Machine Learning Applications for Finance

Clustering

Hao Xing

Unsupervised learning

Unsupervised vs Supervised Learning

- You have learned many supervised learning methods before such as regression and classification
- In that setting we observe both a set of features $X_1, X_2, \ldots, X_p$ for each object, as well as a response outcome variable Y. The goal is to predict Y using $X_1, X_2, \ldots, X_p$
- Here we focus on unsupervised learning, we observe only the features $X_1, X_2, \ldots, X_p$. We are not interested in prediction, because we do not have the associated response variable Y.

Goals of unsupervised learning

The goal is to discover interesting things about the measurements: is there any informative way to visualize the data? Can we discover subgroups among the variables among the obervations?

Clustering: a broad class of methods for discovering unknown subgroups in data

- K-means clustering
- Hierarchical clustering

Challenge of unsupervised learning

Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response

But techniques for unsupervised learning are of growing importance in a number of fields

- Groups of shoppers characterized by their browsing and purchase histories
- movies grouped by the ratings assigned by movie viewers
- Compliance documents grouped by different divisions in banks

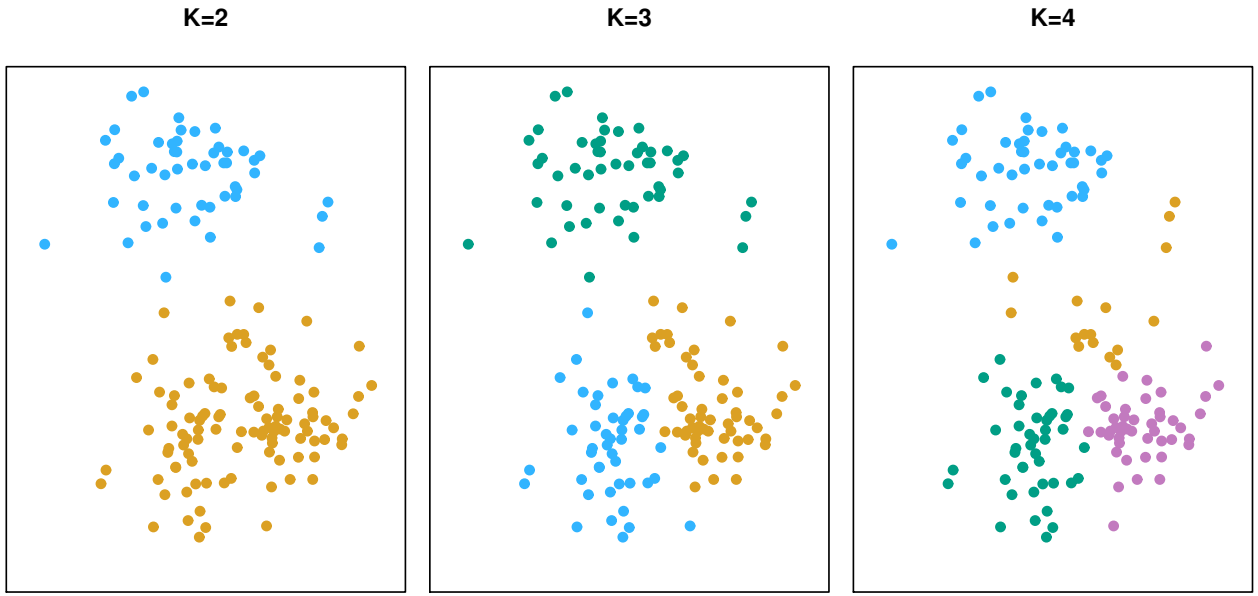
Another advantage: it is often easier to obtain unlabeled data than labeled data, which require human intervention

Two clustering methods

K-means clustering: we seek to partition the observations into K (pre-specified) clusters

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

K-means clustering



Results of applying K-means clustering with different values of K, the number of clusters.

There is no ordering of the clusters, so that the cluster coloring is arbitrary.

Coloring is the outputs of the clustering procedure.

Details of K-means clustering Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

1. $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, n\}$. Each observations belongs to at least one of the K clusters.
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. No observation belongs to more than one cluster.

