Lecture 24 Cluster Analysis I

ECE 625: Data Analysis and Knowledge Discovery

Di Niu

Department of Electrical and Computer Engineering
University of Alberta

April 8, 2021

Outline

Unsupervised Learning

- Most of this course focuses on supervised learning methods such as regression and classification.
- In that setting we observe both a set of features $X_1, \cdots, X_p$ for each object, as well as a response or outcome variable Y. The goal is then to predict Y using $X_1, \cdots, X_p$.
- Here we instead focus on unsupervised learning, we where observe only the features $X_1, \cdots, X_p$.
- We are not interested in prediction, because we do not have an associated response variable Y.

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:
  - subgroups of breast cancer patients grouped by their gene expression measurements,
  - groups of shoppers characterized by their browsing and purchase histories,
  - movies grouped by the ratings assigned by movie viewers.
- The method is largely exploratory, with the intention of following up with a more detailed analysis of the groups.

Unsupervised Learning

- It is often easier to obtain unlabeled data — from a lab instrument or a computer — than labeled data, which can require human intervention.
- For example it is difficult to automatically assess the overall sentiment of a movie review: is it favorable or not?
- The goal of unsupervised learning is to discover interesting things about the measurements: is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?
- We discuss the following methods:
  - clustering, a broad class of methods for discovering unknown subgroups in data
  - principal components analysis, a tool used for data visualization or data pre-processing before supervised techniques are applied,

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,
- To 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.
- Major methods include distances for continuous variables and similarity coefficients for discrete variables.

Distances Between Pairs of Items

- Common measures are $d(\mathbf{x}, \mathbf{y}) =$

$$\|\mathbf{x}-\mathbf{y}\| = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} \text{(Euclidean distance)},$$

$$\sqrt{(\mathbf{x}-\mathbf{y})^T \mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})} \text{ (Mahalanobis distance)},$$

$$\left(\sum |x_i - y_i|^m\right)^{1/m} \text{ (Minkowski distance; } L^m\text{)},$$

$$\sum |x_i - y_i| \text{ (City-block distance; } L^1\text{)}.$$

- True distances must satisfy the following properties:

$$\begin{aligned} \text{(i) } d(\mathbf{x}, \mathbf{y}) &\geq 0, \text{ equality iff } \mathbf{x} = \mathbf{y}, \\ \text{(ii) } d(\mathbf{x}, \mathbf{y}) &= d(\mathbf{y}, \mathbf{x}), \\ \text{(iii) } d(\mathbf{x}, \mathbf{z}) &\leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}). \end{aligned}$$

