# IG.3510-Machine Learning

## Lecture 4: Unsupervised Learning: Clustering

Dr. Patricia CONDE-CESPEDES

patricia.conde-cespedes@isep.fr

October 23rd, 2023

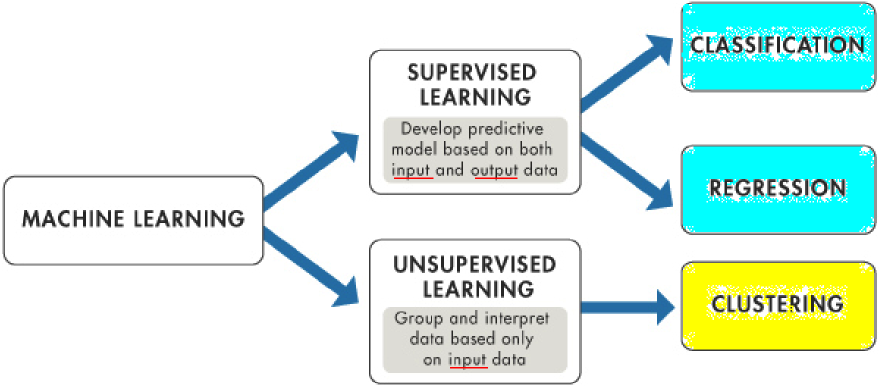
# Plan

1. Introduction to clustering methods
2. What is clustering?
3. $K$-means clustering
4. Hierarchical clustering
5. Practical issues in clustering
6. References

# Outline

1. Introduction to clustering methods
2. What is clustering?
3. *K*-means clustering
4. Hierarchical clustering
5. Practical issues in clustering
6. References

# Reminder: **Supervised** vs. **Unsupervised**


MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

source: `https://fr.mathworks.com/help/stats/machine-learning-in-matlab.html`

# Introduction to Clustering

- **Clustering** refers to a very broad set of techniques for finding subgroups called **clusters**, in a data set.

# Introduction to Clustering

- **Clustering** refers to a very broad set of techniques for finding subgroups called **clusters**, in a data set.
- These groups are made so that the observations within each group are quite similar to each other, while observations between groups are quite different from each other.

# Introduction to Clustering

- **Clustering** refers to a very broad set of techniques for finding subgroups called **clusters**, in a data set.
- These groups are made so that the observations within each group are quite similar to each other, while observations between groups are quite different from each other.

**But, what does it mean for two or more observations to be similar or different?**

# Introduction to Clustering

- **Clustering** refers to a very broad set of techniques for finding subgroups called **clusters**, in a data set.
- These groups are made so that the observations within each group are quite similar to each other, while observations between groups are quite different from each other.

**But, what does it mean for two or more observations to be similar or different?**

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

# Some Clustering applications

Clustering can arise in many applications:

- In **medicine**, to identify different unknown subtypes of breast cancer.

# Some Clustering applications

Clustering can arise in many applications:

- In **medicine**, to identify different unknown subtypes of breast cancer.
- In **marketing**, to perform market segmentation by identifying subgroups of people who might probably purchase a particular product.

# Some Clustering applications

Clustering can arise in many applications:

- In **medicine**, to identify different unknown subtypes of breast cancer.
- In **marketing**, to perform market segmentation by identifying subgroups of people who might probably purchase a particular product.
- In **sociology**, to detect communities in social networks, that is, groups of users that share common interests, hobbies, past times, etc.
- Many more in biology, bibliometry, etc...

# Some Clustering applications

Clustering can arise in many applications:

- In **medicine**, to identify different unknown subtypes of breast cancer.
- In **marketing**, to perform market segmentation by identifying subgroups of people who might probably purchase a particular product.
- In **sociology**, to detect communities in social networks, that is, groups of users that share common interests, hobbies, past times, etc.
- Many more in biology, bibliometry, etc...

These are unsupervised problems because we are trying to **discover structure**- in this case, distinct clusters- in the data set.

# Clustering versus . . .

## Clustering versus Supervised learning

- The goal in supervised problems is to try to predict some outcome variable such as survival time or response to drug treatment, salary, sales, income, etc.

# Clustering versus . . .

## Clustering versus Supervised learning

- The goal in supervised problems is to try to predict some outcome variable such as survival time or response to drug treatment, salary, sales, income, etc.

## Clustering versus PCA:

- *PCA* is usually considered an unsupervised learning method.
- *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.

# K-means clustering and hierarchical clustering

In this lecture we focus on two best-known approaches in clustering:

- In **K-means clustering** : partition the observations into a pre-specified number $K$ of clusters.

# K-means clustering and hierarchical clustering

In this lecture we focus on two best-known approaches in clustering:

- In **K-means clustering** : partition the observations into a pre-specified number $K$ of clusters.
- In **hierarchical clustering (HC)** ,
  - the number of clusters is not previously defined.

