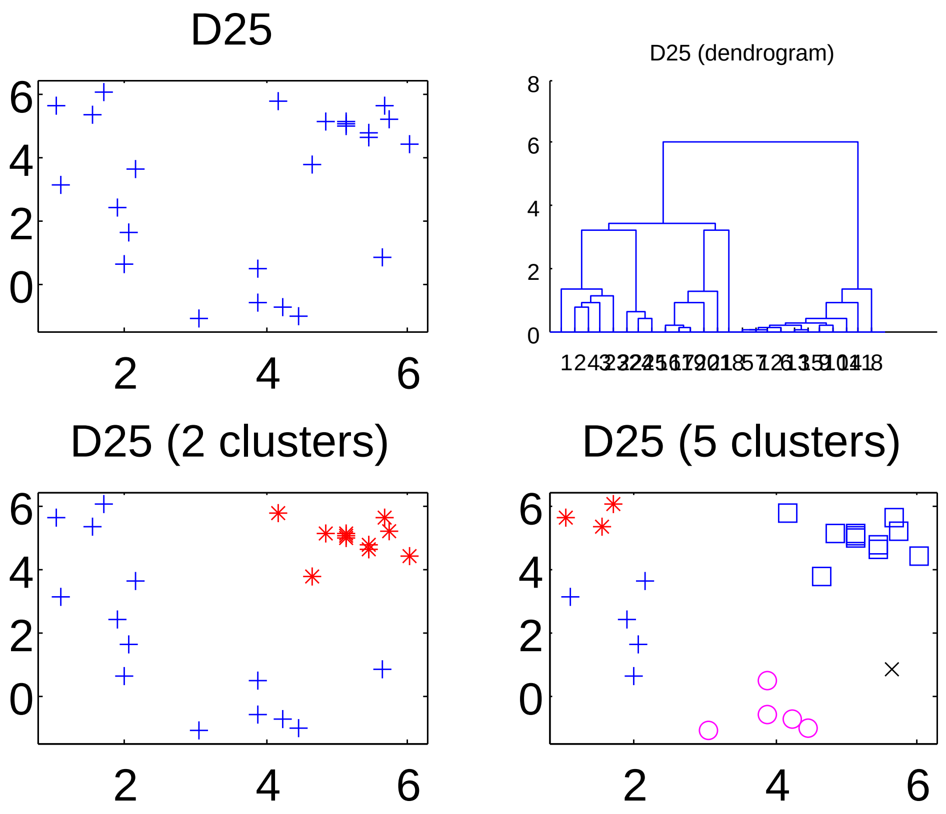
Most of these algorithms are very intuitive, easy to understand and apply and give a very flexible (adjustable) result.

There are basically two kinds: **divisive** (top-down) and **agglomerative** (bottom-up).

Divisive clustering algorithms start from a unique clustering and look for the best way of splitting it into two blocks. The same procedure will be applied recursively to all blocks until there is only one pattern in each block.

Agglomerative clustering start from one cluster per pattern and keep joining neighboring clusters until one ends up with a unique cluster.

Dendrograms



Hierarchical Agglomerative Clustering

General Algorithm

Start with one cluster per pattern
Repeat
  find the two "closest" clusters
  join them and update
Until all patterns are grouped together

Distance between clusters

Measures how close any two clusters are.

The family picture

Single linkage clustering

minimum distance between particular patterns of each cluster

Complete linkage clustering

maximum distance between particular patterns of each cluster

Average linkage clustering

average of distances between pairs of patterns of each cluster

Ward's clustering

takes into account average distances and their variances

Efficiency of agglomerative clustering

$O(n^3)$

The recurrence of Lance and Williams allows to compute the distances to a newly created cluster in a recursive fashion.

Let k be the new cluster resulting from joining i and j. Then, $\forall h$

$$d(h,k) = \frac{1}{2}d(h,i) + \frac{1}{2}d(h,j) \mp \frac{1}{2}|d(h,i) - d(h,j)|$$

This computes de minimum and maximum distances, respectively. In general,

$$d(h,k) = A_i d(h,i) + A_j d(h,j) + Bd(i,j) + C|d(h,i) - d(h,j)|$$

The greedy approach: c-means

The approach consists of establishing a criterion to be optimized.

$$J = \sum_{i=1}^{c} \sum_{x \in X_i} ||x - m_i||^2$$

where m_i is the average of the patterns in the i-th cluster X_i,
$m_i = \frac{1}{|X_i|} \sum_{x \in X_i} x$

The number of clusters, c, needs to be fixed. Otherwise the minimization of J leads to a trivial solution.

The c-means algorithm

Start with an arbitrary c-partition and compute the c means
or start with c arbitrary means

Repeat
  Compute the c partition induced by the current c means
  Compute the the new c means by averaging the patterns at each cluster
Until the means do not change

This algorithm admits many optimizations and extensions.

c-means examples

The c-means (or more correctly, the criterion used) tends to form equally sized hyperspherical clusters.

There is an implicit assumption that the patterns in the clusters have to be normally distributed around their average.

The soft extension: fuzzy c-means

The c-means approach can be extended by considering fuzzy memberships to the c clusters. The criterion is now

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^p ||x_k - m_i||^2$$

where the mean is also generalized as

$$m_i = \frac{\sum_{k=1}^{N} (u_{ik})^p x_k}{\sum_{k=1}^{N} (u_{ik})^p}$$

u_{ik} is the degree of membership of x_k to cluster i, N is the total number of patterns and p is a parameter that controls fuzziness of the result.

The fuzzy c-means algorithm

This criterion can be minimized by using Lagrange multipliers and leads to a well-behaved iterative algorithm.

Once the algorithm has converged to a result, u_{ik}, each pattern is assigned to the cluster whose membership is maximum.

Vector quantization and neural networks

vector quantization – clustering

concepts: quantization error, codebook vectors.

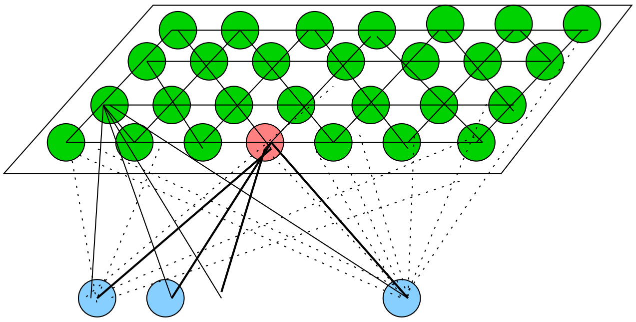
classical vector quantization: Linde, Buzo and Gray (LBG).

Self Organizative Maps (Kohonen)

Can be seen as a vector quantization process coupled with imposing a topology to the codebook vectors (prototypes).

Imagine a **map** of neurons with an explicit topology in form of a (usually regular) graph, in such a way that surroundings of increasing size of each neuron can be defined.

All neurons are connected to D inputs through D weights. So there is a **weight vector** attached to each neuron.



SOM: iterative correction rule

Neuron activation

Correction:

$$w_j = w_j + \eta(t)(x - w_j)$$

It is possible to "see" the neurons (its weight vectors) in the same representation space as the input, x. Is it possible also to represent the topology (the graph) in this space.

...