Eleventh Session

Alireza Moradi

| Linkedin | |

Unsupervised Learning

Unsupervised Learning

- In **supervised learning**, we observe both a set of features $X_1, X_{2,\ldots,} X_p$ for each object, as well as a response or outcome variable Y. The goal is then to predict Y using $X_1, X_{2,\ldots,} X_p$.

- In **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 an associated response variable Y.

- The **main goal of unsupervised learning** is to discover hidden and interesting patterns in unlabeled data.

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.

- It is intrinsically **more difficult** than supervised learning because there is no gold standard and no single objective (like test set accuracy).

- It is often **easier to obtain unlabeled data** than labeled data, which can require human intervention.

    - Example: Movies grouped by the ratings assigned by movie viewers.

- Unsupervised learning can be a useful pre-processor for supervised learning.

Clustering

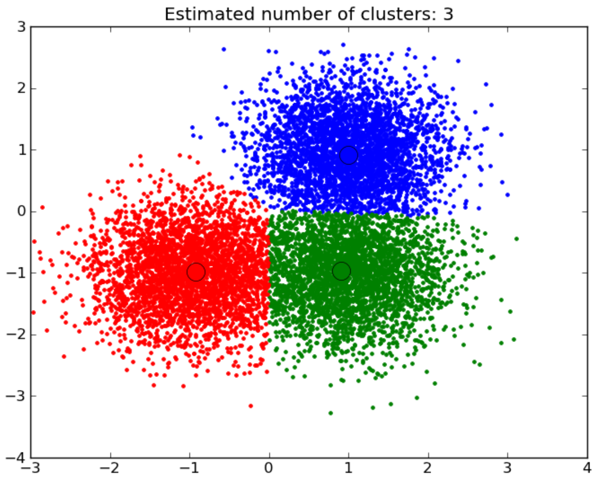
Clustering

- **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 like each other.

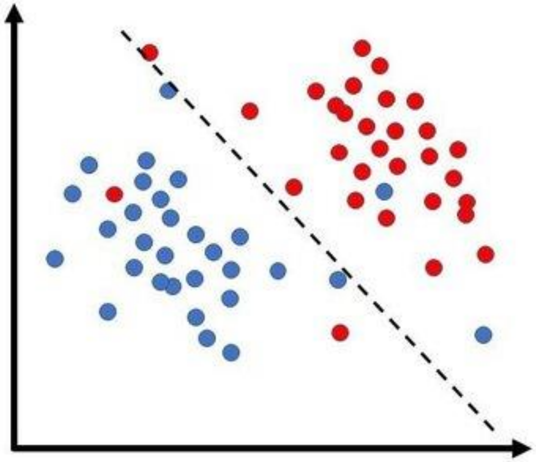
- We must define what it means for two or more observations to be **similar** or **different**.

- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

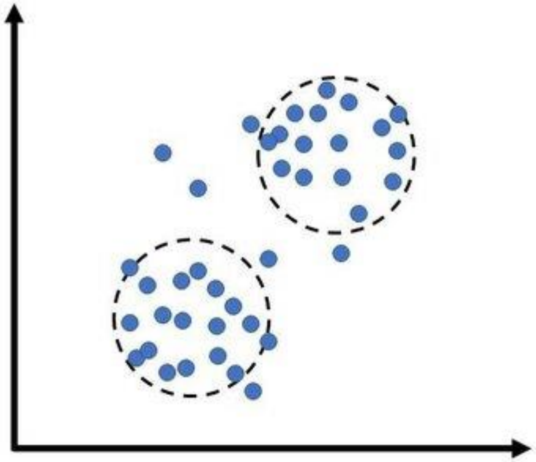


Clustering vs Classification

Classification



Clustering



K-means

K-means

- The idea behind K-means clustering is that **a good clustering** is one for which the within-cluster variation is as small as possible.

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.

