Machine Learning and Intelligent Systems

Hierarchical Clustering

Maria A. Zuluaga

Jan 26, 2023

EURECOM - Data Science Department

Table of contents

Preliminaries

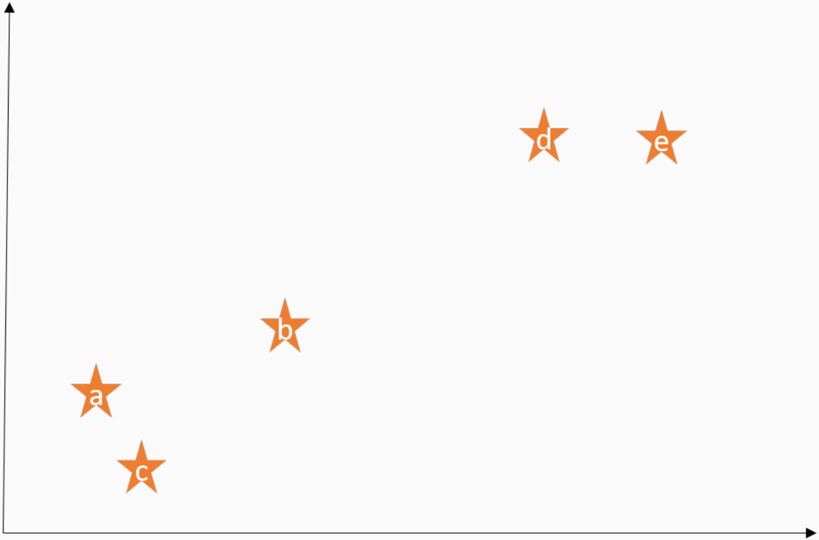
K-means: Limitations

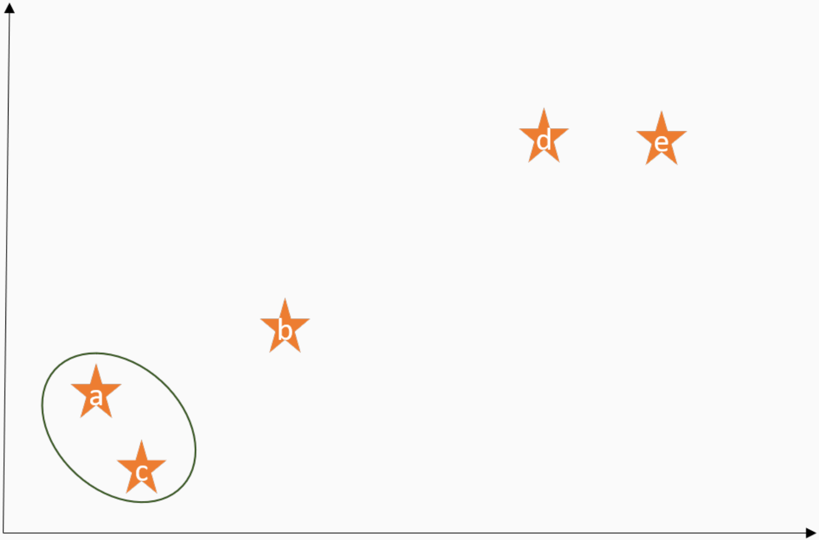
- K-means clustering requires to pre-specify the number of clusters K.
- Results from K-means are somehow random: they depend on the random initialization from which the algorithm started

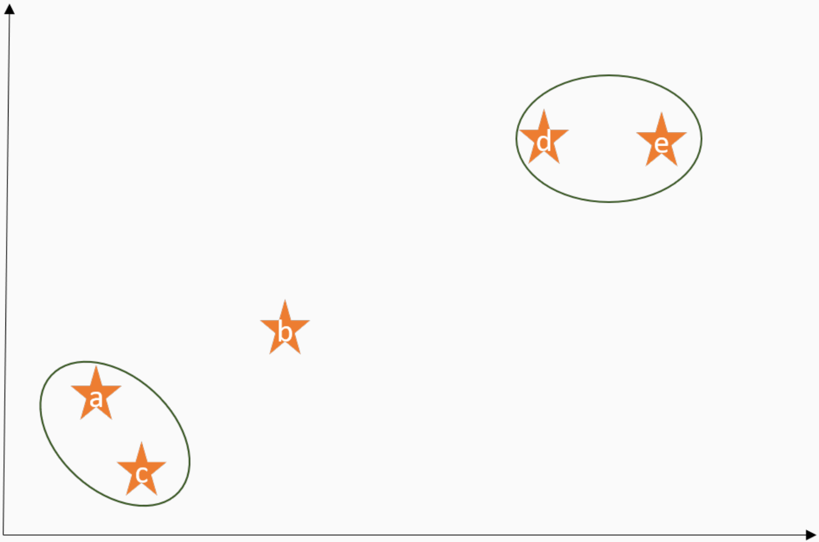
- Hierarchical clustering is an alternative approach that:
  1. Does not require a pre-defined K
  2. Provides a deterministic answer
- Two types:
  1. Bottom-up or agglomerative
  2. Top-down or divisive

