Prof. Flavio Venturini

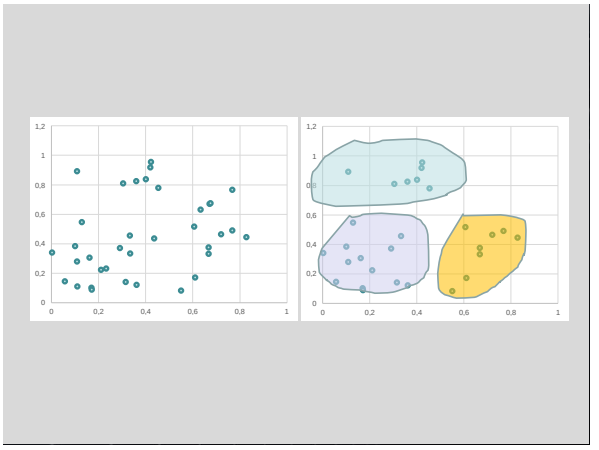
Lez.07 - Introduzione al Machine Learning - Seconda parte

Outline

- Clustering concepts
- K-means clustering
- Hierarchical clustering

What is clustering ?

- Objective: Discover groups in data such that samples within a group are more similar to each other than samples across groups



Hard clustering

- Each data item x_i belongs to only 1 cluster Cj
- Clusters are mutually exclusive

Soft clustering

$$\gamma = \{\gamma_{ki}\}_{k=1:K}^{i=1:n}$$

γ_{ki} the degree of membership of point i to cluster k

$\sum_k \gamma_{ki} = 1$ for all i

(usually associated with a probabilistic model)

Notation for clustering

- D = {x1,x2,…xn} a dataset
- n = number of data points
- K = number of clusters (K << n)
- $\Delta = \{C_1, C_2, \ldots, C_K\}$ a partition of D into disjoint subsets
- $L(\Delta)$ = cost function (loss) of Δ (to be minimized)

Clustering key characteristics

- Several possible cost functions (probabilistic or not)
- Performance on current data is what matters
- K is unknown
- Goal is Data exploration and not prediction

Applications: image segmentation

http://cs.brown.edu/~pff/segment/

Applications: image compression

Ingredients for clustering

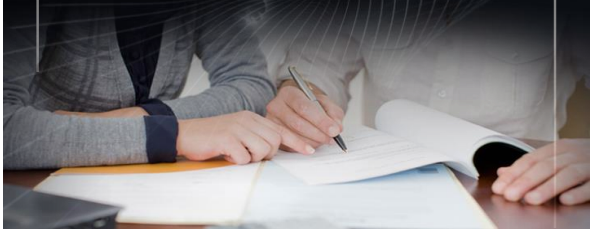
- Data points and their types
- A dissimilarity function between the data points
- A loss function to evaluate clusters
- An algorithm to optimize this loss function

Type of data points

- Categorical (i.e. enumerated types)
  - "Red", "white", "black"
- Ordinal
  - "Old" < "renovated" < "new"
- Quantitative

Dissimilarity functions

- Distance is a way of considering dissimilarity



- Minknowski metric ($p>0$):

$$d(x,y) = \left[\sum_{i=1}^{N} \| x_i - yi \|^p\right]^{1/p}$$

- Euclidean (p=2), Manhattan/Hamming (p=1)

Loss function

Several possible loss functions, characteristic of each algorithm

- K-means uses the distance of each data point belonging to a cluster from its cluster centre

- Hierarchical single linkage uses the shortest distance from any member of one cluster to any member of the other clusters

Review questions

- What is the difference between hard and soft clustering?
- What is needed to perform cluster analysis?

Review questions

- What are the key characteristics of clustering?
- Is clustering supervised or unsupervised?

K-MEANS CLUSTERING

K-means idea

- Choose the number of clusters, k
- Generate k random points as cluster centroids

- Assign each point to the nearest cluster centroid
- Recompute the new cluster centroid
- Repeat steps 3 and 4 until convergence is met

K-means in math terms

- K clusters each summarized by a prototype μ_k
- Assignment of data x_i represented by responsibilities

  $r_{ik} \in \{0,1\}\ with\ \sum_{k=1}^{K} r_{ik} = 1$
- Loss function

  $J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \parallel xi - \mu_k \parallel^2$

- J is minimized in two steps
- E-step: fix μ_k minimize J w.r.t. r_{ik}
  - Assign each data point to its nearest prototype

- M-step: fix r_{ik}, minimize J w.r.t. μ_k
  - Set each prototype to the mean of the points in that cluster

$$\mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}$$

- This procedure is guaranteed to converge to a local minimum

How to initialize ?