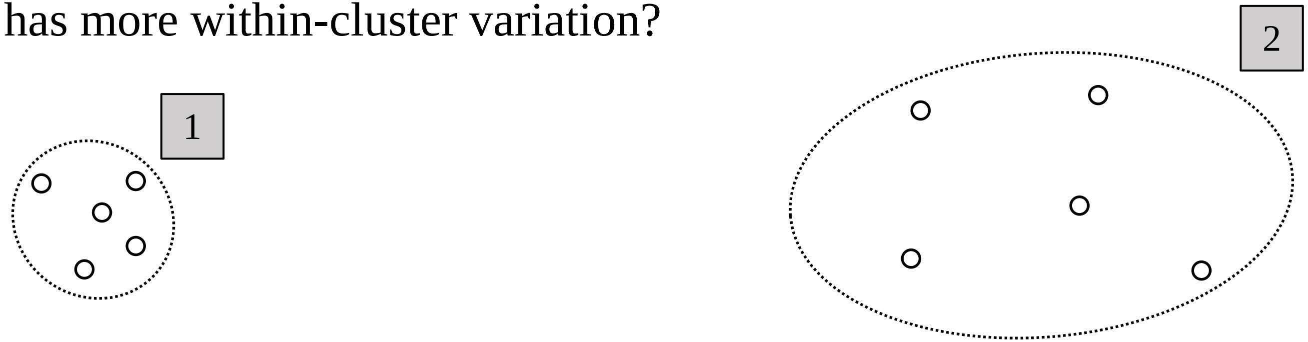
- The within-cluster variation for cluster C_k is a measure of the amount by which the observations within a cluster differ from each other. ($\mathbf{WCV}(\boldsymbol{C_k})$)

- Hence, we want to solve this optimization problem:

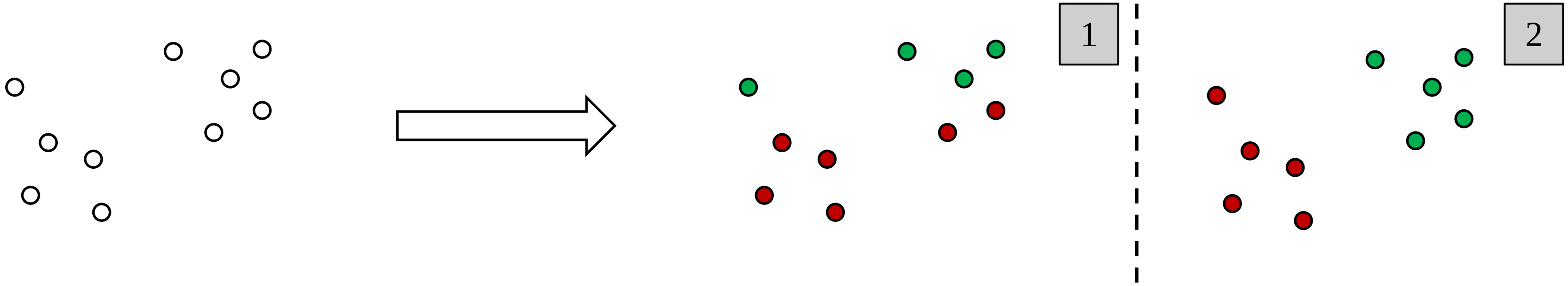
$$\min_{C_1, C_2, \ldots, C_k} \left\{ \sum_{k=1}^{K} WCV(C_k) \right\}$$

K-means

- ✓ Which cluster has more within-cluster variation?



- ✓ Which clustering is better?



K-means

- Typically, we use Euclidean distance:

$$d(a, b) = \sqrt{\sum_{j=1}^{p} (a_j - b_j)^2}$$

- The within-cluster variation for cluster C_k will be:

$$WCV(C_k) = \frac{1}{|C_k|} \sum_{\forall a,b \in C_k} \sum_{j=1}^{p} (a_j - b_j)^2$$

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

K-means

- The optimization problem that defines K-means clustering:

$$\min_{C_1,C_2,\ldots,C_k}\left\{\sum_{k=1}^{K} WCV(C_k)\right\}$$

$$\Downarrow$$

$$\min_{C_1,C_2,\ldots,C_k}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{\forall a,b\in C_k}\sum_{j=1}^{p}\left(a_j-b_j\right)^2\right\}$$

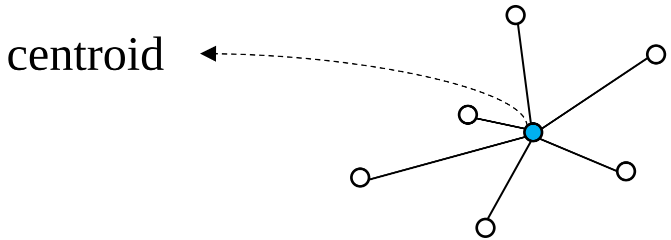
K-Means Algorithm

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:

   i. 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.