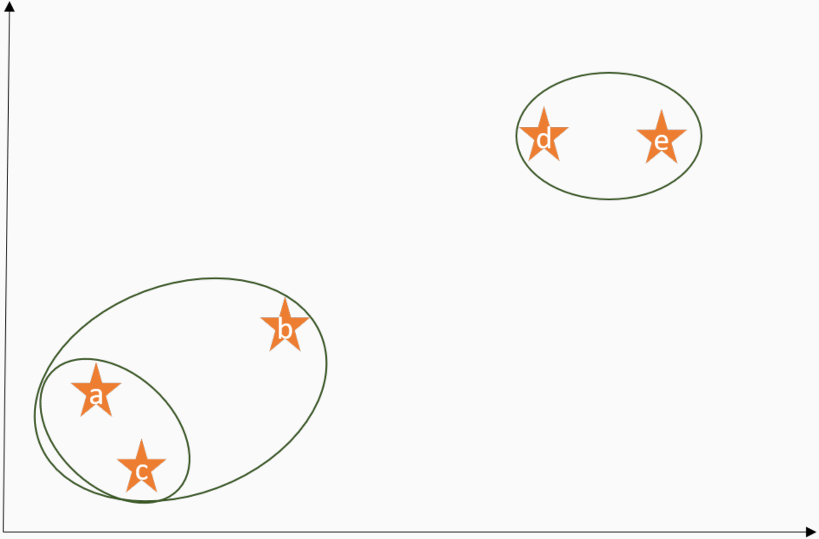
Algorithm

Walk-through Example




a
b
c
d
e


a
b
c
d
e


a
b
c
d
e

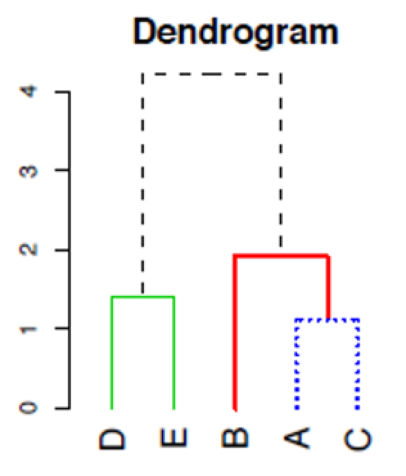
Agglomerative Hierarchical Clustering

Algorithm

1. Start with each point in its own
2. Identify the two closest clusters. Merge
3. Repeat until all points are in a single cluster

Results can be visualized using a dendrogram

The Y-axis reflects the distance between the clusters that got merged at that step



Linkage

Let $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ denote the **dissimilarity** between samples $\mathbf{x}_i, \mathbf{x}_j$.

Linkage

Let $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ denote the **dissimilarity** between samples $\mathbf{x}_i, \mathbf{x}_j$.

- At the first step of hierarchical clustering, each cluster is a single point
- The merging is done over two observations showing the lowest dissimilarity
- After that, we need to think about dissimilarities (distances) between sets (clusters) not single points

Linkage

Let $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ denote the **dissimilarity** between samples $\mathbf{x}_i, \mathbf{x}_j$.

- At the first step of hierarchical clustering, each cluster is a single point
- The merging is done over two observations showing the lowest dissimilarity
- After that, we need to think about dissimilarities (distances) between sets (clusters) not single points

Linkage: Dissimilarity between two clusters:

$$d(G, H)$$

denotes the dissimilarity between sets G and H.