## K-means clustering and hierarchical clustering

In this lecture we focus on two best-known approaches in clustering:

- In **K-means clustering** : partition the observations into a pre-specified number $K$ of clusters.
- In **hierarchical clustering (HC)** ,
  - the number of clusters is not previously defined.
  - **dendrogram**: a tree-like visual representation of the observations. It allows to view at once the clusterings obtained for each possible number of clusters, from 1 to $n$ (the number of observations).

# K-means clustering and hierarchical clustering

In this lecture we focus on two best-known approaches in clustering:

- In **K-means clustering** : partition the observations into a pre-specified number $K$ of clusters.
- In **hierarchical clustering (HC)** ,
  - the number of clusters is not previously defined.
  - **dendrogram**: a tree-like visual representation of the observations. It allows to view at once the clusterings obtained for each possible number of clusters, from 1 to $n$ (the number of observations).

In general, clustering is performed on the observations on the basis of the features. However, It is also possible to cluster features on the basis of the observations. In this lecture we will focus on the first case.

# Outline

1. Introduction to clustering methods
2. What is clustering?
3. *K*-means clustering
4. Hierarchical clustering
5. Practical issues in clustering
6. References

# Definition of a partition or clustering

A **partition**, also called **clustering**, on a set of $n$ objects (observations) is a set of nonempty subsets $C_1, \ldots, C_K$ of the objects, called **clusters** that satisfy two properties:

- $C_1 \cup C2 \cup ... \cup C_K = \{1, ..., n\}$. That is, each observation belongs to at least one of the $K$ clusters.

# Definition of a partition or clustering

A **partition**, also called **clustering**, on a set of $n$ objects (observations) is a set of nonempty subsets $C_1, \ldots, C_K$ of the objects, called **clusters** that satisfy two properties:

- $C_1 \cup C2 \cup ... \cup C_K = \{1, ..., n\}$. That is, each observation belongs to at least one of the $K$ clusters.
- $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. The clusters are non-overlapping: no observation belongs to more than one cluster.

Each observation belongs to exactly one cluster.

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

# Stirling number of the second kind

**How many partitions on a set of $n$ objects into $K$ clusters exist?**

# Stirling number of the second kind

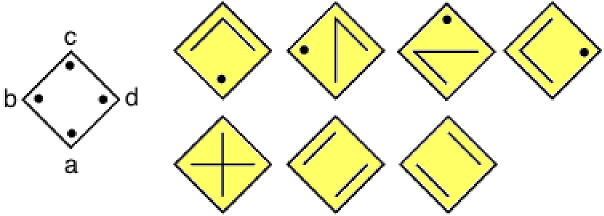
**How many partitions on a set of $n$ objects into $K$ clusters exist?**

The number of ways to partition a set of $n$ objects into $K$ clusters is given by the **Stirling number of the second kind** $S(n, K)$:

$$S(n, K) = \frac{1}{K!} \sum_{j=1}^{K} (-1)^{K-j} \binom{K}{j} j^n$$

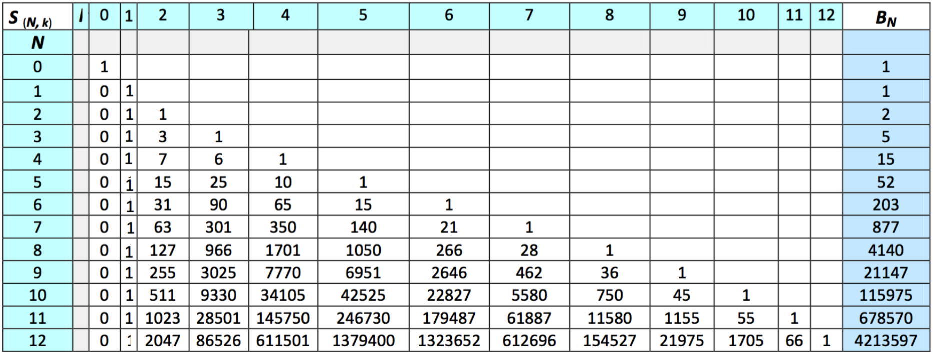
For example, $S(4, 2) = 7$, there are 7 ways to partition 4 objects, $\{a, b, c, d\}$, into 2 groups, namely:

1) $\{(a), (bcd)\}$
2) $\{(b)(acd)\}$
3)$\{(c), (abd)\}$
4)$\{(d), (abc)\}$,
5) $\{(ac), (bd)\}$
6) $\{(ad)(bc)\}$
7)$\{(ab), (cd)\}$.