For instance, if the i-th observation is in the k-th cluster, then $i \in C_k$.

- 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

$$\min_{C_1,\ldots,C_k} \left\{ \sum_{k=1}^{K} WCV(C_k) \right\}. \tag{1}$$

- 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.

Within-cluster varation

Typically we use Euclidean distance

$$WCV(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{2}$$

where $|C_k|$ denotes the number of observations in the k-th cluster

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

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \tag{3}$$

K-means clustering 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
    1. For each of the K clusters, compute the cluster centroid. The k-th cluster centroid is the vector of the p feature means for the observations in the k-th cluster
    2. Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance)

Properties of the algorithm

This algorithm is guaranteed to decrease the value of the objective (3) at each step. Why?

Properties of the algorithm

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

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\overline{x}_{kj})^2,$$

where $\overline{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ is the mean for feature j in the cluster C_k.

However, it is not guaranteed to give the global minimum Why?

Properties of the algorithm

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

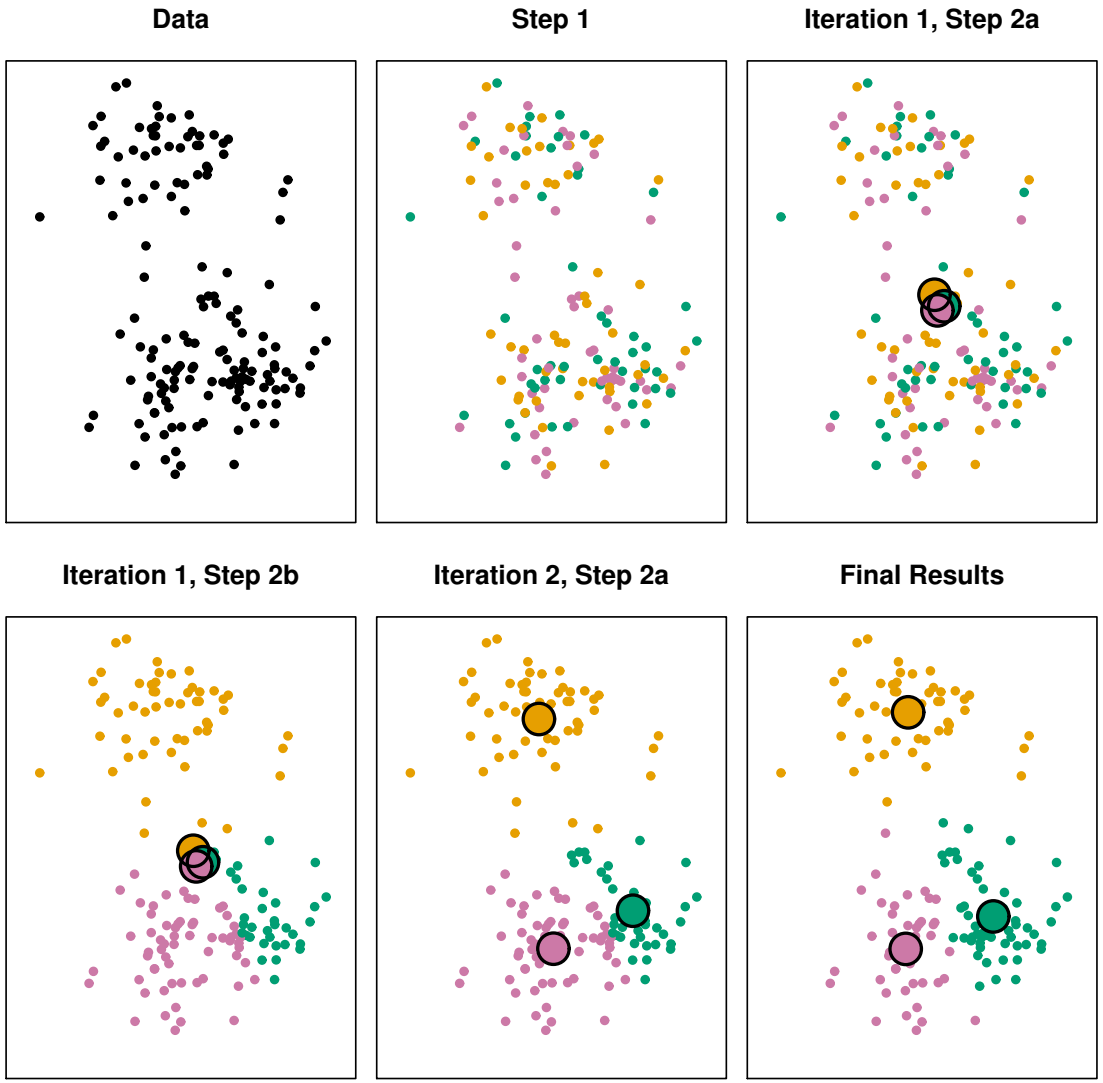
$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\overline{x}_{kj})^2,$$