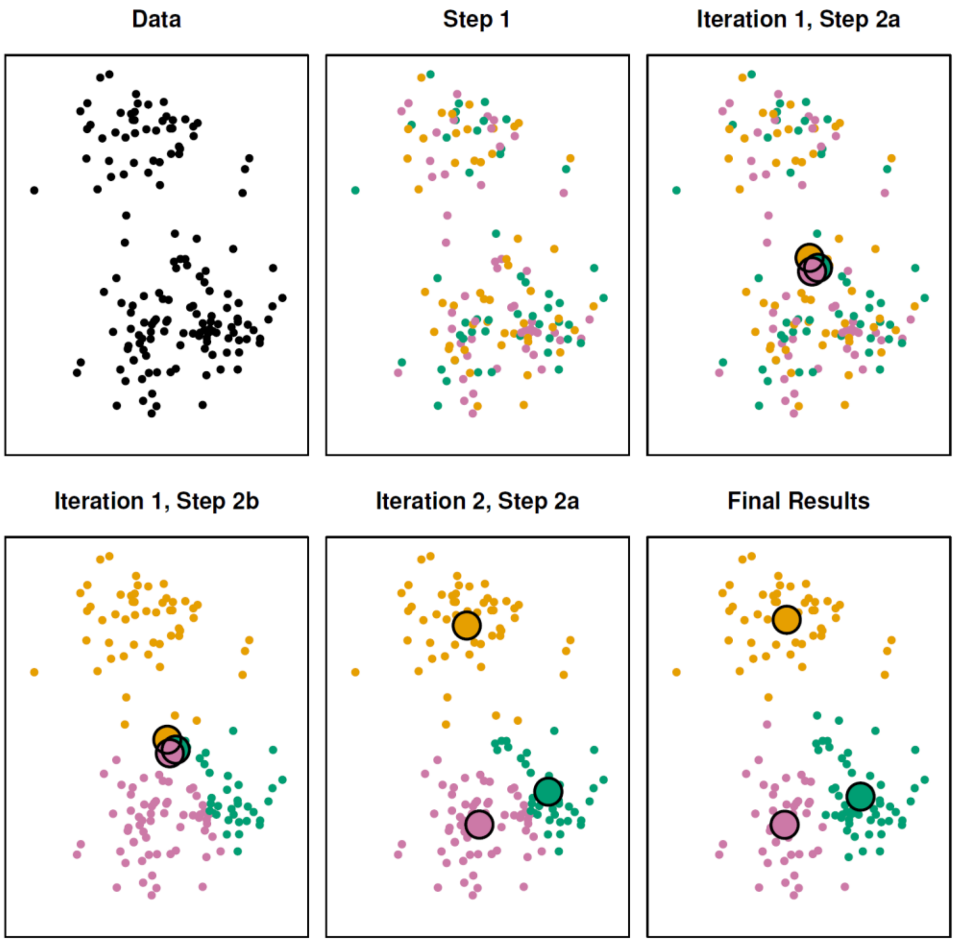
   ii. Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).



K-Means Algorithm

- The progress of the K-means algorithm with K=3.

- The results obtained after 10 iterations.



K-means algorithm

- The optimization problem that defines K-means clustering:

$$\frac{1}{|C_k|}\sum_{\forall a,b\in C_k}\sum_{j=1}^{p}\left(a_j-b_j\right)^2=2\sum_{\forall a\in C_k}\sum_{j=1}^{p}\left(a_j-\bar{a}_j\right)^2$$