Market Segmentation

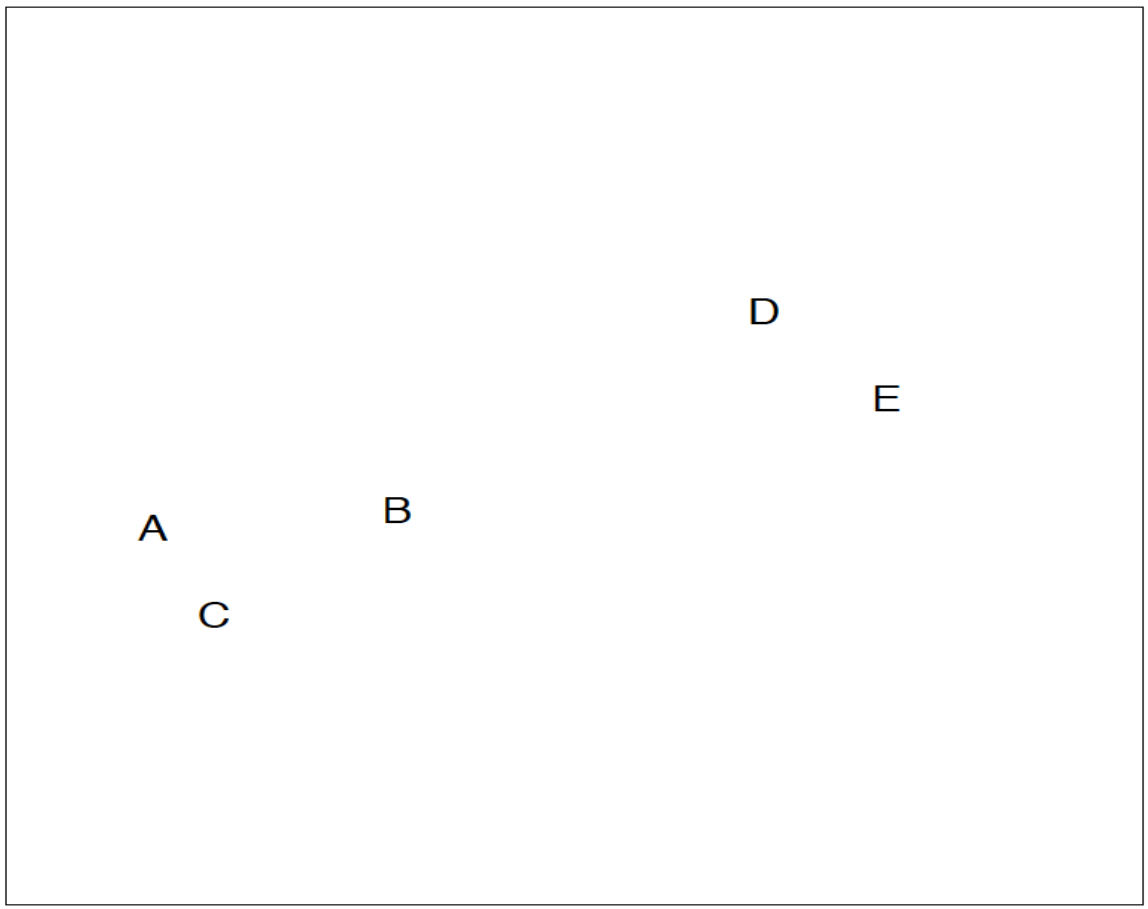
- Suppose we have access to a large number of measurements (e.g. median household income, occupation, distance from nearest urban area, and so forth) for a large number of people.
- Our goal is to perform market segmentation by identifying subgroups of people who might be more receptive to a particular form of advertising, or more likely to purchase a particular product.
- The task of performing market segmentation amounts to clustering the people in the data set.