Linkage types

Complete Maximal inter-cluster dissimilarity. Computes all pairwise dissimilarities between observations of cluster G and cluster H. Records the largest

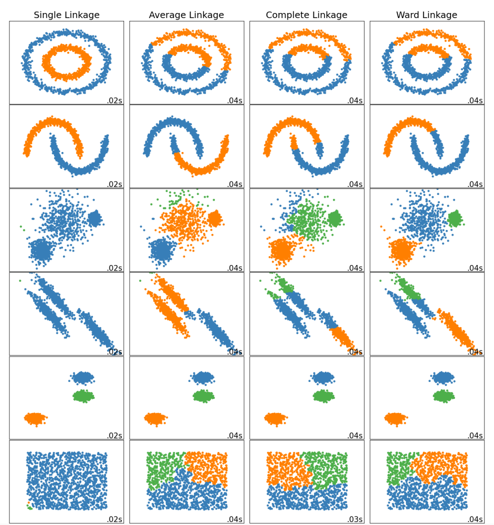
Single Minimal inter-cluster dissimilarity. Computes all pairwise dissimilarities between observations of cluster G and cluster H. Records the smallest.

Avergae Mean inter-cluster dissimilarity. Computes all pairwise dissimilarities between observations of cluster G and cluster H. Records the average.

Centroid Dissimilarity between the centroid for cluster G and centroid for cluster H.

Ward Minimizes the variance of the clusters to be merged. At each step find the pair of clusters that leads to minimum increase in total within-cluster variance after merging

Examples



Source: scikit-learn

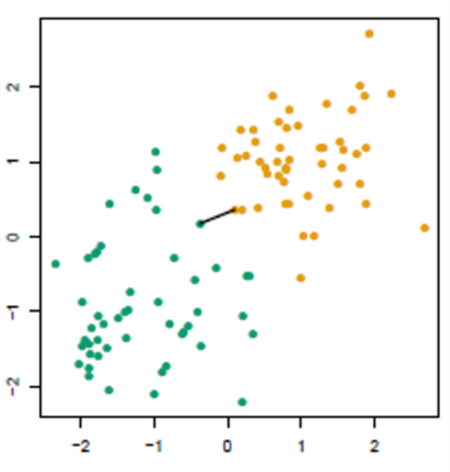
Single Linkage

The dissimilarity between G, H is the smallest dissimilarity between two points in different groups.

It minimizes the distance between the closest observations of pairs of clusters.

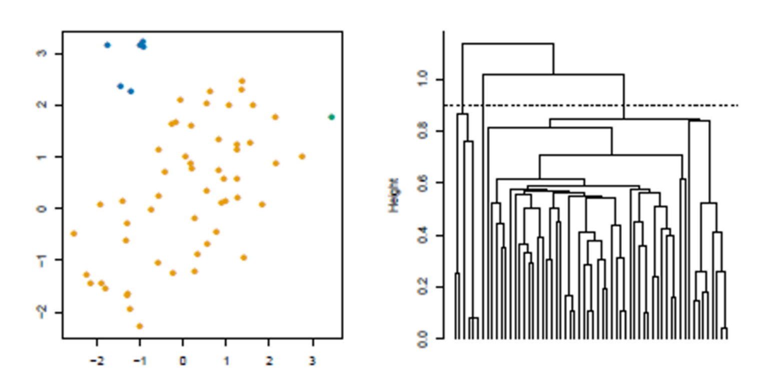
It is a nearest-neighbor linkage

$$d_{single}(G, H) = \min_{i \in G, j \in H} d(\mathbf{x}_i, \mathbf{x}_j)$$



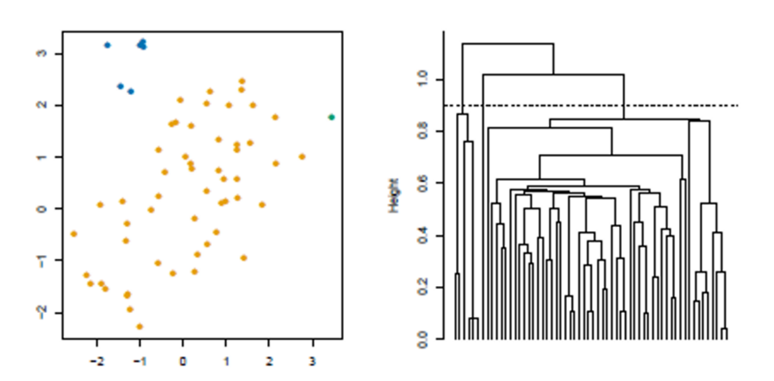
Example

Setup: $N = 60$, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, cut at $h = 0.9$.