where $\overline{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ is the mean for feature j in the cluster C_k.

However, it is not guaranteed to give the global minimum Why?

The result depends on starting values

Example


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Example: different starting values

Details of previous figure

K-means clustering performed six times on the data from previous figure with $K = 3$, each time with a different random assignment of the observations in Step 1 of the algorithm

Above plot is the value of the objective (3)

Three different local optima were obtained, one of which resulted in a smaller value of the objective and provides best separation between the clusters

Those labeled in red all achieved the same best solution, with an objective value of 235.8

Hierarchical clustering

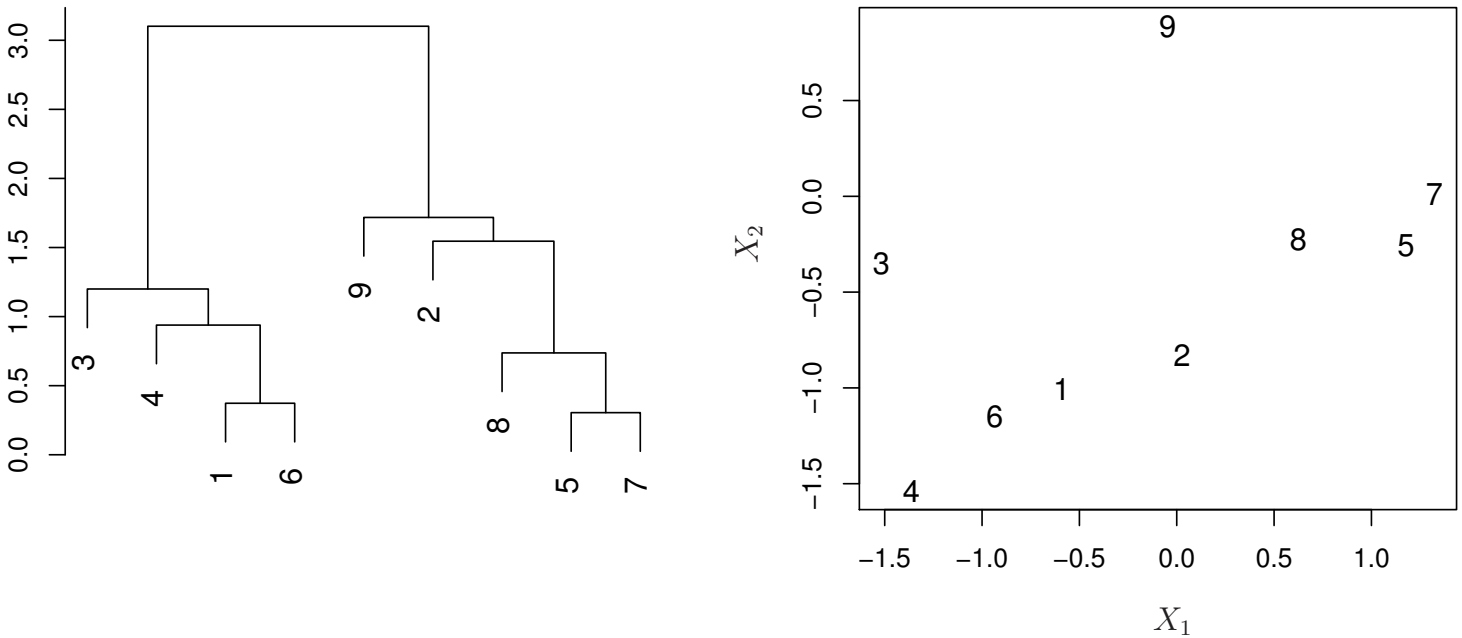
K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage

Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.