Two clustering methods

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

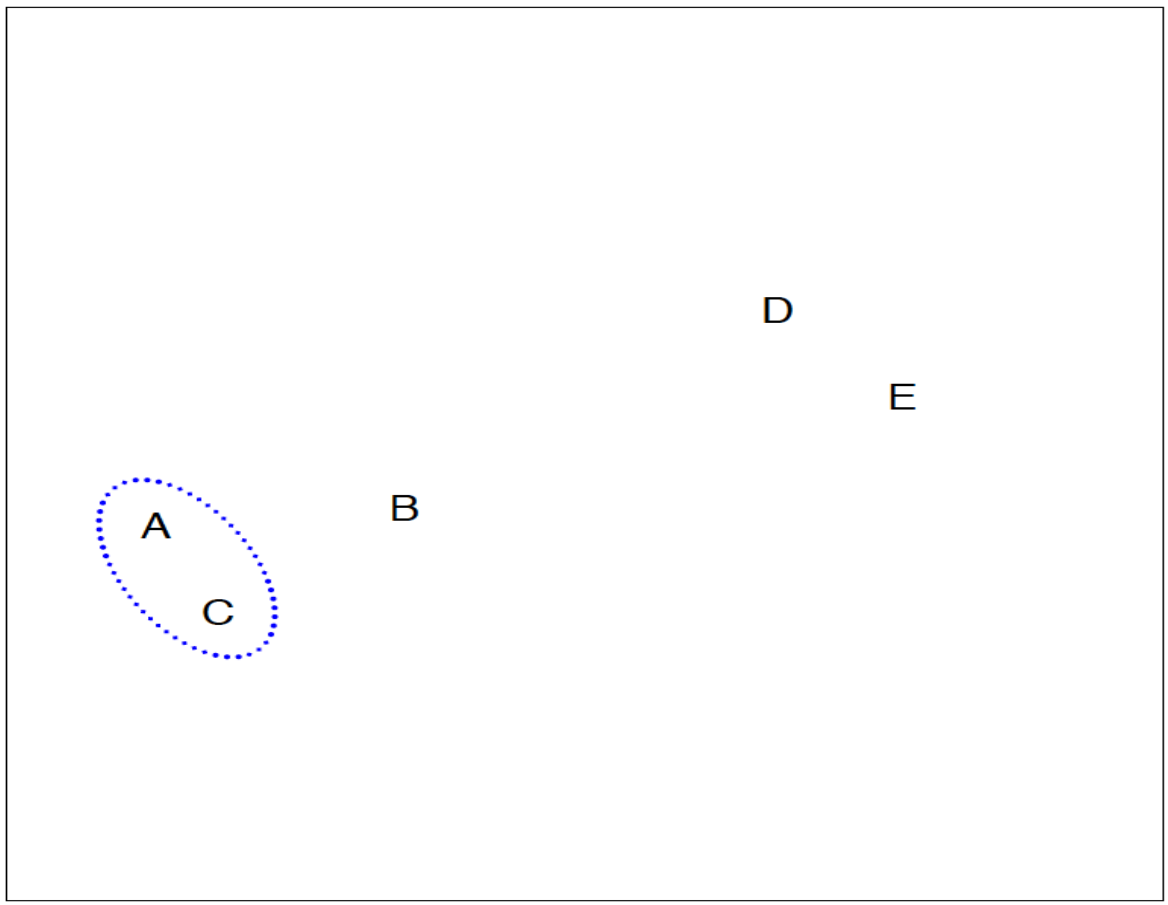
Hierarchical clustering

Builds a hierarchy in a bottom-up fashion...



