STATISTICAL LEARNING

COURSE 1 - PRINCIPAL COMPONENTS ANALYSIS

ECOLE CENTRALE DE LYON - BACHELOR 2ND YEAR
2024-2025

Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other,
- It make this concrete, we must define what it means for two or more observations to be similar or different.
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied

PCA vs Clustering

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
- Clustering looks for homogeneous subgroups among the observations.

Two clustering methods

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.
- In hierarchical clustering, we do not know in advance how many clusters we want ; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to *n*.

K-means clustering

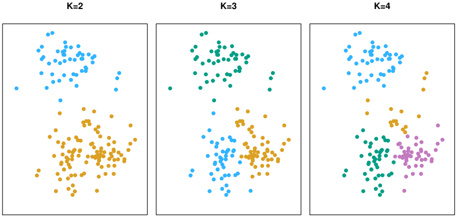


Figure – from An introduction to Statistical Learning

A simulated data set with 150 observations in 2-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering ; instead, they are the outputs of the clustering procedure.

Details of K-means clustering

Let $C_1, ..., C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties :

1. $C_1 \cup C_2 \cup ... \cup C_K = \{1, ..., n\}$. In other words, each observation belongs to at least one of the K clusters.
2. $C_k \cap C_k' = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping : no observation belongs to more than one cluster. For instance, if the ith observation is in the kth cluster, then $i \in C_k$.

Details of K-means clustering : continued

- The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.
- The within-cluster variation for cluster C_k is a measure $WCV(C_k)$ of the amount by which the observations within a cluster differ from each other.
- Hence we want to solve the problem

$$minimize_{C_1,\ldots,C_K} \sum_{k=1}^{K} WCV(C_k) \tag{1}$$

- In words, this formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

How to define within-cluster variation ?

- We need is a notion of distance between our data.
- Typically we use Euclidean distance.
- Many others distance metrics exist :
  - For strings or DNA sequences, one might use edit distance.
  - For bit vectors, it might be sensible to use Hamming distance

How to define within-cluster variation : continued

- In case of Euclidean distance $\|x - y\|_2$

$$WCV(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \|x_i - x_{i'}\|_2^2, \qquad (2)$$

where $|C_k|$ denotes the number of observations in the kth cluster.

- Combining 1 and 2 gives the optimization problem that defines K-means clustering,

$$minimize_{C_1,\ldots,C_K} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \|x_i - x_{i'}\|_2^2. \qquad (3)$$

K-Means Clustering (Lloyd's Algorithm)

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing :
    2.1 For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

    2.2 Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance)

Two steps of the algorithm

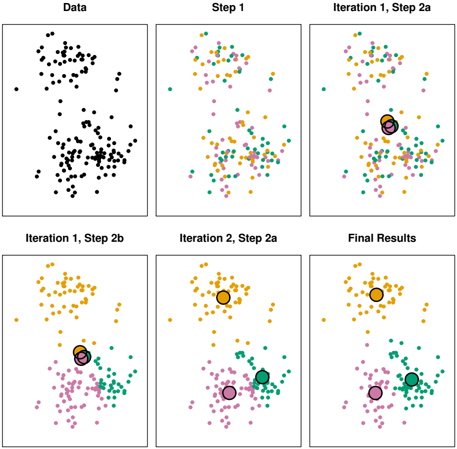


Figure – from An introduction to Statistical Learning. Progress of the Lloyd's Algorithm

Properties of the Lloyd's Algorithm

- This algorithm is guaranteed to decrease the value of the objective 3 at each step. Why ? Note that

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\|x_i - x_{i'}\|_2^2 = 2\sum_{i\in C_k}\|x_i - \bar{x}_k\|_2^2,$$

  where $\bar{x}_k = \frac{1}{|C_k|}\sum_{i\in C_k} x_i$ is the mean (cluster centroid) in cluster C_k.

- however it is not guaranteed to give the global minimum. Why not ?
- highly non-convex optimization problems
- use random restarts and choose the best i.e. that for which the objective (3) is smallest.

Two steps of the algorithm

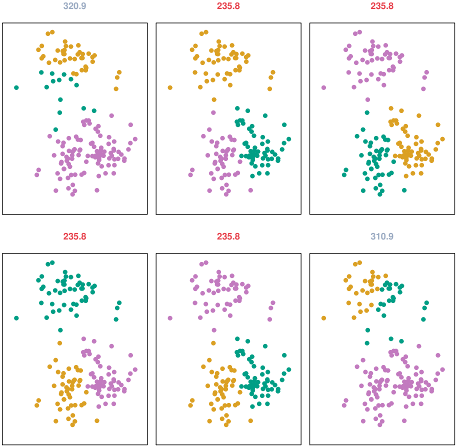


Figure – from An introduction to Statistical Learning.

Example

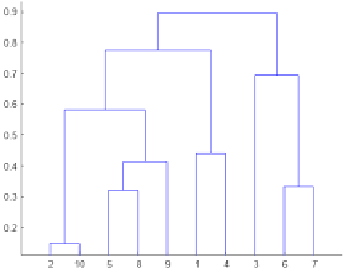
See Notebook

Hierarchical Clustering

- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage (later we discuss strategies for choosing K)
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.
- In this section, we describe bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

Goals : build a tree structure that :

1. shows hierarchical links between individuals or groups of individuals
2. detects a “natural” number of classes individuals or groups of individuals



Choice of Dissimilarity Measure

- So far have used Euclidean distance.
- An alternative is correlation-based distance which considers two observations to be similar if their features are highly correlated.
- This is an unusual use of correlation, which is normally computed between variables ; here it is computed between the observation profiles for each pair of observations. See Details

How to compute dissimilarity between groups

- Dissimilarity between pairs of observation is obvious
- Question : How to extend it to a pair of groups of observations ?
- Doing by the notion of *linkage*.
- Several notion of linkage.

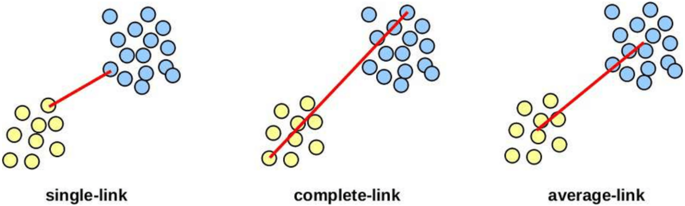
Types of Linkage

Linkage	Description
Complete	Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.
Single	Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.
Average	Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions.

Which linkage ?

Similarity between groups of individuals :

- minimum jump or single linkage(smallest distance)
- complete linkage (largest distance)



- Complete and average are preferred
- Single linkage tend to yield a unbalanced dendograms
- Centroid is often used in genomics .

Practical issues

- Scaling of the variables matters !. Should the observations or features first be standardized in some way ? For instance, maybe the variables should be centered to have mean zero and scaled to have standard deviation one.
- In the case of hierarchical clustering,
- What dissimilarity measure should be used ?
- What type of linkage should be used ?
- How many clusters to choose ? (in both K-means or hierarchical clustering). Difficult problem. No agreed-upon method.
- Which features should we use to drive the clustering ?