$$\bar{a}_j=\frac{1}{|C_k|}\sum_{\forall a\in C_k}a_j$$

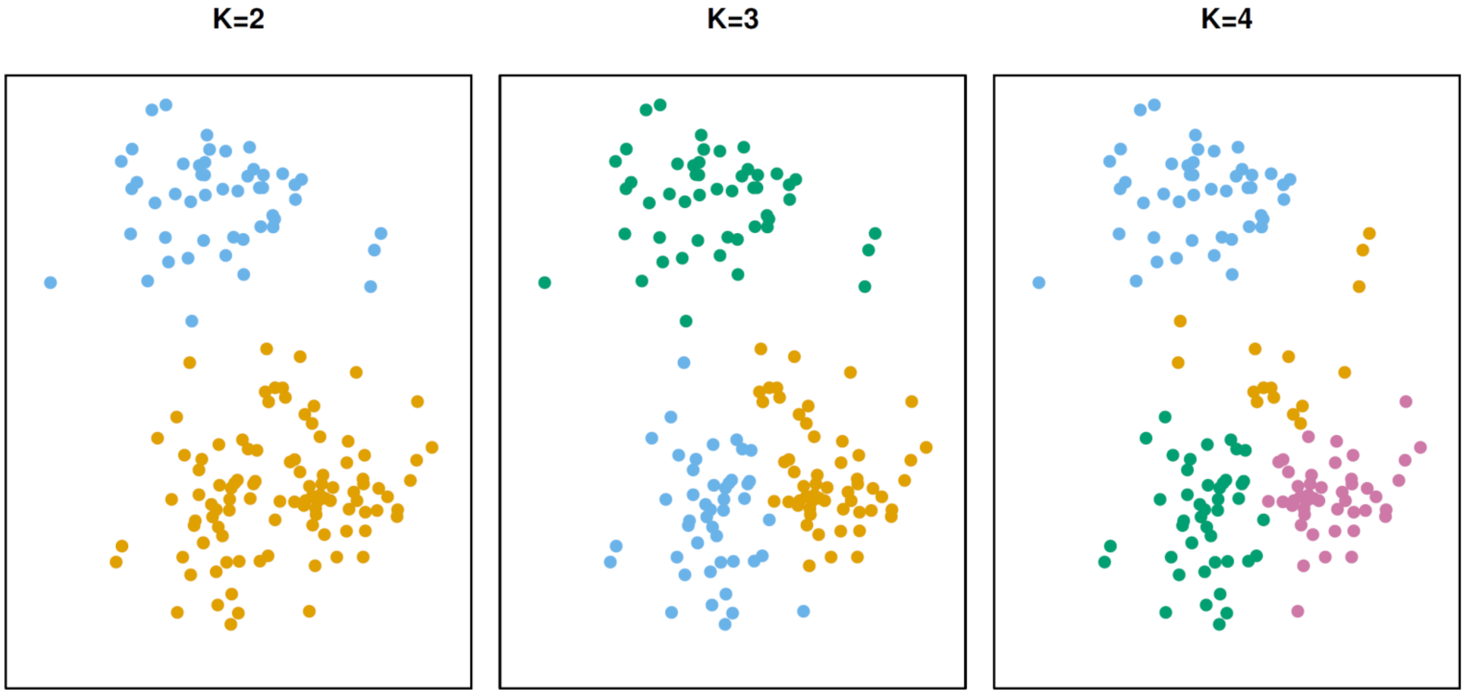
Different initializations

- 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.

✓ Which is the best clustering?

Different K in K-means

The color of each observation indicates the cluster to which it was assigned.

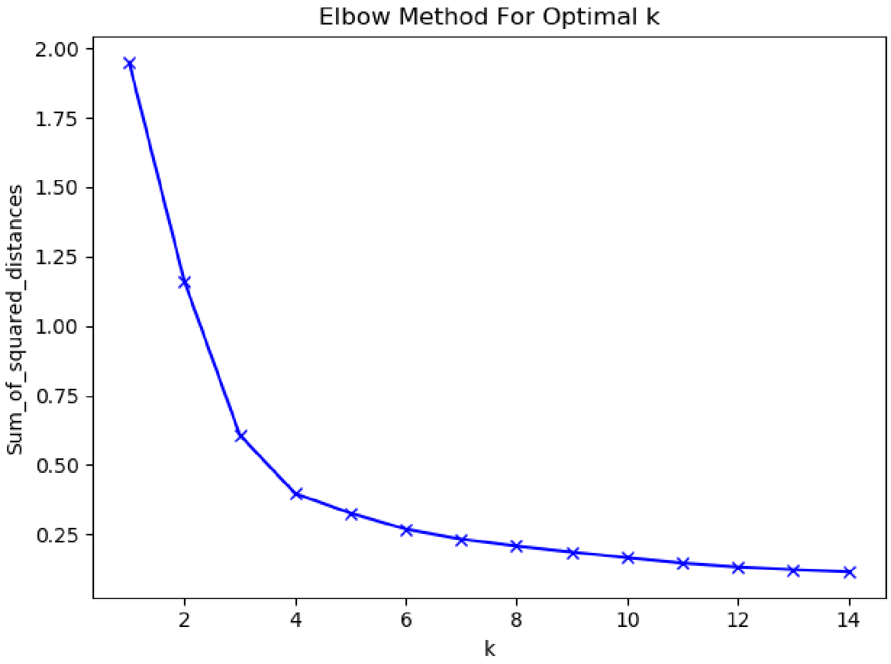


Choosing K

- The **elbow method** is a graphical method for finding the optimal K value in a k-means clustering algorithm.