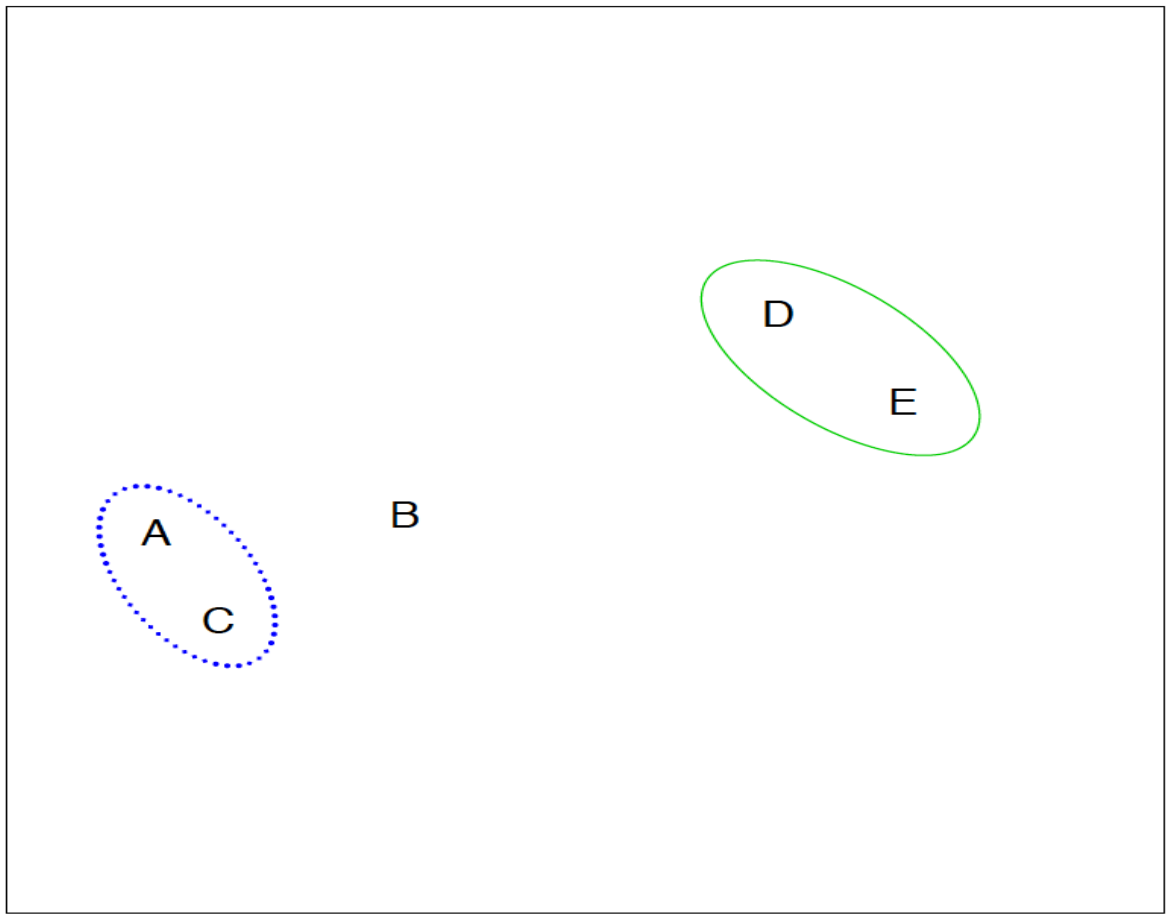
Hierarchical clustering

Builds a hierarchy in a bottom-up fashion...



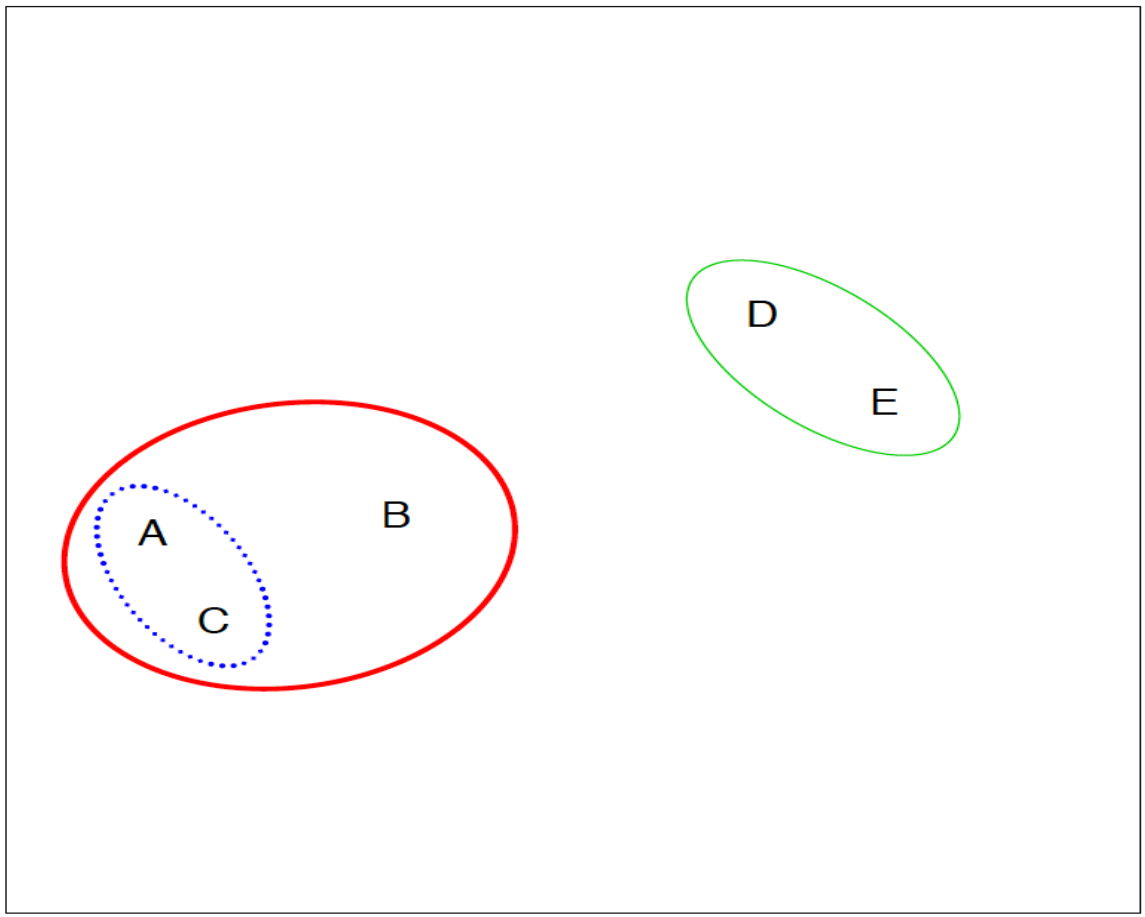
Hierarchical clustering

Builds a hierarchy in a bottom-up fashion...