# Stirling Number of 2nd kind up to $N = 12$ objects

**Bell number** $B_N$: number of possible partitions on a set of $N$ objects.



| $S_{(N,k)}$ | / | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | $B_N$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | | | | | | | | | | | | | | | |
| 0 | | 1 | | | | | | | | | | | | | 1 |
| 1 | | 0 | 1 | | | | | | | | | | | | 1 |
| 2 | | 0 | 1 | 1 | | | | | | | | | | | 2 |
| 3 | | 0 | 1 | 3 | 1 | | | | | | | | | | 5 |
| 4 | | 0 | 1 | 7 | 6 | 1 | | | | | | | | | 15 |
| 5 | | 0 | 1 | 15 | 25 | 10 | 1 | | | | | | | | 52 |
| 6 | | 0 | 1 | 31 | 90 | 65 | 15 | 1 | | | | | | | 203 |
| 7 | | 0 | 1 | 63 | 301 | 350 | 140 | 21 | 1 | | | | | | 877 |
| 8 | | 0 | 1 | 127 | 966 | 1701 | 1050 | 266 | 28 | 1 | | | | | 4140 |
| 9 | | 0 | 1 | 255 | 3025 | 7770 | 6951 | 2646 | 462 | 36 | 1 | | | | 21147 |
| 10 | | 0 | 1 | 511 | 9330 | 34105 | 42525 | 22827 | 5580 | 750 | 45 | 1 | | | 115975 |
| 11 | | 0 | 1 | 1023 | 28501 | 145750 | 246730 | 179487 | 61887 | 11580 | 1155 | 55 | 1 | | 678570 |
| 12 | | 0 | 1 | 2047 | 86526 | 611501 | 1379400 | 1323652 | 612696 | 154527 | 21975 | 1705 | 66 | 1 | 4213597 |

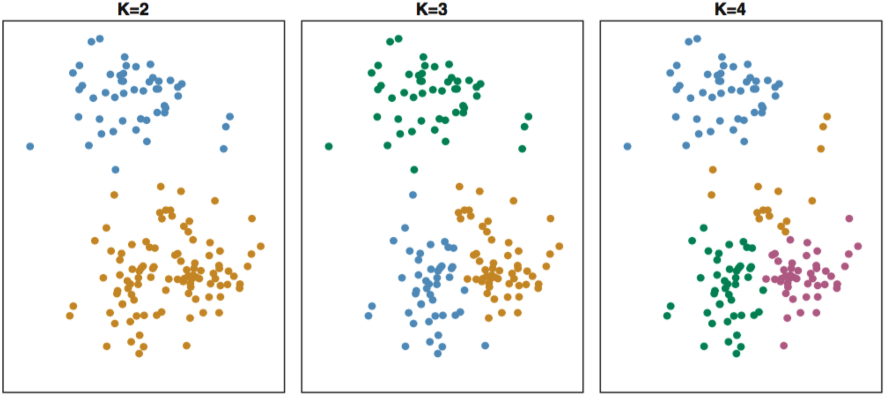
The total is called the **Bell number**.

# Outline

1. Introduction to clustering methods
2. What is clustering?
3. $K$-means clustering
4. Hierarchical clustering
5. Practical issues in clustering
6. References

# *K*-means clustering example

First specify the desired number of clusters $K$; then the $K$-means algorithm will assign each observation to exactly one of the $K$ clusters.



Example of $K$-means clustering for different values of $K$ on simulated data with 150 observations in 2-dimensional space.

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

# How to define which one is the best clustering?

- 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 $C_k$, denoted $WCV(C_k)$, measures the amount by which the observations within a cluster differ from each other.

# How to define which one is the best clustering?

- 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 $C_k$, denoted $WCV(C_k)$, measures the amount by which the observations within a cluster differ from each other.
- *Idea :* choose the partition of the $n$ observations into $K$ clusters such that the total within-cluster variation (for all the clusters), is as small as possible.

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

The clustering that minimizes this expression is the best among all the possibilities given by the $S(n, K)$! (Stirling number of the 2nd kind).

# Definition of within-cluster variation

The most common choice is the *Euclidean distance*:

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

where $|C_k|$ denotes the number of observations in the $k$th cluster and $p$ the number of features.

# Definition of within-cluster variation

The most common choice is the *Euclidean distance*:

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

where $|C_k|$ denotes the number of observations in the $k$th cluster and $p$ the number of features.

This gives the optimization problem that defines $K$-means clustering,

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left( \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right). \qquad (1)$$

This is a very difficult problem to solve exactly! Since the $S(n,K) \sim \frac{K^n}{K!}$. becomes very very Huge rapidly! $S(12,5) = 1379400$

Fortunately, the $K$-means algorithm provides an approximation.