Example

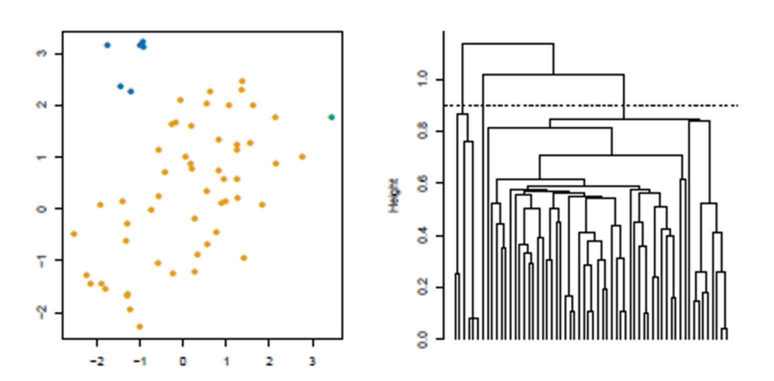
Setup: $N = 60$, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, cut at $h = 0.9$.



Interpretation?

Example

Setup: $N = 60$, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, cut at $h = 0.9$.



Interpretation?

Answer: For each point, there is another point in its cluster such that $d \leq 0.9$

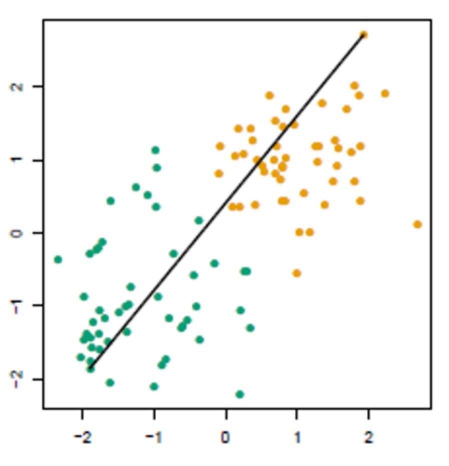
Complete Linkage

The dissimilarity between G, H is the largest dissimilarity between two points in different groups.

It minimizes the maximum distance between observations of pairs of clusters.

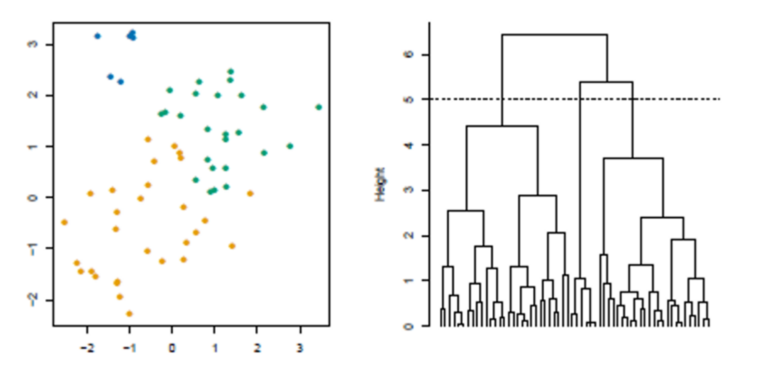
It is a farthest-neighbor linkage

$$d_{complete}(G, H) = \max_{i \in G, j \in H} d(\mathbf{x}_i, \mathbf{x}_j)$$



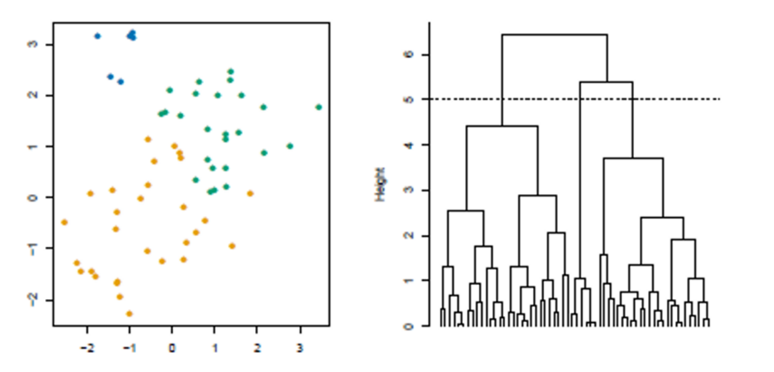
Example

Setup: $N = 60$, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, cut at $h = 5$.