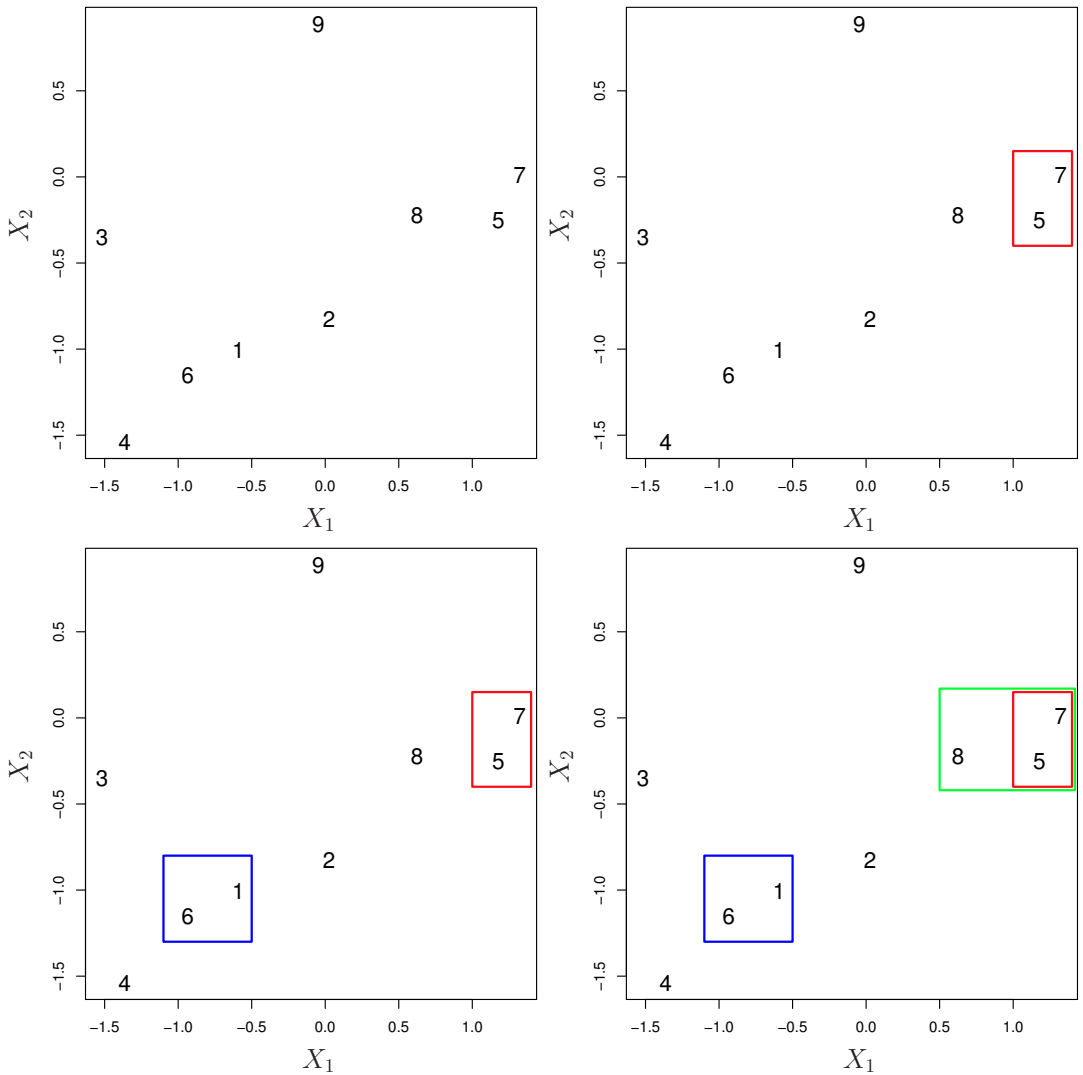
We describe bottom-up clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