Heuristics:

- Randomly pick K data points as prototypes
- Pick prototype $i+1$ to be the farthest from the first i prototypes

How to choose K ?

The objective of cluster analysis is to minimize within cluster variance (Var_w) and to maximise between cluster variance (Var_b)

$Var_w(X) = E[(X-\mu)^2]$

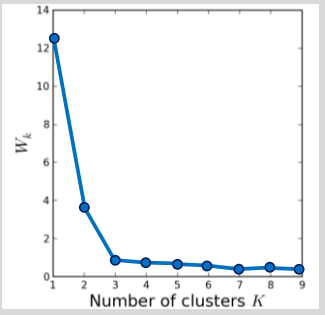
$Var_b(Y) = E[(Y-\mu)^2]$

Elbow criterion

Therefore, as the number of clusters is increasing, the ratio of Var_w to Var_b will keep decreasing

At some point the marginal gain of adding a new cluster will drop, giving an angle in the graph

Elbow chart



Better than Elbow ?

Davies-Bouldin index

$$DB \equiv \frac{1}{N}\sum_{i=1}^{N} D_i$$

$$D_i \equiv \max_{j \neq i} R_{i,j}$$

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$$

$M_{i,j}$ the separation between the i^{th} and the j^{th} cluster, which ideally has to be as large as possible

S_i, the within cluster scatter for cluster i, which has to be as low as possible

Review questions

- Describe the k-means algorithm
- How can you choose the right k?
- Do you believe there is a correct k for each case?
- Using euclidean distance as a metric, is scaling of variables an issue or not?

HIERARCHICAL CLUSTERING

Agglomerative HC

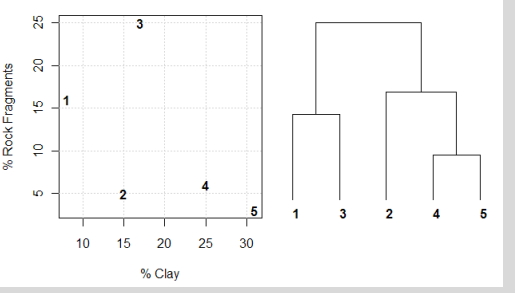
- Start with each point as individual cluster
- At each step, merge the closest pair of clusters until only one cluster (or k clusters) left

Divisive HC

- Start with one, all inclusive, cluster
- At each step, eliminate from each cluster its farest point until each cluster contains a single point (or there are k clusters)

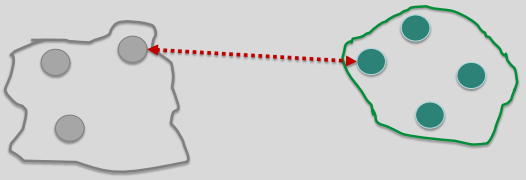
HC results

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram



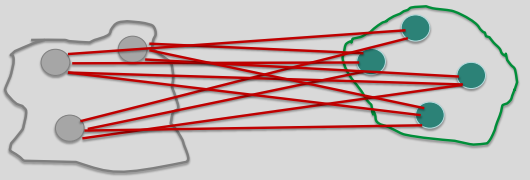
Proximity measures

Min (single link)



Proximity measures

Group average intercluster similarity

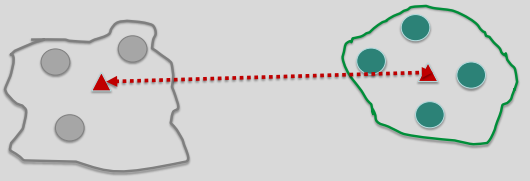


Proximity measures

Max (complete link)

Proximity measures

Distance between centroids



Review questions

- Describe the hierarchical agglomerative and divisive algorithm
- Describe the proximity measures used in HC
- What is the HC key difference with k-means?

SUMMARY QUESTIONS

- What is the difference between hard and soft clustering?
- What is needed to perform cluster analysis?
- What are the key characteristics of clustering?
- Is clustering supervised or unsupervised?

- Describe the k-means algorithm
- How can you choose the right k?
- Do you believe there is a correct k for each case?
- Using euclidean distance as a metric, is scaling of variables an issue or not?

- Describe the hierarchical agglomerative and divisive algorithm
- Describe the proximity measures used in HC
- What is the HC key difference with k-means?