Example

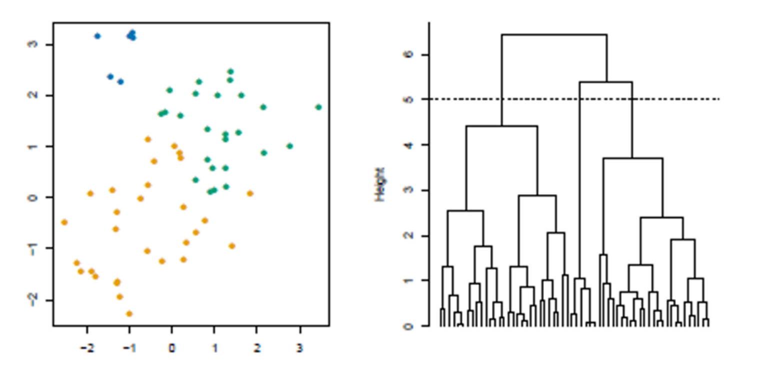
Setup: $N = 60$, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, cut at $h = 5$.



Interpretation?

Example

Setup: $N = 60$, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, cut at $h = 5$.



Interpretation?

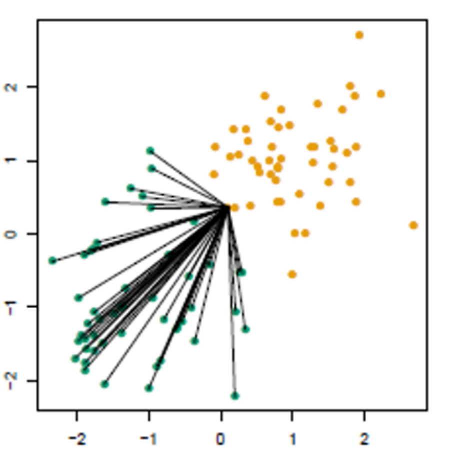
Answer: For each point, every other point satisfies $d_{ij} \leq 5$

Average Linkage

The dissimilarity between G, H is is the average dissimilarity over all points in opposite groups.

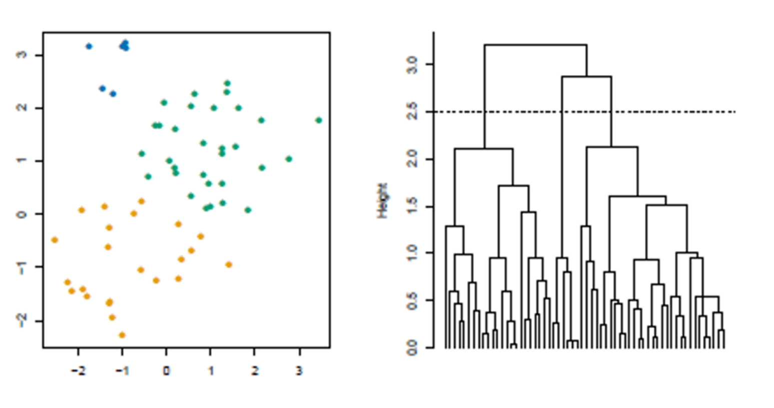
It minimizes the average of the distances between all observations of pairs of clusters.

$$d_{AVG}(G, H) = \frac{1}{|G||H|} \sum_{i \in G, j \in H} d(\mathbf{x}_i, \mathbf{x}_j)$$



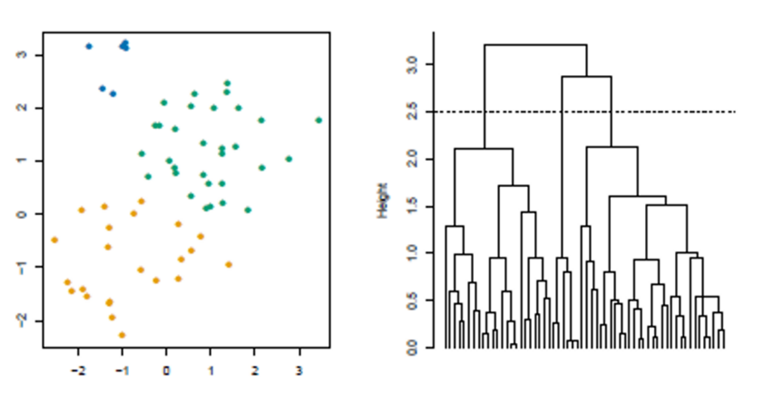
Example

Setup: $N = 60$, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, cut at $h = 2.5$.



