

# Unsupervised Learning

## *Unsupervised vs Supervised 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  $X_1, X_2, \dots, X_p$  for each object, as well as a response or outcome variable  $Y$ . The goal is then to predict  $Y$  using  $X_1, X_2, \dots, X_p$ .
- Here we instead focus on *unsupervised learning*, where we observe only the features  $X_1, X_2, \dots, X_p$ . We are not interested in prediction, because we do not have an associated response variable  $Y$ .

# The Goals of Unsupervised Learning

- The goal 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 two methods:
  - *principal components analysis*, a tool used for data visualization or data pre-processing before supervised techniques are applied, and
  - *clustering*, a broad class of methods for discovering unknown subgroups in data.

# Principal Components Analysis

- PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

# 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.

## PCA vs Clustering

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
- Clustering looks for homogeneous subgroups among the observations.

## Two clustering methods

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

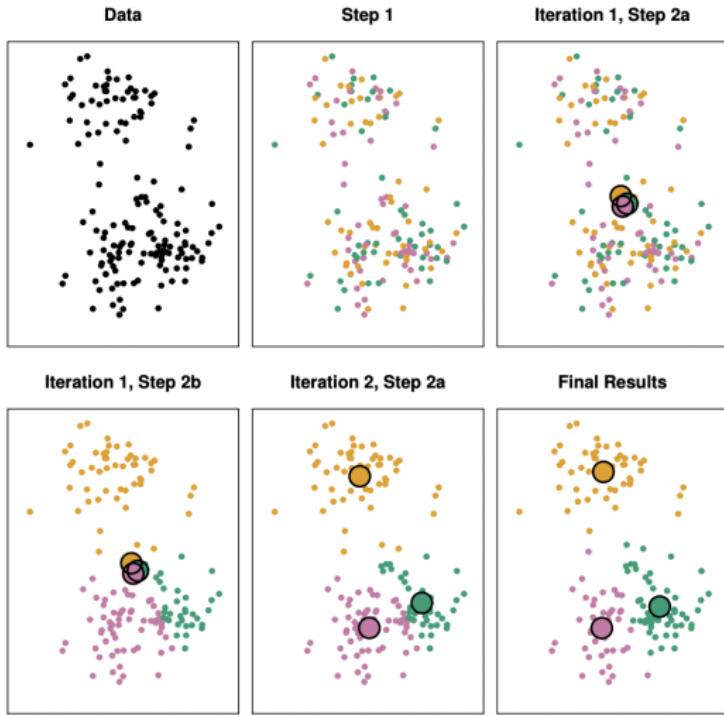
## Details of $K$ -means clustering: continued

- 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  $\text{WCV}(C_k)$  of the amount by which the observations within a cluster differ from each other.
- Hence we want to solve the problem

$$\underset{C_1, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K \text{WCV}(C_k) \right\}. \quad (2)$$

- In words, this formula says that 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.

# Example



## Details of Previous Figure

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

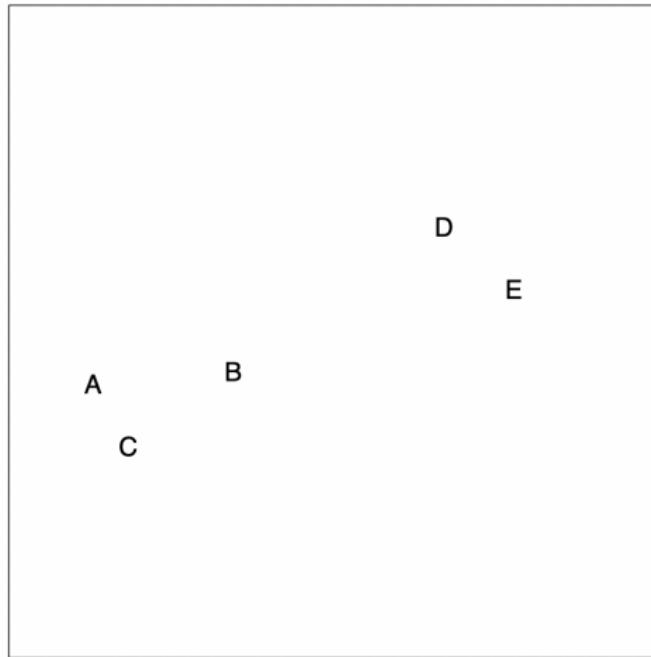
- *Top left:* The observations are shown.
- *Top center:* In Step 1 of the algorithm, each observation is randomly assigned to a cluster.
- *Top right:* In Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.
- *Bottom left:* In Step 2(b), each observation is assigned to the nearest centroid.
- *Bottom center:* Step 2(a) is once again performed, leading to new cluster centroids.
- *Bottom right:* The results obtained after 10 iterations.

