Data Exploration - Clustering

Pietro Franceschi
pietro.franceschi@fmach.it

FEM - UBC

Clustering

Clustering

The term “cluster” has the meaning of “concentrated” group. It usually refers to the objects (in the variable space), but is also used for variables (in the space of the objects), or for both, variables and objects simultaneously.

In a nutshell

Cluster analysis tries to find groups containing similar objects, it is thus a method for **UNSUPERVISED LEARNING**

Key Ideas

- One needs a measure of similarity among the objects
- The right “similarity” depends on the problem
- The similarity measure gives high freedom
- . . . one can use it to look for a specific result . . .

Methods

- HIERARCHICAL METHODS : Objects and partitions are arranged in a hierarchy. An appropriate graphical representation of the result is a tree-like dendrogram. It allows to determine manually the "optimal" number of clusters as well as to see the hierarchical relations between different groups of objects
- PARTITIONING METHODS : Each object is assigned to exactly one group
- FUZZY CLUSTERING METHODS : Each object is assigned by a membership coefficient to each of the found clusters. Usually, the membership coefficients are normalized to fall in the interval [0, 1], and thus they can be interpreted as probability that an object is assigned to anyone of the clusters
- MODEL-BASED CLUSTERING : The different clusters are supposed to follow a certain model, like a multivariate normal distribution with a certain mean and covariance

Similarity and Distance

Similarity can be seen as the inverse of a *distance measure*

A function d on a set of points is called distance if:

$$d(x, y) = 0 \Leftrightarrow x = y$$

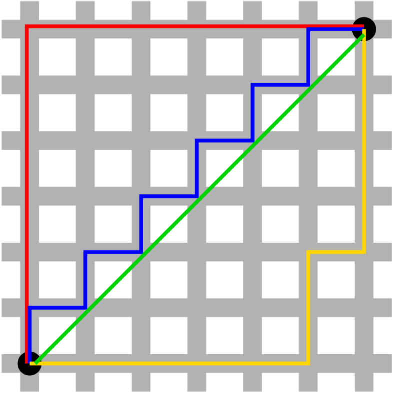
$$d(x, y) = d(y, x)$$

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

Clustering and Scaling

When we rely on a distance measure the result will be heavily **depending on scaling**

Flexibility



- red, yellow and blue lines have the same taxicab distance
- in euclidean metric the green is the unique shortest path

Clustering the Variables

Most of the standard clustering algorithms can be directly used for **clustering the variables**. In this case, the "distance between the variables" rather than between the objects has to be measured. A popular choice is the **Pearson correlation Distance**

$$d_{CORR}(v_j, v_k) = 1- \mid r_{ik} \mid$$

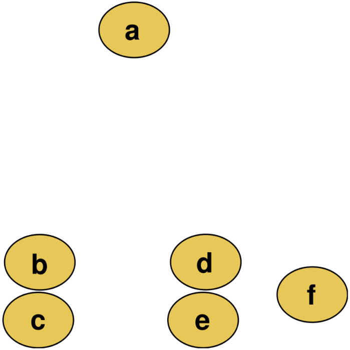
Common Questions

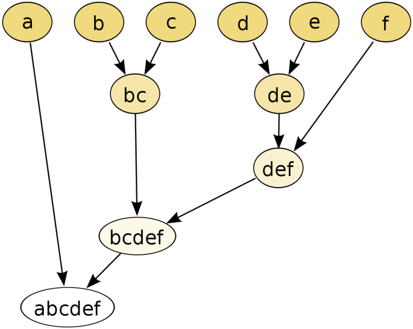
????

- Are my data really clustered?
- What type of clustering should I use?
- How many clusters do I have?

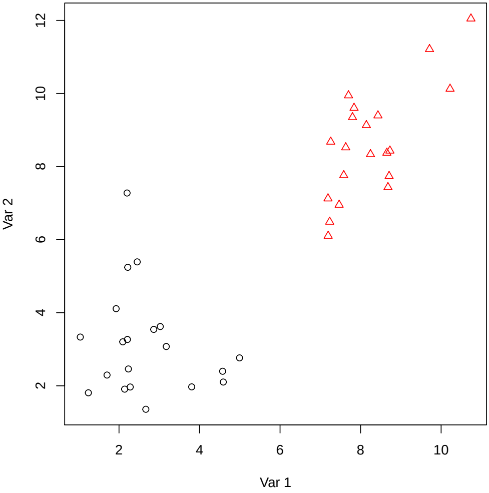
Agglomerative Clustering

In the beginning of the process, **each element is in a cluster of its own**. The clusters are then sequentially **combined into larger clusters**, until all elements end up being in the same cluster. At each step, the two clusters separated by the shortest distance are combined.