Example

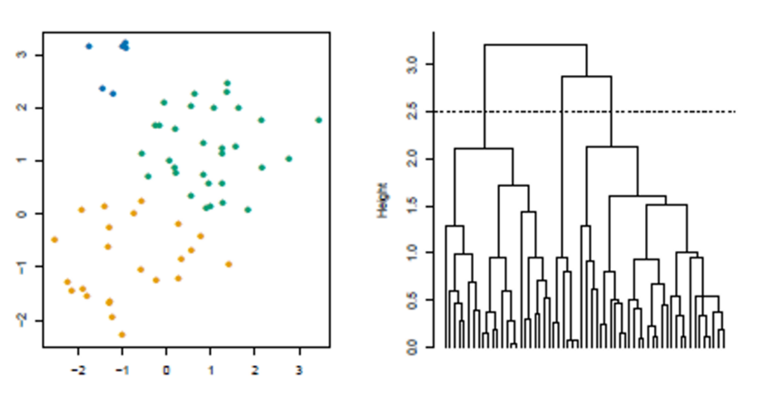
Setup: $N = 60$, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, cut at $h = 2.5$.



Interpretation?

Example

Setup: $N = 60$, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, cut at $h = 2.5$.



Interpretation?

Answer: Not a good interpretation

Chaining and Crowding

Chaining:

- To merge two groups, single linkage only needs for one pair of points to be close.
- Clusters can be too spread out and not compact enough

Chaining and Crowding

Chaining:

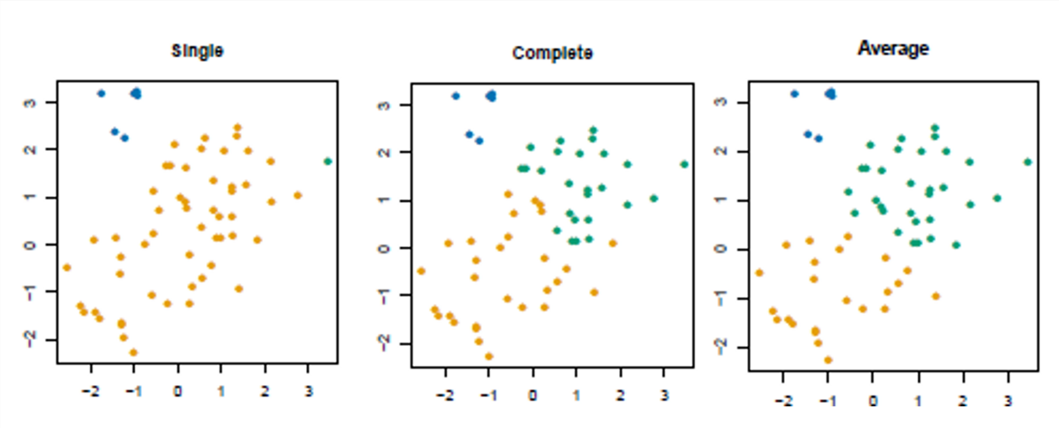
- To merge two groups, single linkage only needs for one pair of points to be close.
- Clusters can be too spread out and not compact enough

Crowding:

- Complete linkage avoids chaining but suffers from crowding
- Because its score is based on the worst-case dissimilarity, a point can be closer to points in other clusters than to points in its own cluster.
- Compact clusters that are not well separated