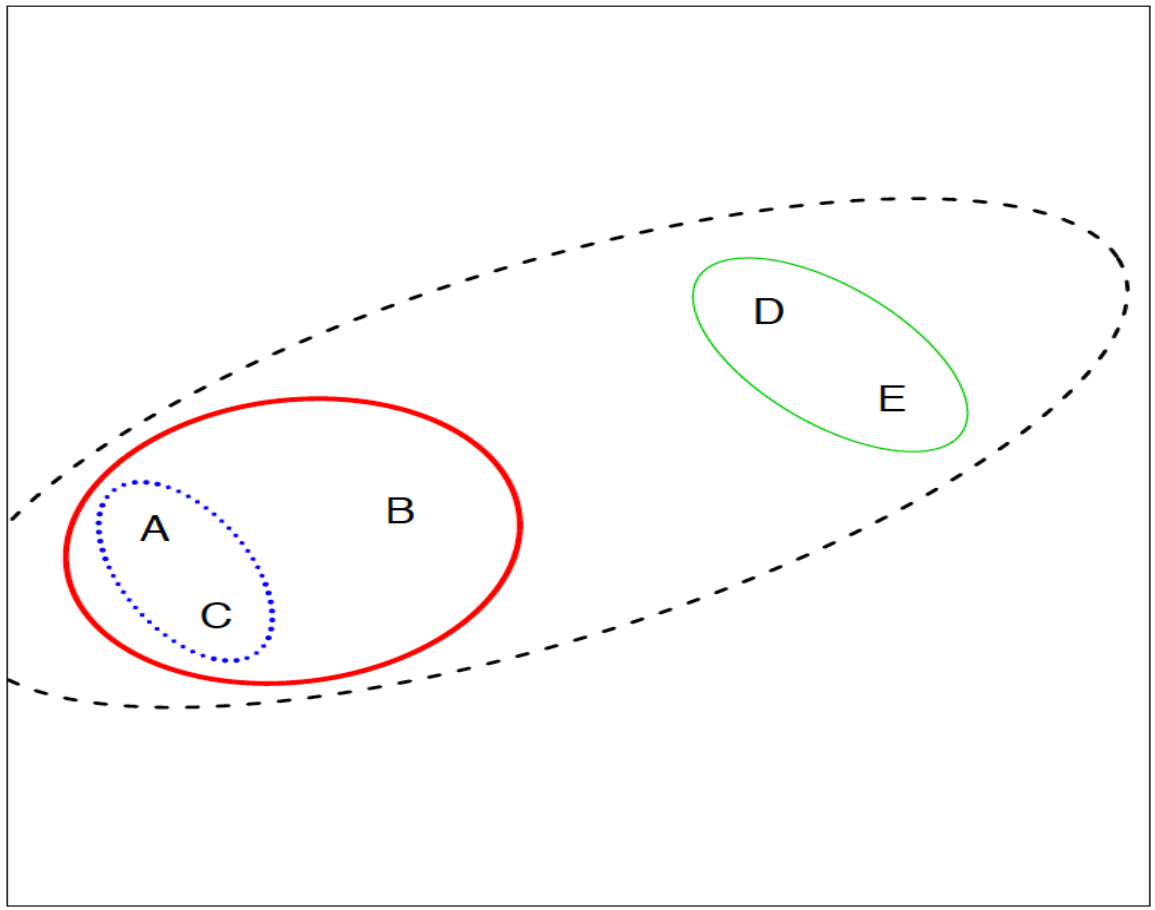
Hierarchical clustering

Builds a hierarchy in a bottom-up fashion...