# $K$-Means Clustering Algorithm

$K$-Means Clustering Algorithm

**Step 1:** Randomly assign a number, from 1 to $K$, to each of the observations.

# $K$-Means Clustering Algorithm

$K$-Means Clustering Algorithm

**Step 1:** Randomly assign a number, from 1 to $K$, to each of the observations.

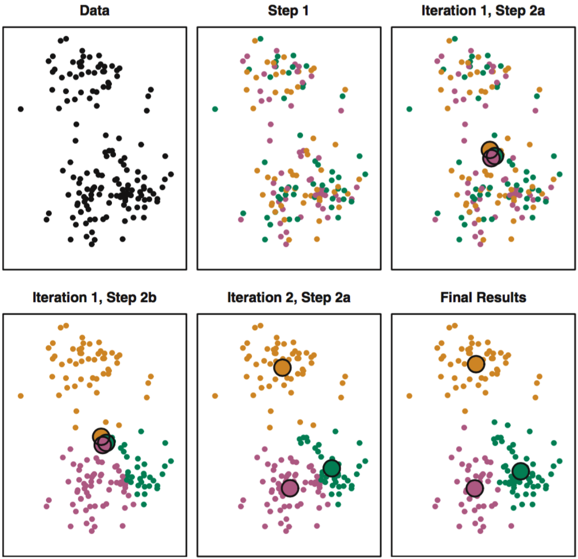
**Step 2:** Iterate until the cluster assignments stop changing:

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

Mathematically, the $k$th cluster **centroid** is $\bar{x}_k = (\bar{x}_{k1}, \bar{x}_{k2}, \dots, \bar{x}_{kp})$. where $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ is the mean for feature $j$ in cluster $C_k$.

This is where the name $K$-means is derived from.

# Example: steps of the $K$-means algorithm, $K = 3$



# Details of Previous example

The steps 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 2a), 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 2b), each observation is assigned to the nearest centroid.
- Bottom center: Step 2a) is once again performed, leading to new cluster centroids.
- Bottom right: The results obtained after 10 iterations.

# Properties of the $K$-means Algorithm

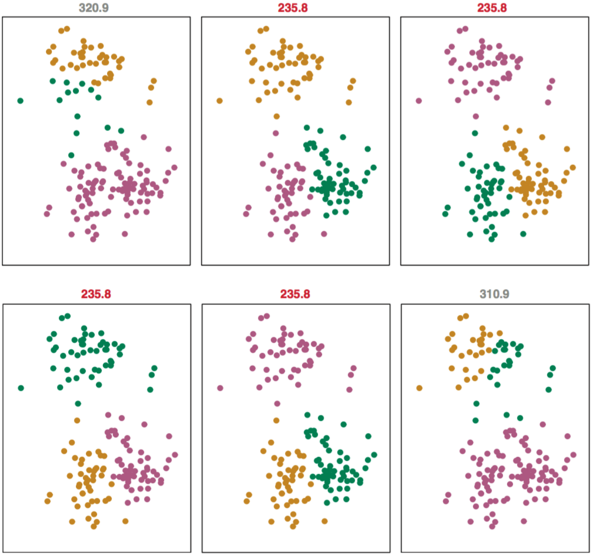
This algorithm is guaranteed to decrease the value of the objective function (1) at each step. Indeed, note that:

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

- In Step 2a) the cluster means for each feature are the constants that minimize the sum-of-squared deviations, and
- In Step 2b) reallocating the observations can only improve the objective function.
- Then, as the algorithm is run, the objective function (1) can only decrease until the result no longer changes;

When the result no longer changes, a **local optimum** has been reached.

# $K$-means with different starting assignements



Results after running K-means 6 times using different initial cluster assignments, $K = 3$.

# Details on the previous example

$K$-means algorithm is a **heuristic**!

It finds a local optimum as it depends on the initial (random) cluster assignment of each observation in Step 1.

# Details on the previous example

*K*-means algorithm is a **heuristic**!

It finds a local optimum as it depends on the initial (random) cluster assignment of each observation in Step 1.
It is important to run the algorithm multiple times.

- For the previous example, above each plot is the value of the objective function.
- Three different local optima were obtained,
- The best clustering is the one that reaches an objective value of 235.8 (labeled in red). It visually provides better separation between the clusters.

# Outline

1. Introduction to clustering methods
2. What is clustering?
3. $K$-means clustering
4. Hierarchical clustering
5. Practical issues in clustering
6. References

# Hierarchical clustering vs. $K$-means

Potential disadvantage of $K$-means: It requires to define the number of clusters $K$ in advance!

# Hierarchical clustering vs. $K$-means

Potential disadvantage of $K$-means: It requires to define the number of clusters $K$ in advance!

- **Hierarchical clustering** is an alternative approach which does not require to pre-define $K$.
- Hierarchical clustering results in a tree-based representation of the observations, called a **dendrogram**.