Hierarchical clustering algorithm

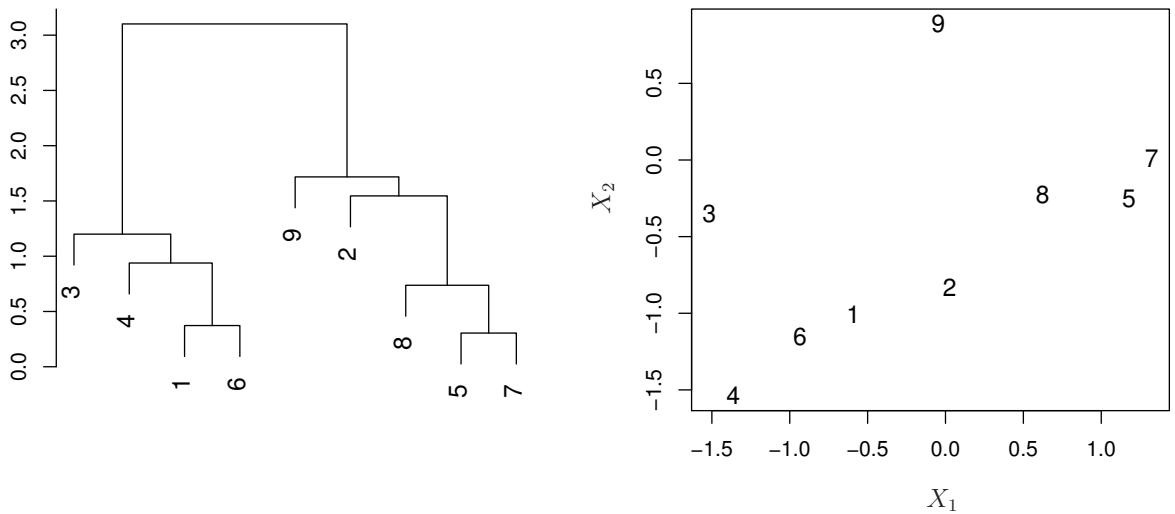
- Start with each point in its own cluster
- Identify the closest two clusters and merge them
- Repeat
- Ends when all points are in a single cluster