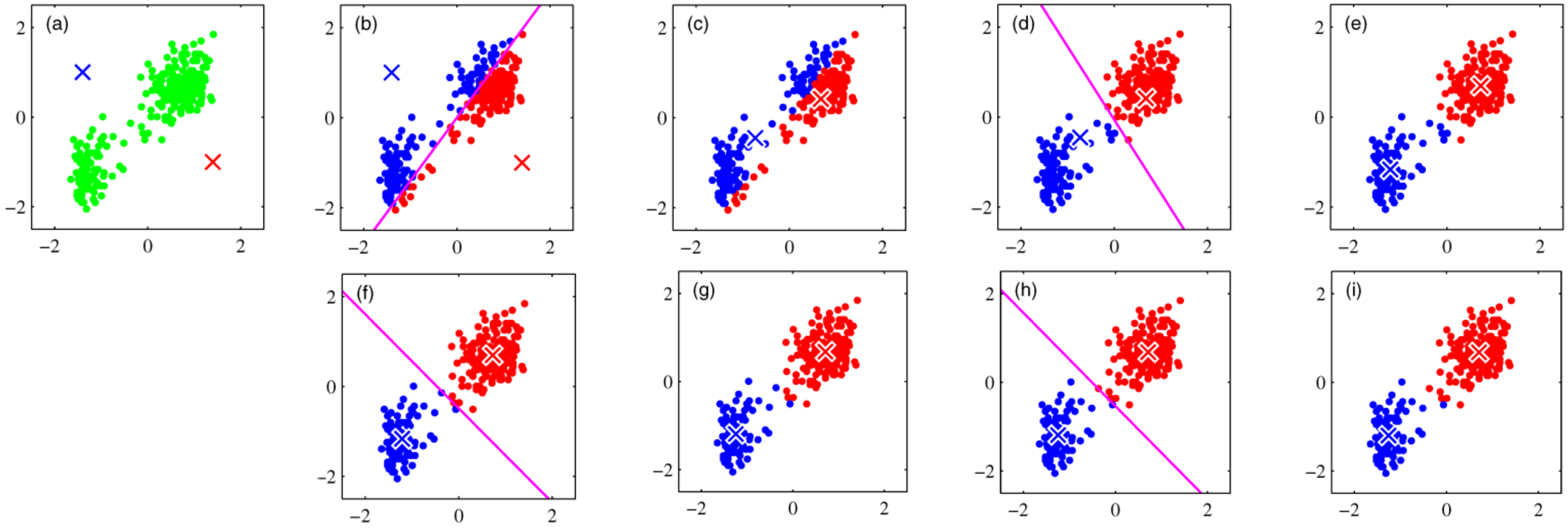
- The elbow graph shows the within-cluster-sum-of-square (WCSS) values on the y-axis corresponding to the different values of K (on the x-axis).

- The **optimal K value** is the point at which the graph forms an elbow.



K-means

- We can also perform k-means clustering with better initialization techniques.



✓ How can you make this even better?

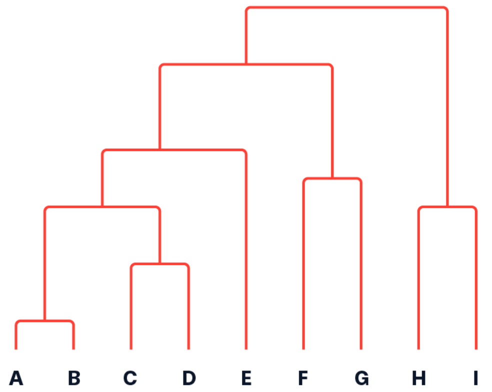
Hierarchical clustering

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.

- 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 clustering obtained for each possible number of clusters, from 1 to n.

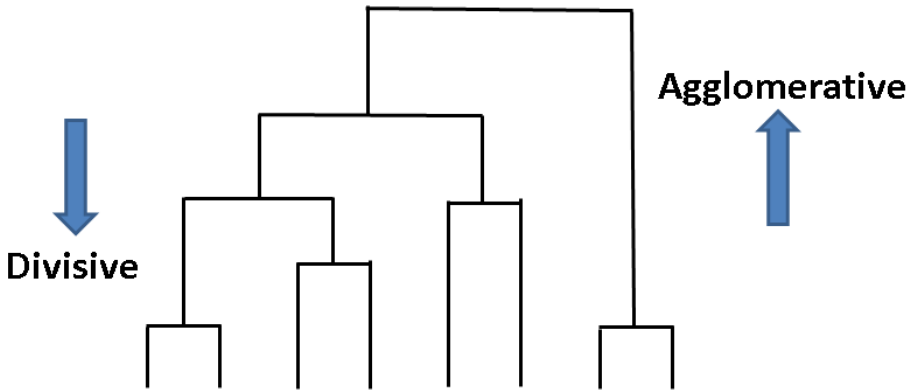


Hierarchical clustering

- There are two major types of approaches in hierarchical clustering; (1) **Divisive** & (2) **agglomerative.**