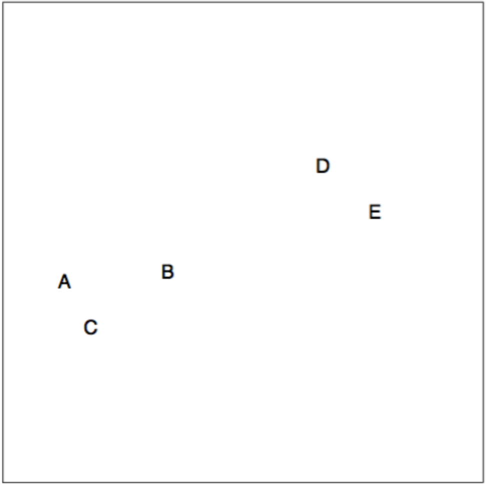
# Hierarchical clustering vs. $K$-means

Potential disadvantage of $K$-means: It requires to define the number of clusters $K$ in advance!

- **Hierarchical clustering** is an alternative approach which does not require to pre-define $K$.
- Hierarchical clustering results in a tree-based representation of the observations, called a **dendrogram**.
- In this lecture, we will study the **bottom-up** or **agglomerative** clustering.
- A dendrogram is built starting from the leaves and combining clusters up to the trunk or root.

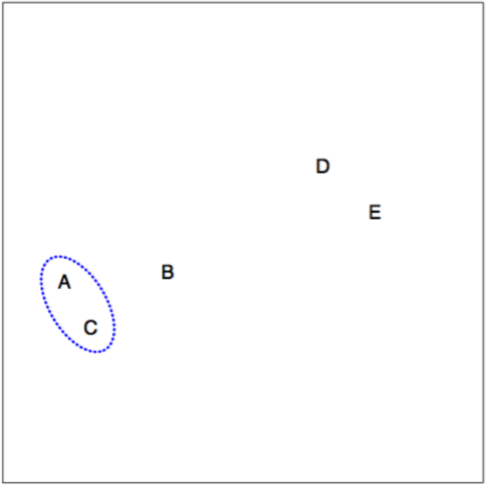
# Hierarchical Clustering: the idea (1/6)

We have 5 objects described by two features (raw data):



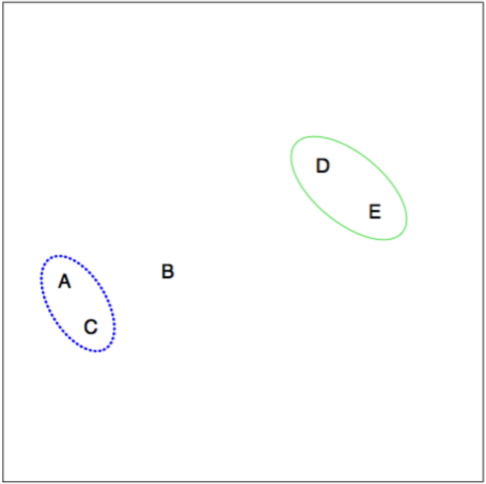
# Hierarchical Clustering: the idea (2/6)

Hierarchical Clustering joins the closest pair of objects (squared distance):



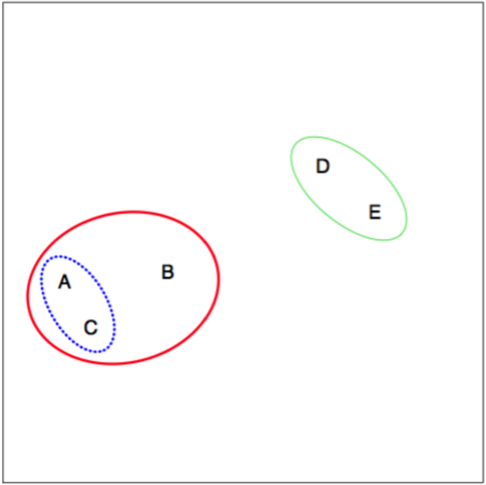
# Hierarchical Clustering: the idea (3/6)

Then, Hierarchical Clustering looks for the next closest pair of objects:



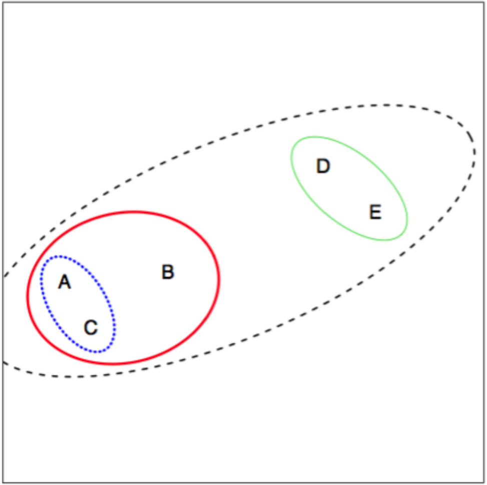
# Hierarchical Clustering: the idea (4/6)