## Hierarchical Clustering

- $K$ -means clustering requires us to pre-specify the number of clusters  $K$ . This can be a disadvantage (later we discuss strategies for choosing  $K$ )
- *Hierarchical clustering* is an alternative approach which does not require that we commit to a particular choice of  $K$ .
- In this section, 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.

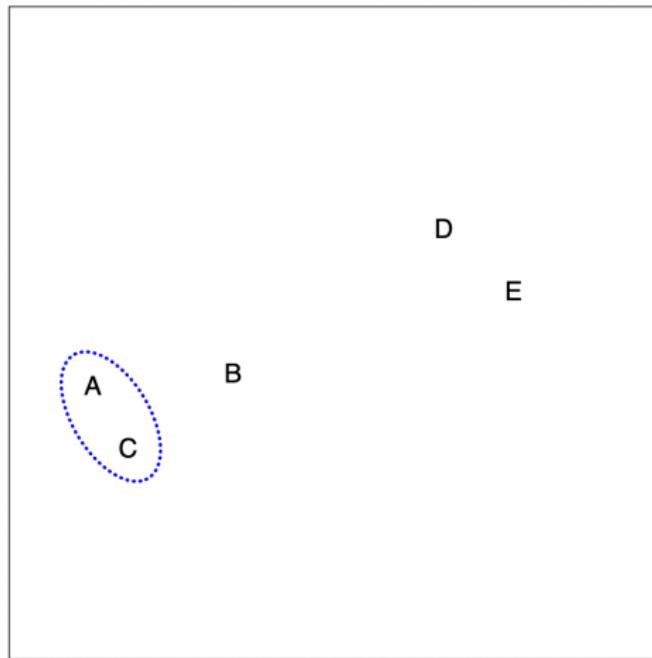
## Hierarchical Clustering: the idea

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



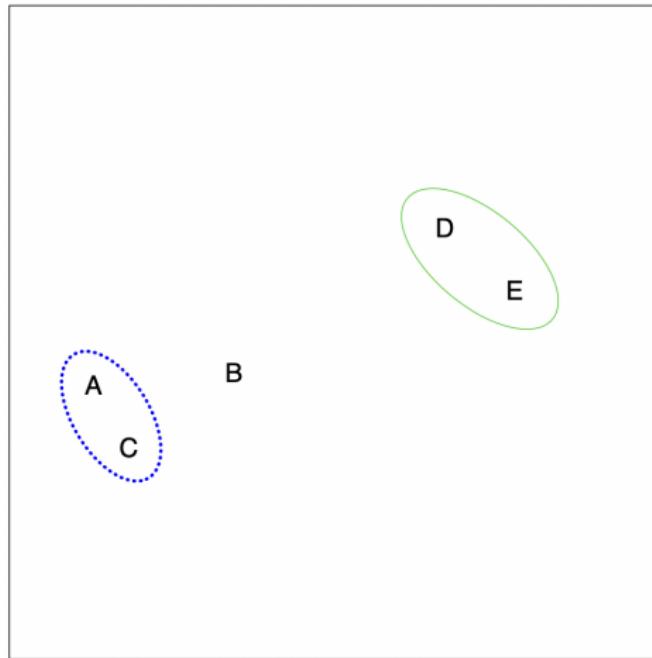
## Hierarchical Clustering: the idea

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



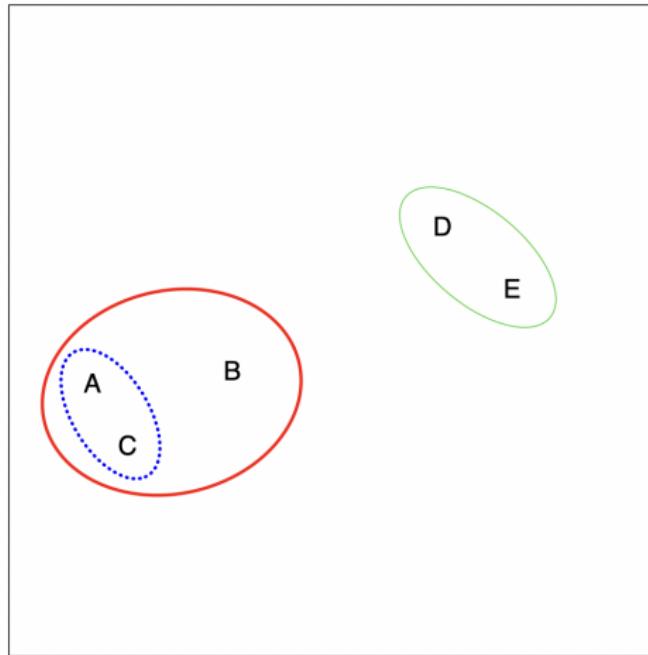