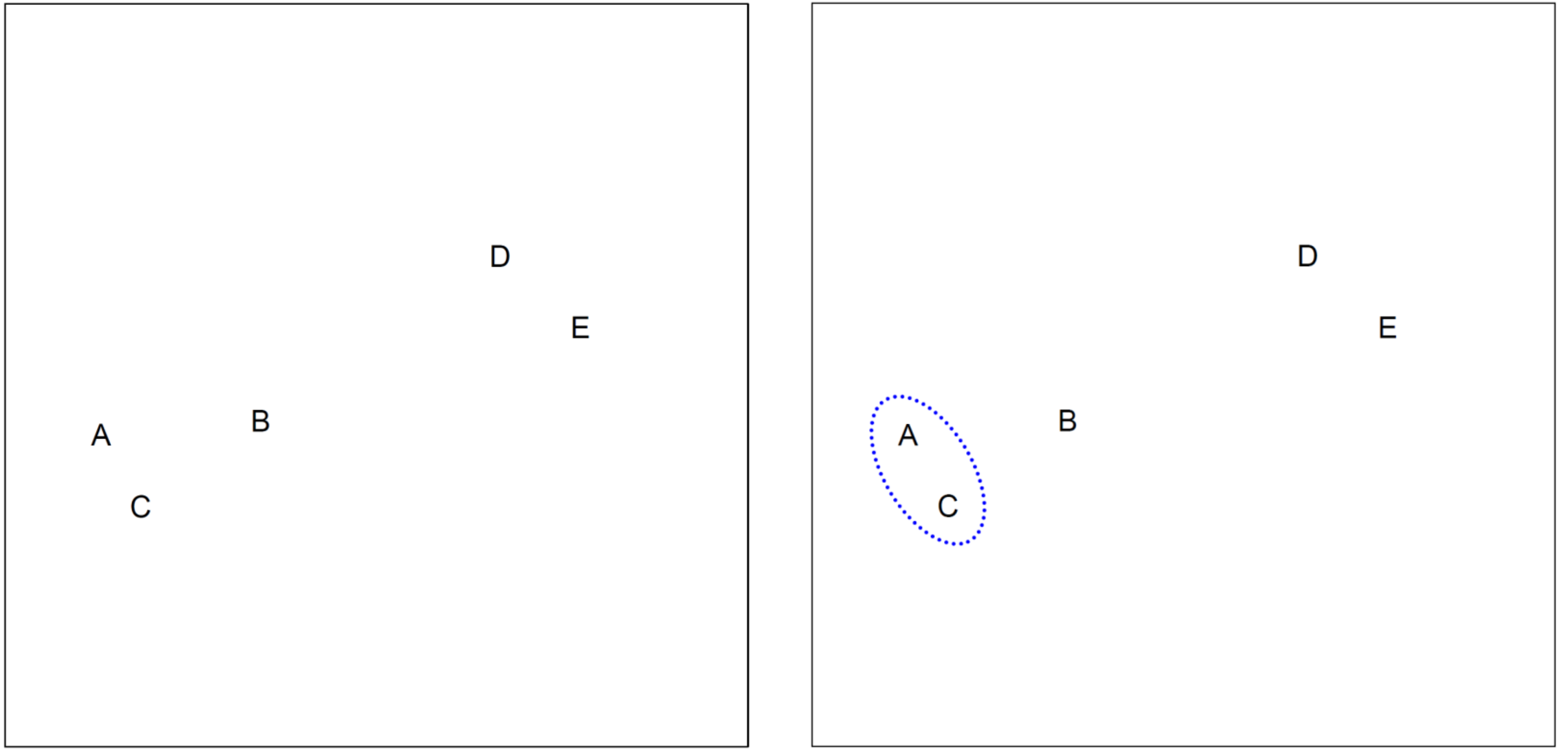
- In this section, we describe **bottom-up** or **agglomerative clustering** which refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

- This is the most common type of hierarchical clustering.