Hierarchical Clustering can join an object and a cluster or even 2 clusters:



# Hierarchical Clustering: the idea (5/6)

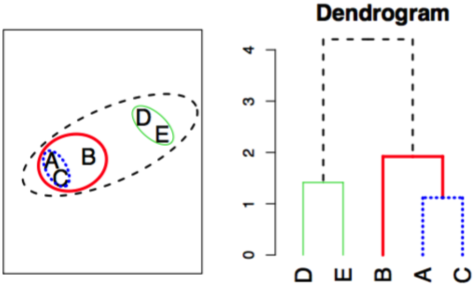
Finally, the 5 objects are put together in one cluster:



# Hierarchical Clustering idea (summary) (6/6)

In summary:

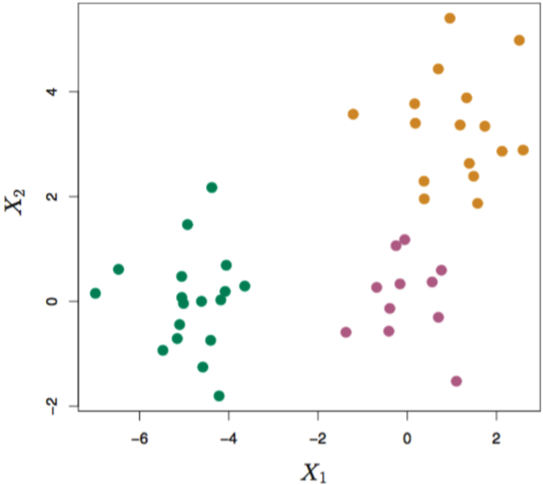
- Start with each point in its own cluster.
- Identify the closest two objects and merge them.
- Repeat.
- End when all points are in a single cluster.



The height of the join in the dendrogram represents the distance between the two involved points or clusters. The total number of joins is $(n-1)$.

# Second example with more observations

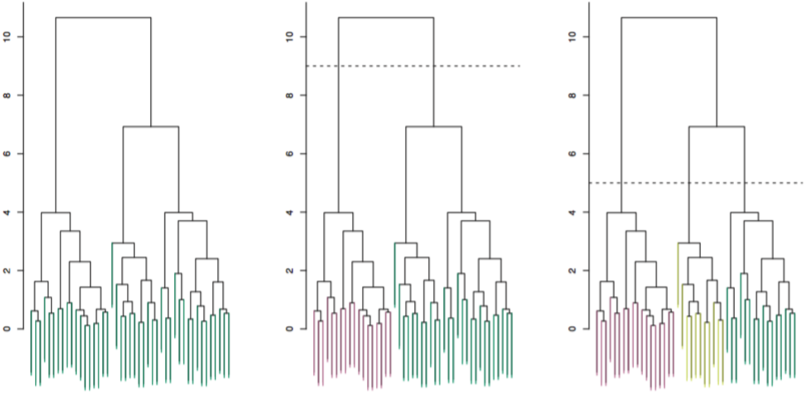
45 observations generated in 2-dimensional space.



Hierarchical clustering was performed to detect the grouping structure.

# How to identify the clusters in a dendrogram?

Make a horizontal cut across the dendrogram. The distinct sets of observations beneath the cut represent the clusters.



From a single dendrogram, one can obtain partitions of any number of clusters from 1 (top) to $n$ (bottom) because there were $(n-1)$ joins. The height of the cut controls the number of clusters obtained.

# More interpretations about a dendrogram

- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other. In contrast, observations that fuse later (near the top of the tree) can be quite different.

# More interpretations about a dendrogram

- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other. In contrast, observations that fuse later (near the top of the tree) can be quite different.
- Given any two observations, the **height** of their fusion, measured on the vertical axis, indicates how different they are.

# More interpretations about a dendrogram

- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other. In contrast, observations that fuse later (near the top of the tree) can be quite different.
- Given any two observations, the **height** of their fusion, measured on the vertical axis, indicates how different they are.
- The term **hierarchical** refers to the fact that clusters obtained by cutting the dendrogram at a given height are necessarily nested within the clusters obtained by cutting the dendrogram at any greater height.
  - However, on an arbitrary data set, this assumption of hierarchical structure might be unrealistic. Drawback!.

# The Hierarchical Clustering Algorithm

*Hierarchical Clustering Algorithm*

**Step 1:** Begin with $n$ observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = \frac{n(n-1)}{2}$ pairwise dissimilarities. Treat each observation as one cluster.

# The Hierarchical Clustering Algorithm

*Hierarchical Clustering Algorithm*

**Step 1:** Begin with $n$ observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = \frac{n(n-1)}{2}$ pairwise dissimilarities. Treat each observation as one cluster.