## Hierarchical Clustering: the idea

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



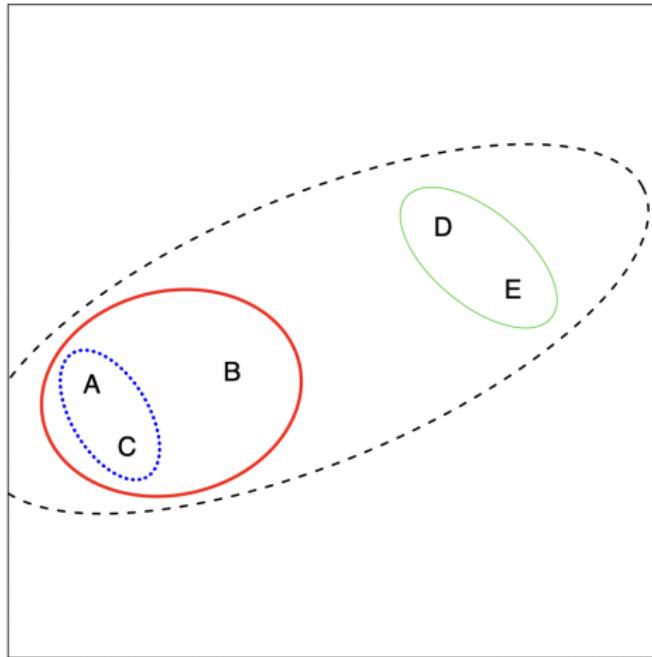
## Hierarchical Clustering: the idea

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



## Hierarchical Clustering: the idea

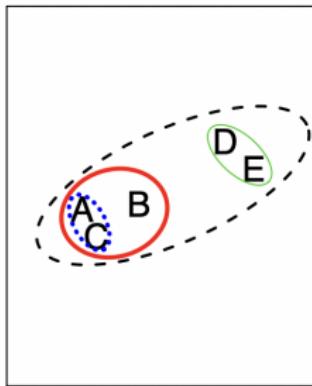
Builds a hierarchy in a “bottom-up” fashion...



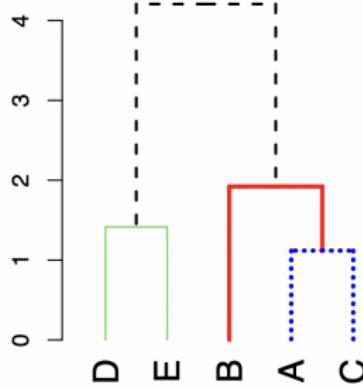
# Hierarchical Clustering Algorithm

The approach in words:

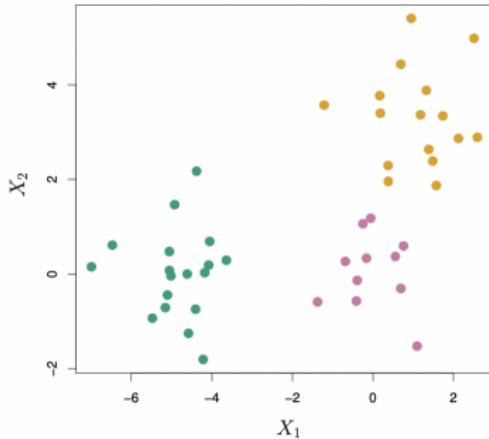
- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.



Dendrogram



## 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.

# Application of hierarchical clustering