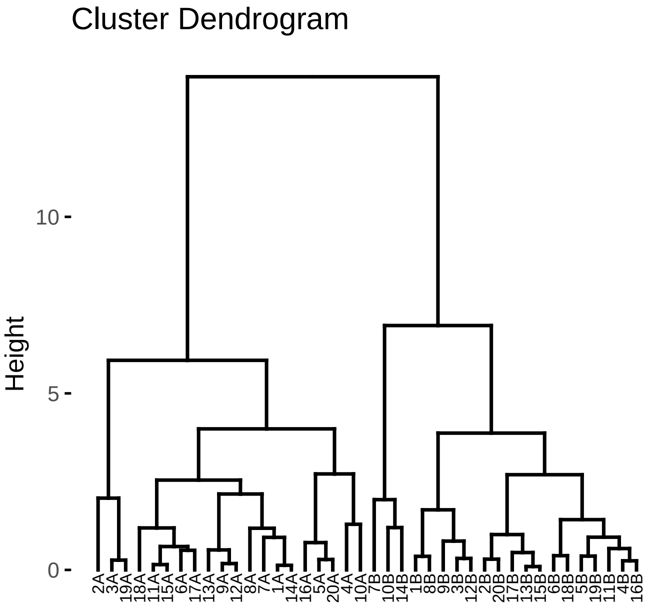

a
b
c
d
e
f


a
b
c
d
e
f
bc
de
def
bcdef
abcdef

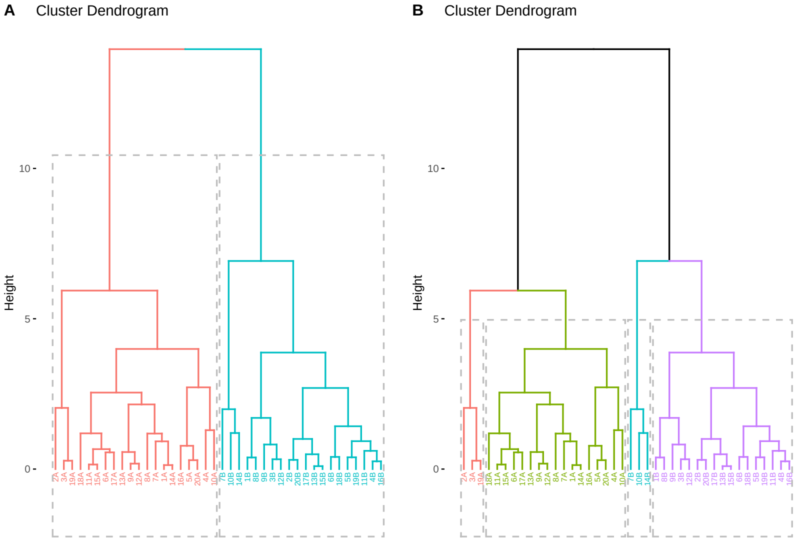
Example data



Dendrogram

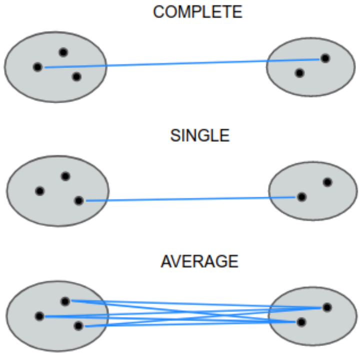

Cluster Dendrogram
Height
10
5
0
2A
3A
19A
18A
11A
15A
6A
17A
13A
9A
12A
8A
7A
1A
14A
16A
5A
20A
4A
10A
7B
10B
14B
1B
8B
9B
3B
12B
2B
20B
17B
13B
15B
6B
18B
5B
19B
11B
4B
16B