**Step 2:** For $i = n, n-1, \ldots, 2$:

**a)** Examine all pairwise inter-cluster dissimilarities among the $i$ clusters and identify the pair of clusters that are least dissimilar. Join these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

# The Hierarchical Clustering Algorithm

*Hierarchical Clustering Algorithm*

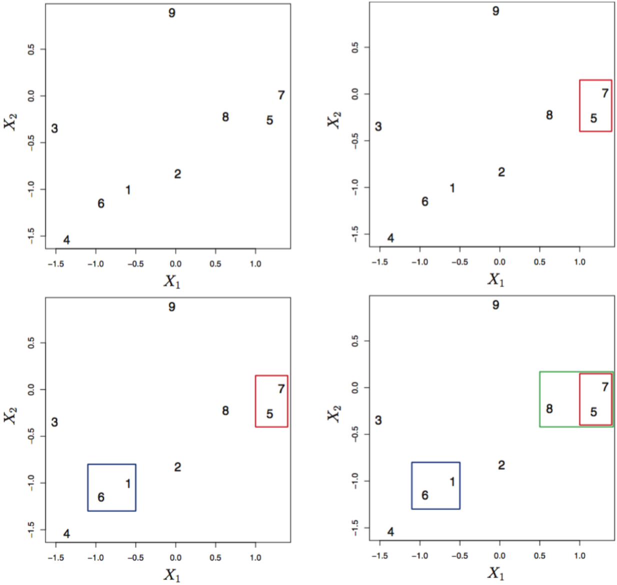
**Step 1:** Begin with $n$ observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = \frac{n(n-1)}{2}$ pairwise dissimilarities. Treat each observation as one cluster.

**Step 2:** For $i = n, n-1, \ldots, 2$:

**a)** Examine all pairwise inter-cluster dissimilarities among the $i$ clusters and identify the pair of clusters that are least dissimilar. Join these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

**b)** Compute the new pairwise inter-cluster dissimilarities among the $(i-1)$ remaining clusters.

# The first few steps of the algorithm, example



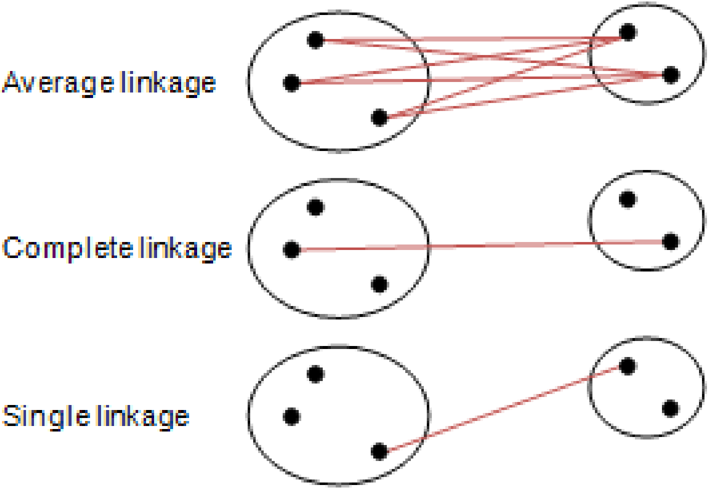
# How to define the dissimilarity between 2 clusters?

The **linkage** defines the dissimilarity between two groups of observations.

| **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. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities. |
| Average | Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities. |
| Centroid | Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. |

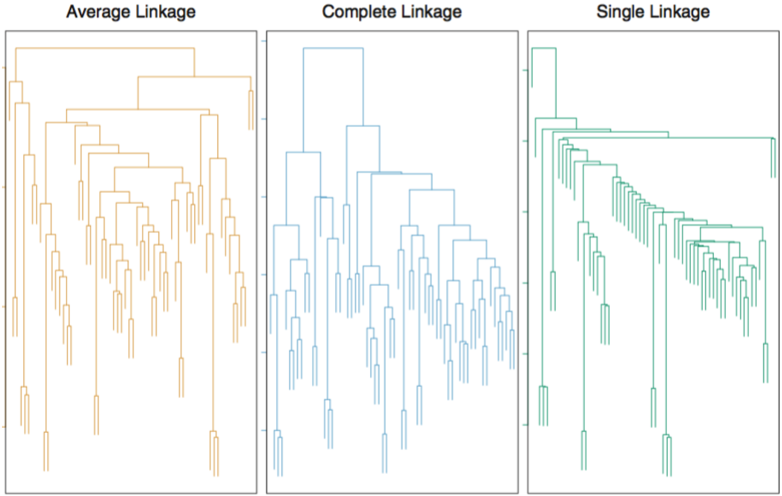
Average and complete linkage tend to yield more balanced dendrograms.