Average linkage tries to balance this

Summary



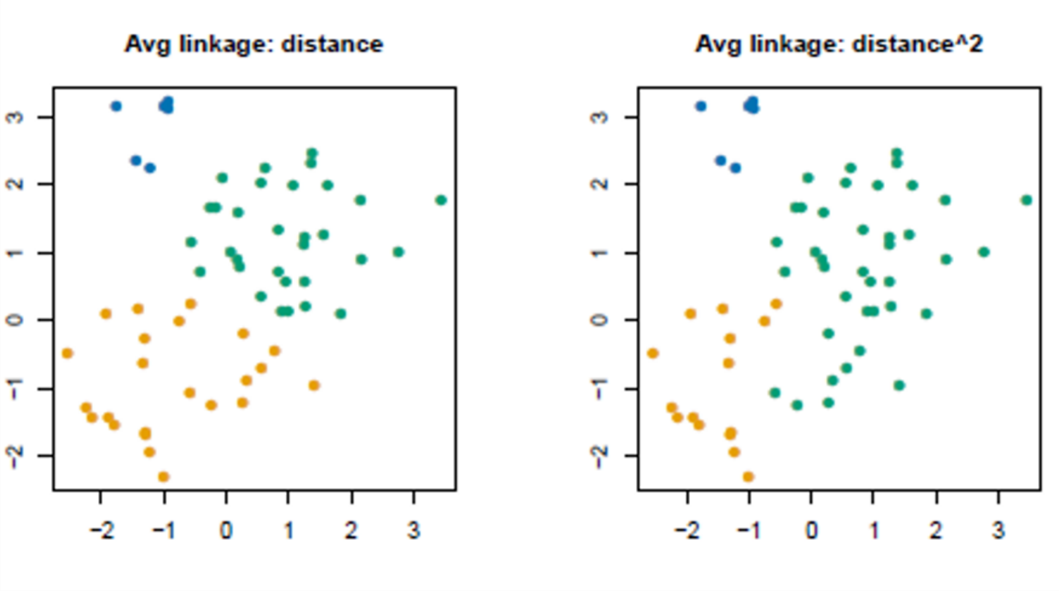
Disadvantages of average linkage

- the dendrogram does not give a nice interpretation of the cut
- Results can change
- **Example:** Apply a monotone increasing transformation to the dissimilarity measure

$$d \to d^2 \qquad \text{or} \qquad d \to \frac{e^d}{1+e^d}$$

- This is problematic when not sure of which measure to use
- Single and complete do not have this problem

Example



Dissimilarity Measure

- The choice of linkage has strong effects on structure and quality of resulting clusters
- The choice of the dissimilarity/similarity measure is as or more important than the linkage

Dissimilarity Measure

- The choice of linkage has strong effects on structure and quality of resulting clusters
- The choice of the dissimilarity/similarity measure is as or more important than the linkage
- Using a dissimilarity measure is easy.
- Finding the good measure is the real challenge

Dissimilarity Measure

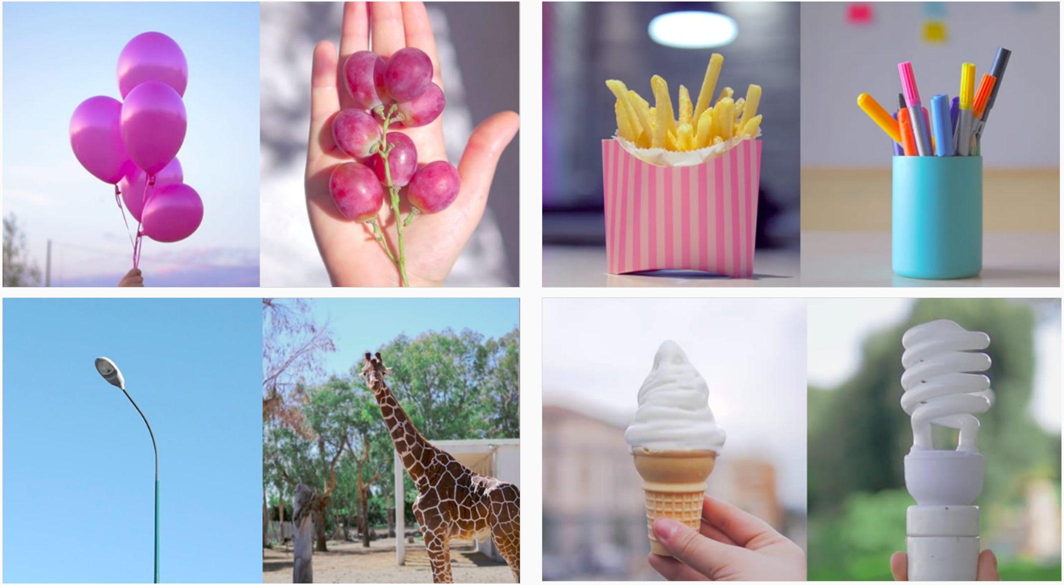
- The choice of linkage has strong effects on structure and quality of resulting clusters
- The choice of the dissimilarity/similarity measure is as or more important than the linkage
- Using a dissimilarity measure is easy.
- Finding the good measure is the real challenge
- Warning: In the literature, you may find both references to similarity or dissimilarity measures. Large dissimilarity implies, small similarity and vice versa.

Dissimilarity Measure

- The choice of linkage has strong effects on structure and quality of resulting clusters
- The choice of the dissimilarity/similarity measure is as or more important than the linkage
- Using a dissimilarity measure is easy.
- Finding the good measure is the real challenge
- Warning: In the literature, you may find both references to similarity or dissimilarity measures. Large dissimilarity implies, small similarity and vice versa.

Philosophical question:

What does it mean for two observations to be similar?