Cutting a dendrogram

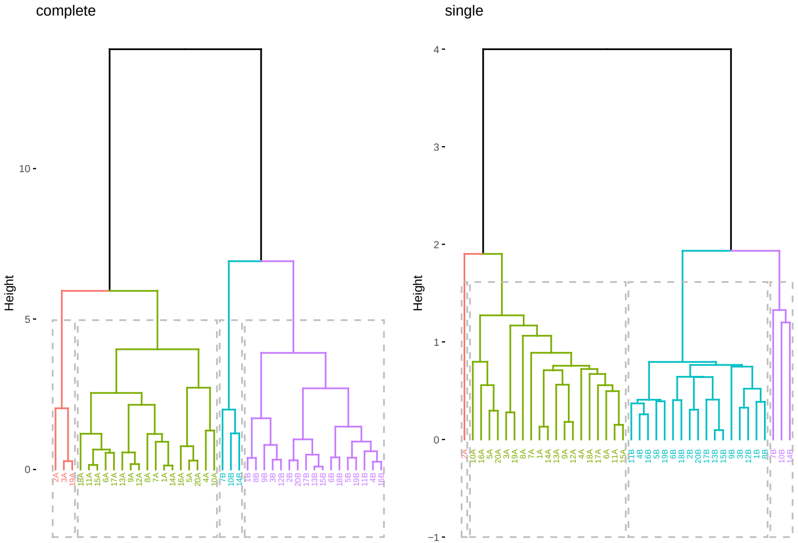


Linkage

The **linkage** defines the way we calculate the distance between two groups or between one unit and a group



Effects of linkage



Cophenetic distance

The cophenetic distance between two observations that have been clustered is defined to be the intergroup dissimilarity at which the two observations are first combined into a single cluster.

It can be argued that a dendrogram is an appropriate summary of some data if the correlation between the original distances and the cophenetic distances is high.

In the toy example:

```
## [1] 0.9002377
```

Notes

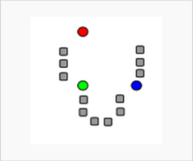
- You need to calculate the distance among all the elements
- Once done, you can cut the tree wherever you want
- If you read it from the top, if two elements are split, they will be in different groups until the end ...
- With big datasets it becomes quite computationally demanding ...

Partitional clustering

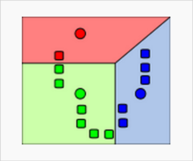
In the beginning of the process, the dataset is decomposed in a **set of disjoint groups**. Given a data set of N points, a partitioning method constructs K (N $\geq$ K) partitions of the data, with each partition representing a cluster

- An euclidean (weighted) distance measure
- An hypothesis on the number of clusters
- . . . a reasonably good computer . . .

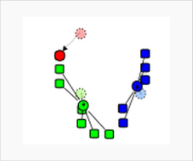
K-means



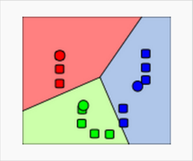
1) *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3) The centroid of each of the *k* clusters becomes the new mean.



4) Steps 2 and 3 are repeated until convergence has been reached.

Notes

- You do not need need to calculate the distance among all the elements
- If you change your mind you should re-calculate everything
- It is fast also with big datasets
- If you change the starting points the class membership could change
- *k-means* algorithm tends to be sensitive to outliers. A more robust alternative is **k-medoids**

Is my dataset really clustered?

Before applying any clustering method on your data, it's important to evaluate whether the data sets contains meaningful clusters (i.e.: non-random structures) or not. If yes, then how many clusters are there. This process is defined as the assessing of clustering tendency or the feasibility of the clustering analysis

- Inspection of the data
- Hopkins Statistics [0,1]
- Visual methods

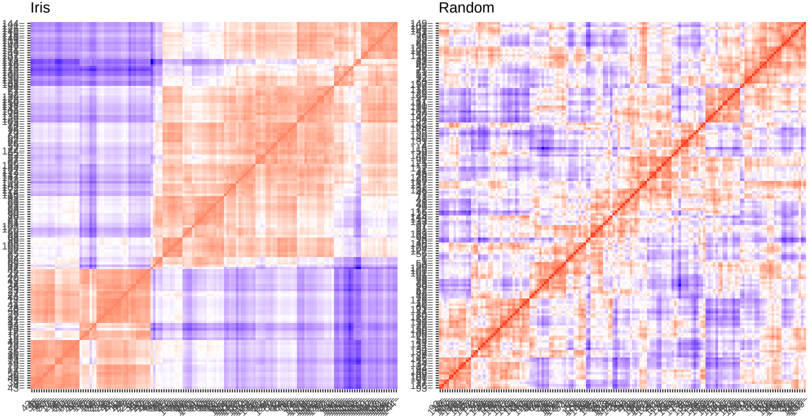
Hopkins Statistics

Is the distribution of the distance among the different objects different from the one I would obtain for randomly distributed data in the same space?