Merge in the previous example



Number of clusters



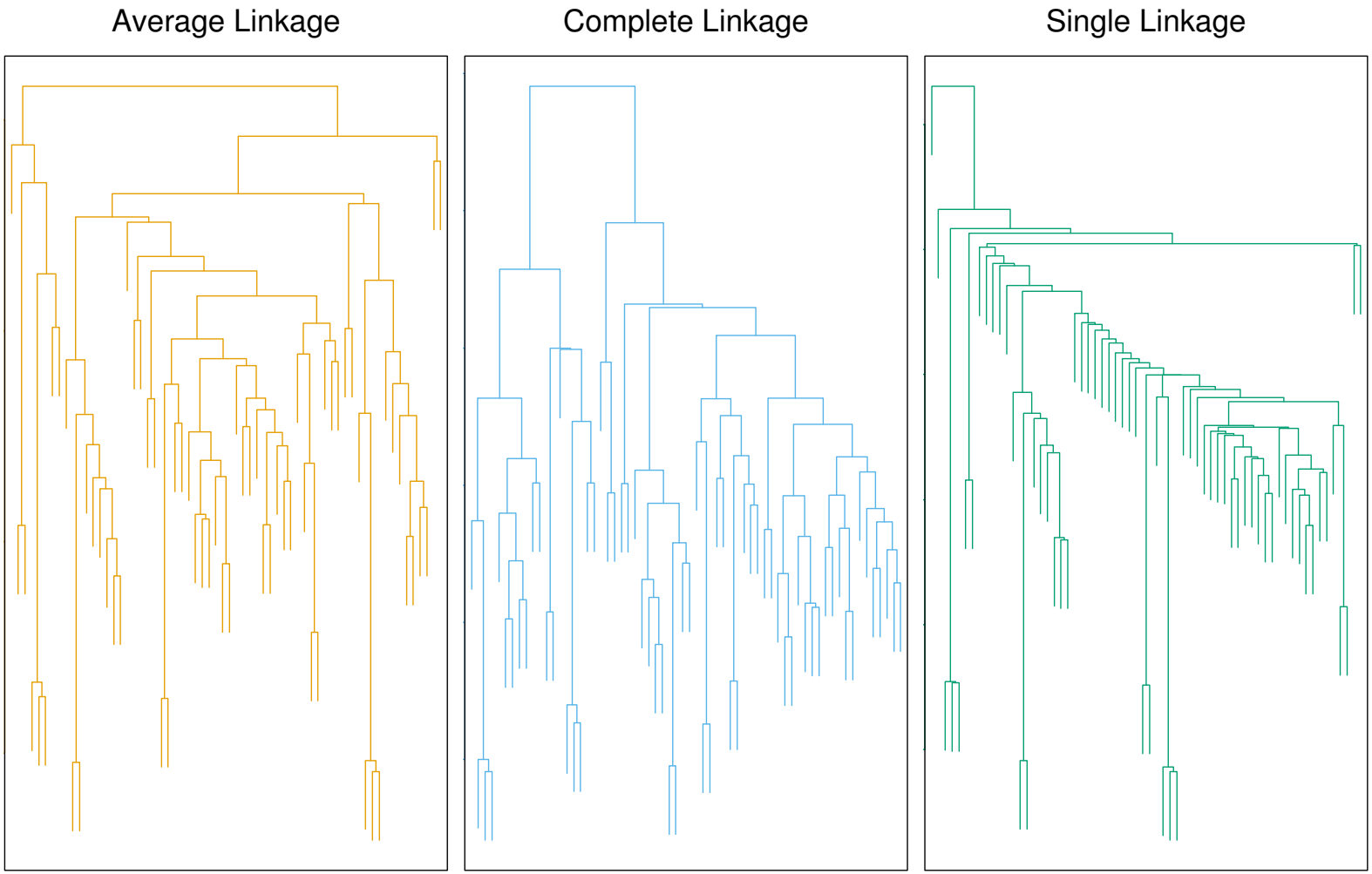
- Horizontal distance does not matter
- Vertical hight measures the difference between different cluster.
- Vertical hight can give cluster for a given K. For example, $K = 2$, cut at 2.5; $K = 3$ a cut at 1.6.

Distance between clusters: 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. smallest pairwise dissimilarities
Average	Mean inter-cluster dissimilarity. average of all pairwise dissimilarities
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Dissimilarities between centroid.

Different linkage

Average and Complete are generally preferred. They tend to yield more balanced dendrograms.