Source: `https://mymodernmet.com/mirror-short-film-tanello-production/`

Clustering: Comparison

K-means

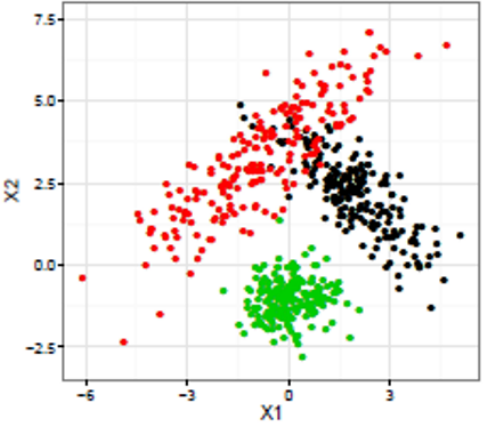
- Low memory usage
- Good implementation: O(n)
- Sensitive to initialization
- Number of clusters is predefined
- No categorical variables

Hierarchical Clustering

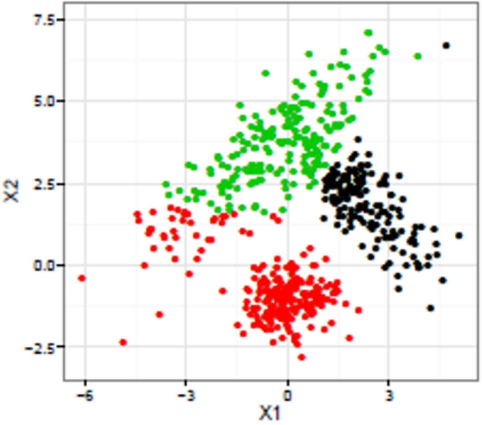
- Computationally expensive
- Dendrogram for visualization
- Deterministic
- Number of clusters can be varied
- Can handle missing and categorical values

Practical Tips on Clustering

Scaling

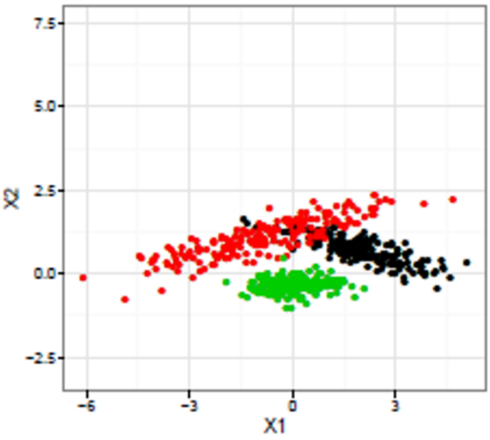


Labeled data

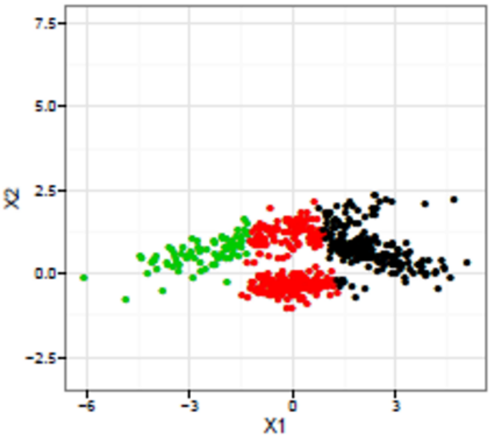


$K = 3$-means clustering

Re-scaling



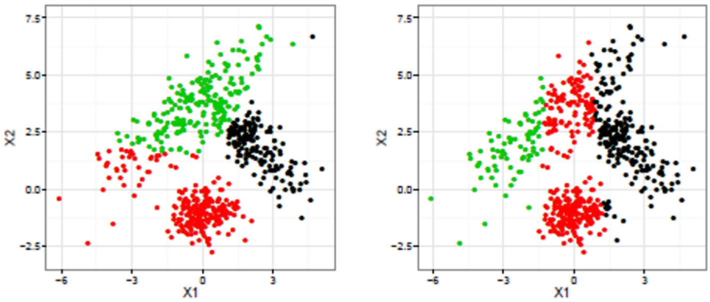
Labeled rescaled data



K-means clustering

Comparing Results

- For easier comparison of results rescale your data
- **Tip:** rescale all variables to have variance 1



K-means from *unscaled* data K-means from *scaled* data

Wrap-up

Wrap-up

- We reviewed two clustering techniques: K-means and hierarchical clustering
- We sketched the expectation maximization algorithm
- We reviewed different linkage techniques and studied their pro's and cons
- We compared both studied clustering techniques and saw some tips on how to use them

Key Concepts

- Dendogram
- Similarity/dissimilarity measure
- Linkage

References

Further Reading and Useful Material

Source	Notes
The Elements of Statistical Learning	Section 14.3