# Types of linkage, illustration



# Comparison of linkage measures

The resulting dendrogram depends strongly on the type of linkage used



# Outline

1. Introduction to clustering methods
2. What is clustering?
3. $K$-means clustering
4. Hierarchical clustering
5. Practical issues in clustering
6. References

# The optimal number of clusters

No unanimous answer. Two possibilities:

- Domain knowledge
- Data driven approach.

Elbow method: The WCV is a measure of compactness of the cluster.

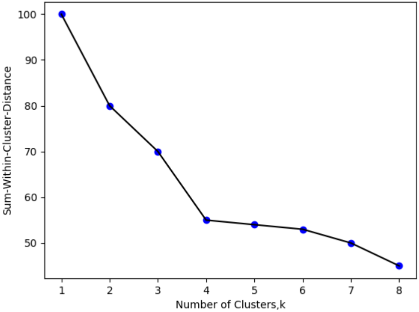
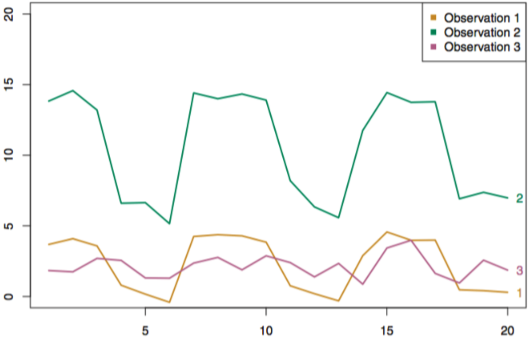


Figure: Source: `https://towardsdatascience.com/clustering-evaluation-strategies-98a4006fcfc`

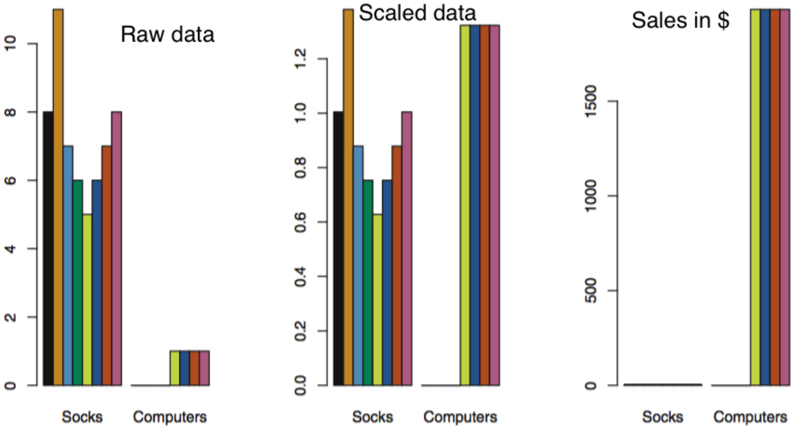
# Choice of Dissimilarity Measure

- An alternative to Euclidean distance is **correlation-based distance** which considers two observations to be similar if their features are highly correlated.
- In this case, correlation is computed between the observation profiles for each pair of observations. Correlaton is usually used for recommender systems.



# Should the variables be scaled?

Should the observations or features first be standardized (mean zero and standard deviation one)?
Some items may be purchased more frequently than others; for instance, a shopper might buy ten pairs of socks a year, but a computer very rarely.



# Clustering performance evaluation

- **the Silhouette coefficient:** for each observation calculate :

$$s = \frac{b-a}{\max(a,b)}$$

  where $a$ is the mean distance between the observation and all other points in the same class and $b$ is the mean distance between the observation and all other points in the next nearest cluster.
  $s$ lies from $-1$ to $+1$; $-1$ for incorrect clustering and $+1$ for highly dense clustering. $s = 0$ for overlapping clusters.

- **Davies-Bouldin Index (DB):** average similarity between each cluster $i$ and its most similar one $j$.

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

  where $s_i$ is the average distance between each point of cluster $i$ and its centroid and $d_{ij}$ is the distance between cluster centroids $i$ and $j$. Then, the DB index is:

$$DB = \frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j} R_{ij}.$$

  Zero is the lowest possible score. Values closer to zero indicate a better partition.

# Outline

1 Introduction to clustering methods

2 What is clustering?

3 *K*-means clustering

4 Hierarchical clustering

5 Practical issues in clustering

6 References

# References

- James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. ”An Introduction to Statistical Learning with Applications in R”, 2nd edition, New York : ”Springer texts in statistics”, 2021. Site web: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf.
- Hastie, Trevor; Tibshirani, Robert and Friedman, Jerome (2009). ”The Elements of Statistical Learning (Data Mining, Inference, and Prediction), 2nd edition”. New York: ”Springer texts in statistics”. Site web : http://statweb.stanford.edu/~tibs/ElemStatLearn/
- Towards data science (2019): ”Clustering Evaluation strategies”. Visited on October 20th. Site web : https://towardsdatascience.com/clustering-evaluation-strategies-98a4006fcfc