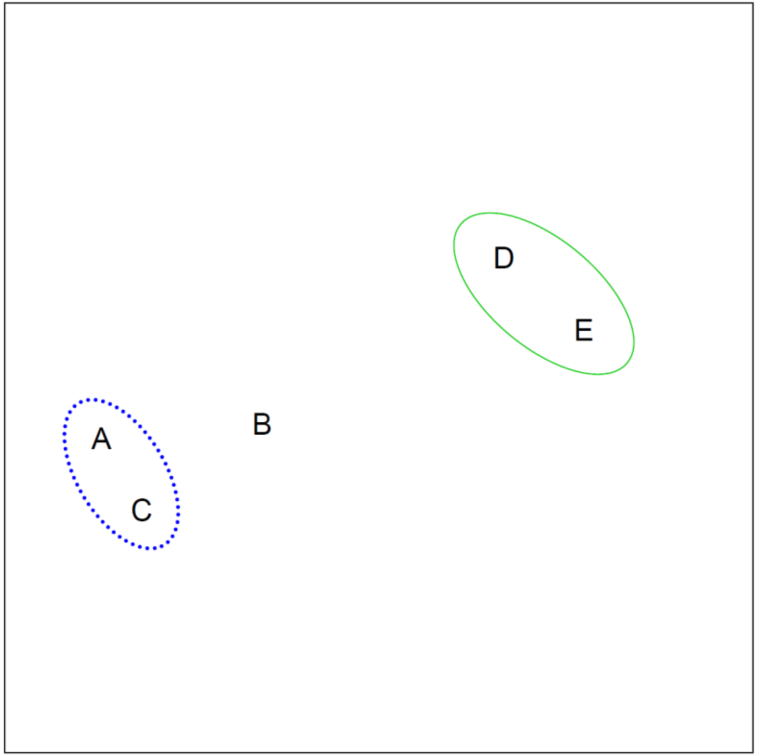


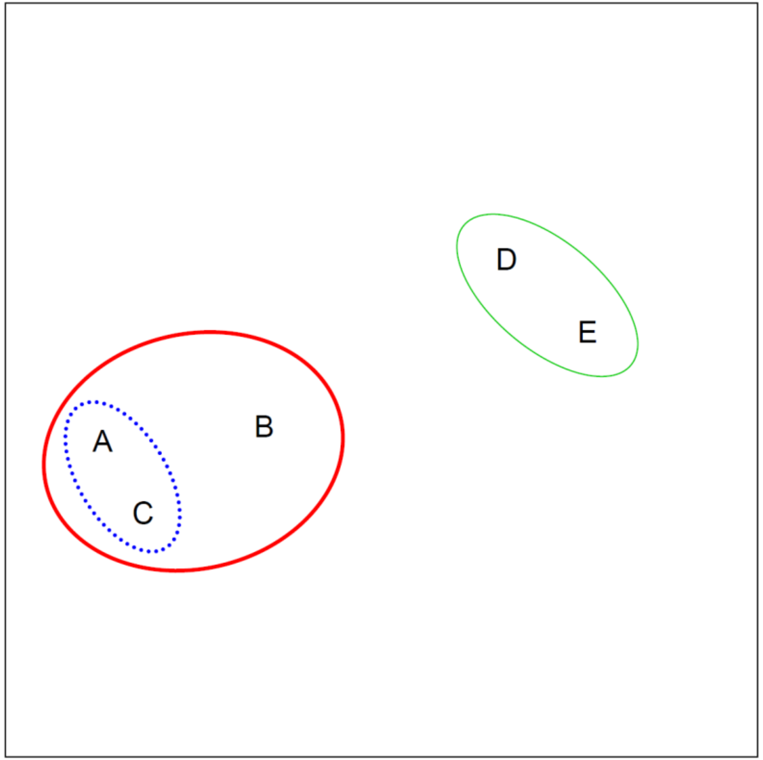
Hierarchical clustering; bottom-up

Example:



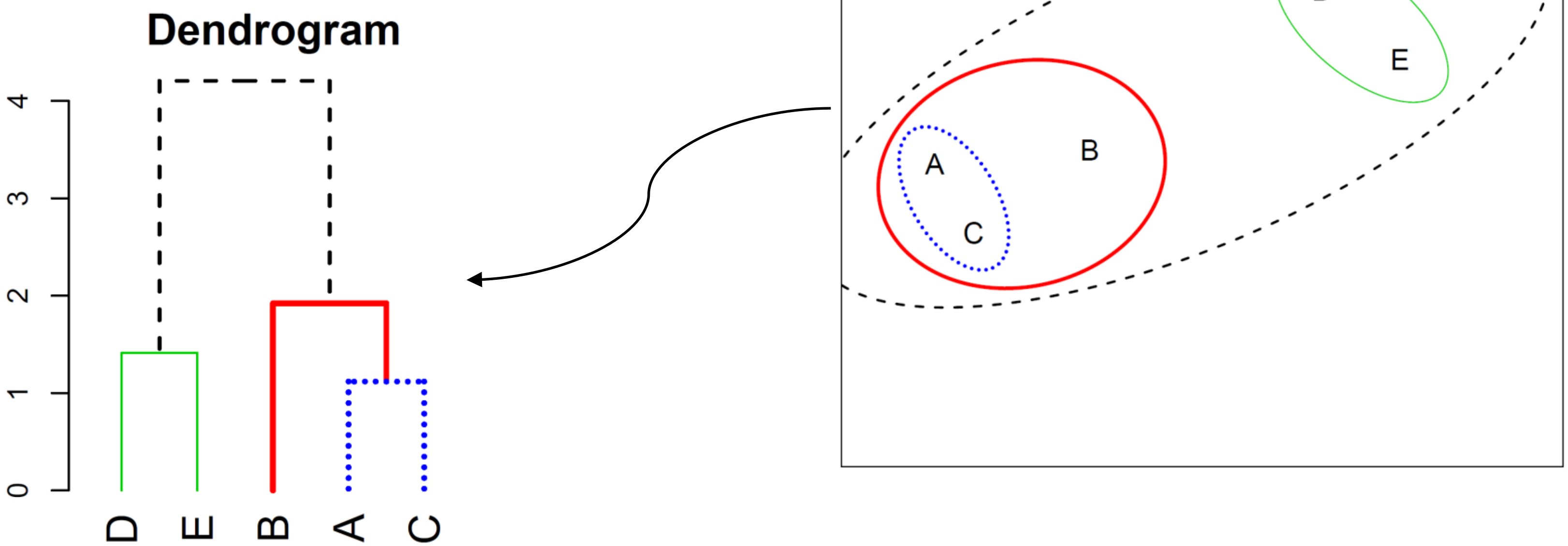
Hierarchical clustering; bottom-up





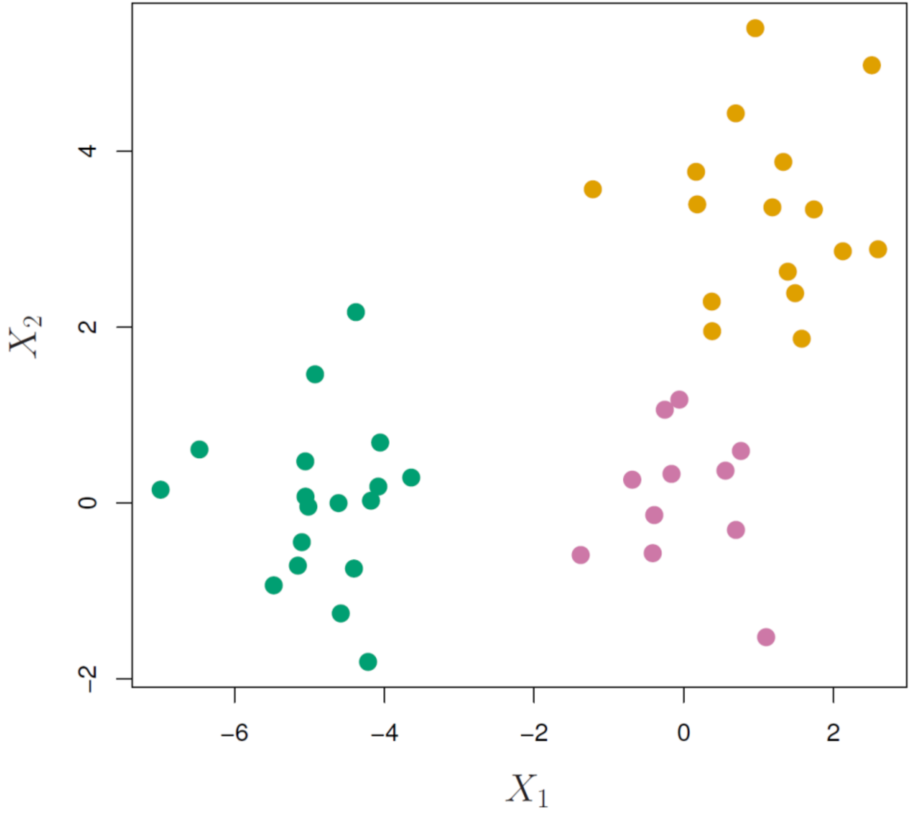
Hierarchical clustering; bottom-up

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



Hierarchical clustering; bottom-up

- Example:



Hierarchical clustering; bottom-up