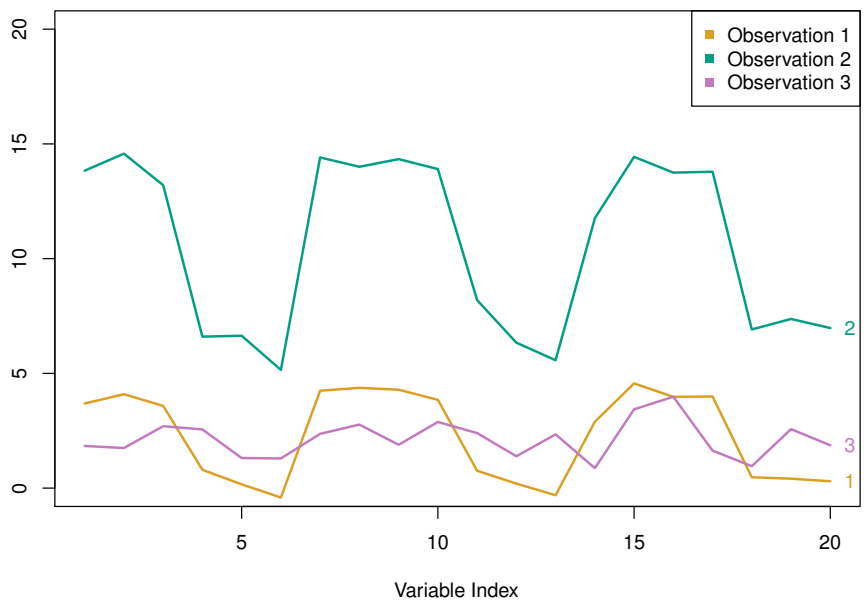


Choice of dissimilarity measure

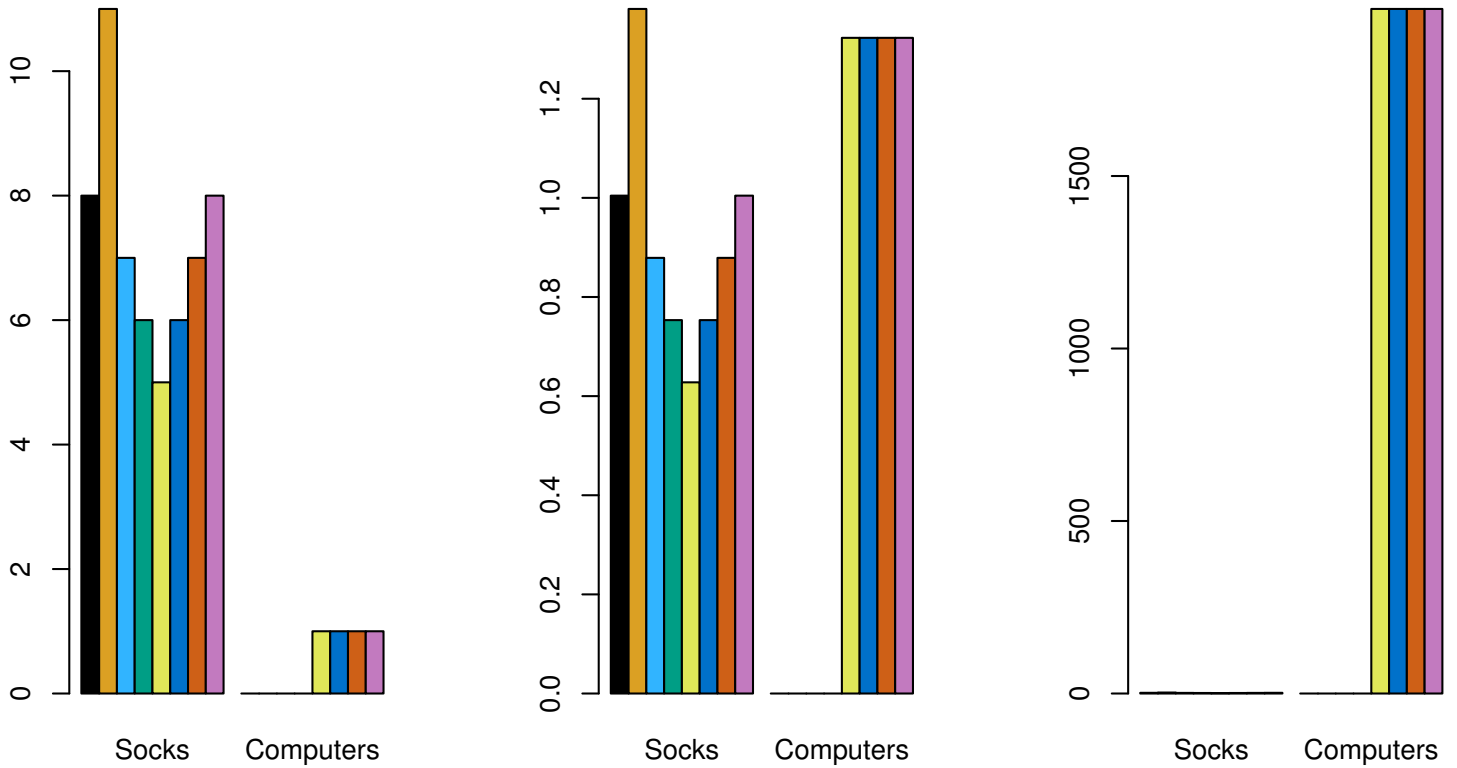
So far we have used Euclidean distance

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



Scaling of the variables matters



We may need to rescale or standarize variables depending on the goal of our model.

Practical issues

- 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?

Practical issues

- 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?

These are difficult questions. There is no single right answer — any solution that exposes some interesting aspects of the data should be considered.