Hierarchical clustering

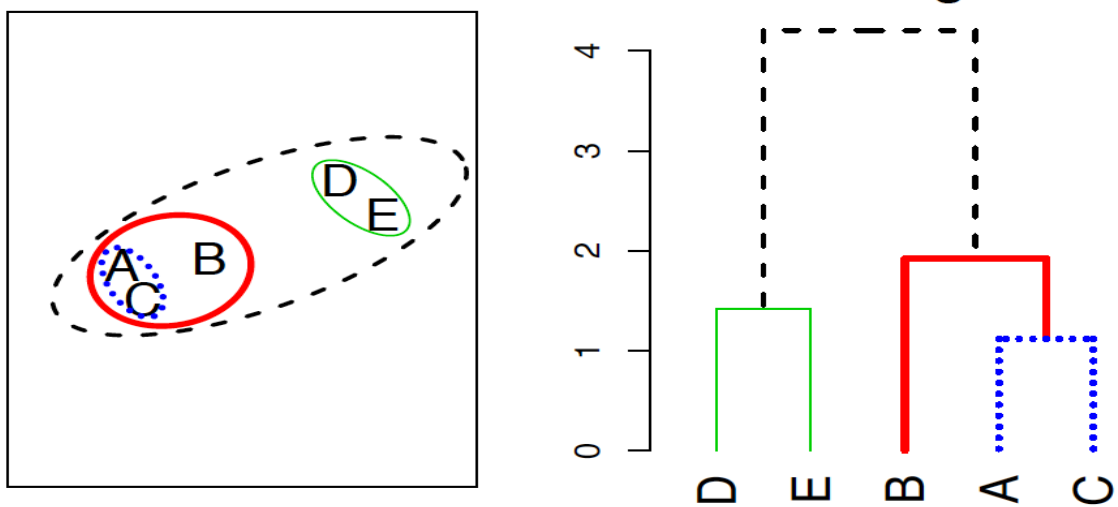
Builds a hierarchy in a bottom-up fashion...



Hierarchical Clustering Algorithm

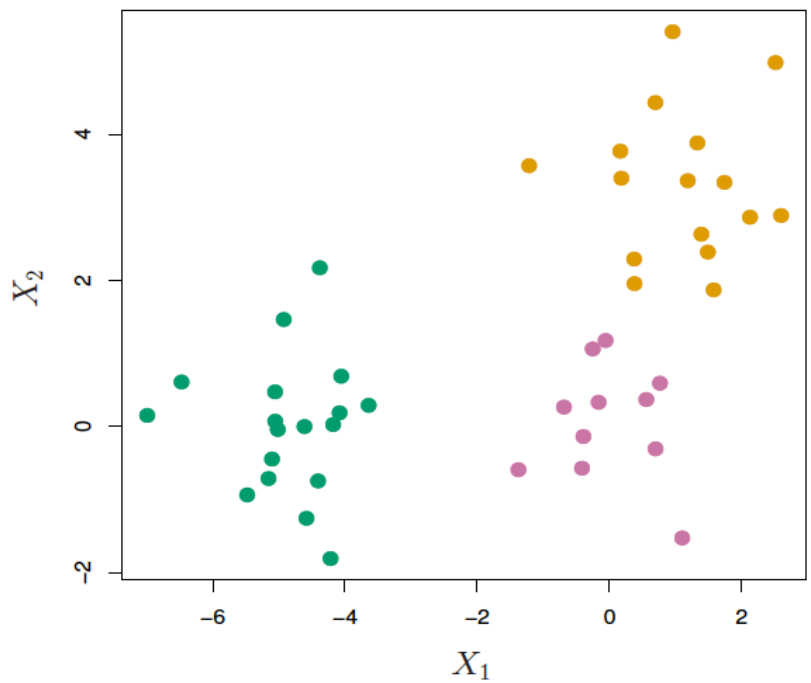
The approach in words:

- 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.