Data compatible with a random distribution will get $H \sim 0.5$. Higher values are suggesting the presence of clusters

Nice idea . . . but I really need a substantial amount of points ;-)

Visual Methods

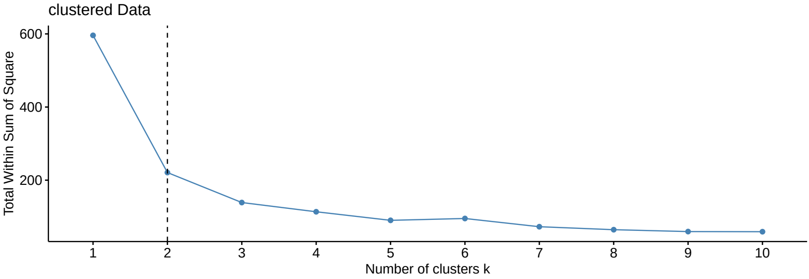


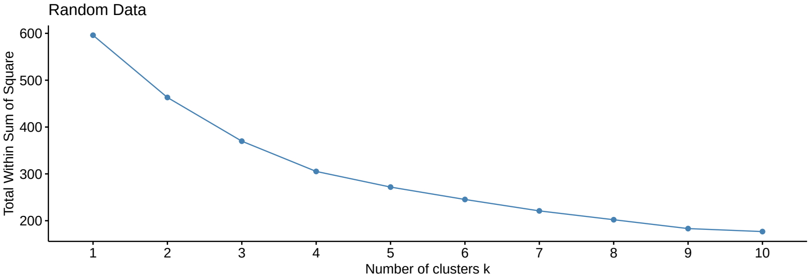
Ok, but how many clusters?

If my data show a tendency to cluster, we need to to find the “best” number of clusters here we list three possible approaches:

- monitor the **within cluster sum of squares (WSS)** as a function of the number of clusters
- inspect the **silhouette** plot
- calculate the **Dunn Index** as a function of the number of clusters

Within Cluster Sum of Squares





Silhouette

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

For each element *i* of the dataset:

- b_i is the smallest distance between *i* and the elements belonging to another cluster
- a_i is the average distance between *i* and the elements of its cluster
- s_i close to one means that the datum is appropriately clustered
- s_i negative indicate that *i* would be more appropriate if it was clustered in its neighboring cluster

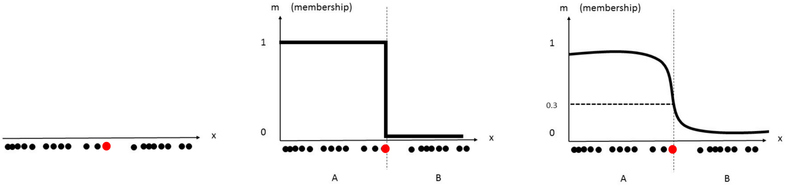
Dunn Index

- For each cluster, compute the distance between each of the objects in the cluster and the objects in the other clusters
- Use the minimum of these pairwise distances as a measure of the inter cluster separation (*min.separation*)
- For each cluster compute the distance between the objects belonging to the cluster
- Use the maximal intra cluster distance ($max.diameter$) as a measure of the cluster compactness

$$D = \frac{min.separation}{max.diameter}$$

Fuzzy Clustering

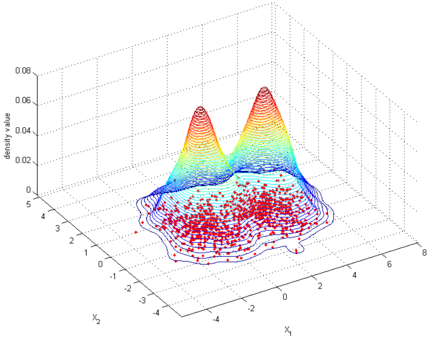
Fuzzy clustering allows for a *fuzzy* assignment meaning that an observation is not assigned to exclusively one cluster but at some part to all clusters by calculating a "membership"



Model Based Clustering

Model-based clustering assumes a statistical model of the clusters.

Each cluster is supposed to be represented by a statistical distribution, like the multivariate normal distribution, with certain parameters for mean and covariance.



LIVE