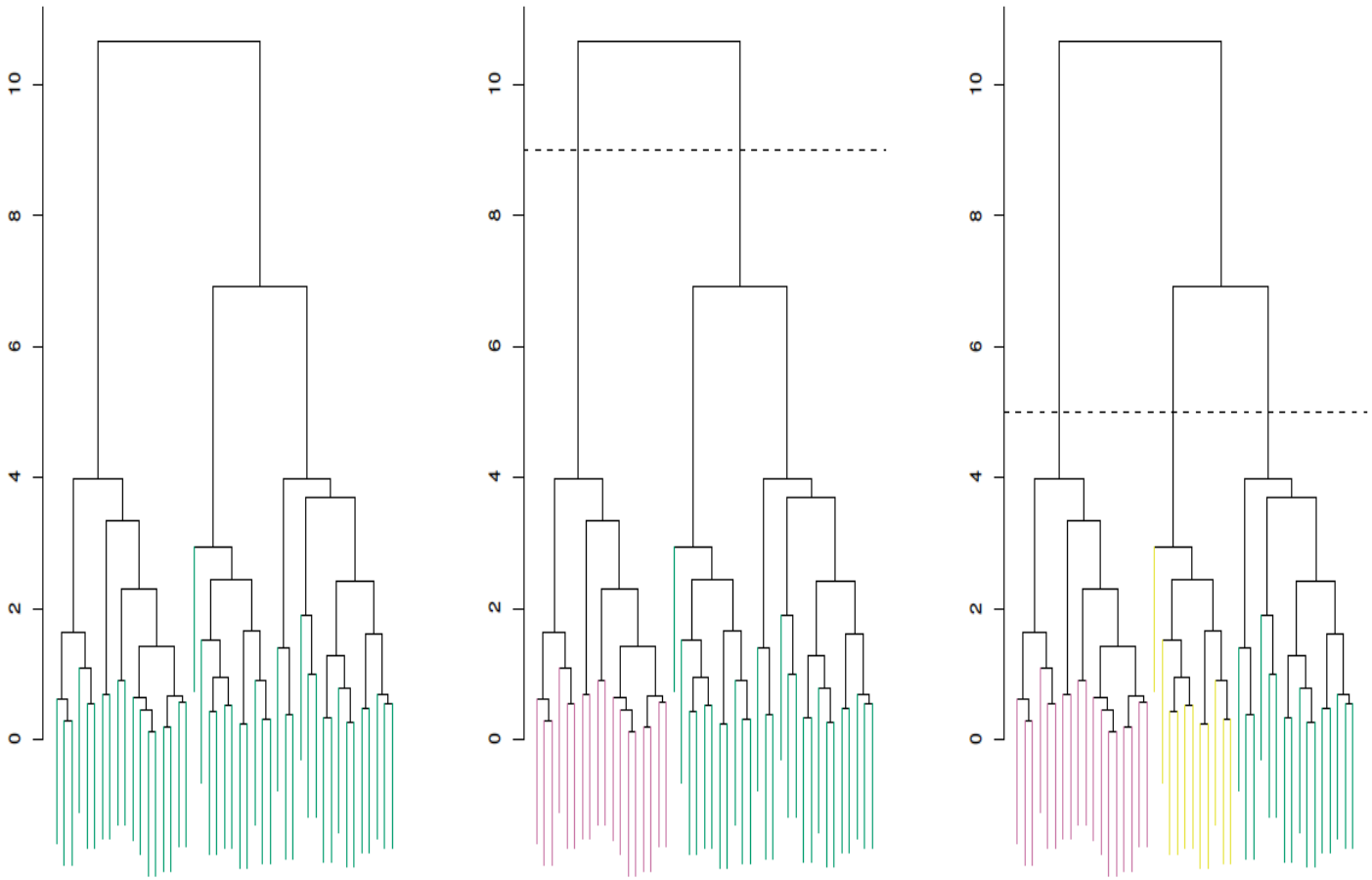


An Example

- 45 observations generated in 2-dimensional space.
- In reality there are three distinct classes, shown in separate colors.
- However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.



An Example



An Example

- Left: Dendrogram obtained from hierarchically clustering the data from previous slide, with complete linkage and Euclidean distance.
- Center: The dendrogram from the left-hand panel, cut at a height of 9 (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.
- Right: The dendrogram from the left-hand panel, now cut at a height of 5. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure

Summary and Remark

- Unsupervised Learning
- Clustering
- Hierarchical clustering
- Read textbook Chapter 14 and `R` code
- Do `R` lab