Unsupervised Learning Methods

Faculty of Information Technology, Monash University, Australia

FIT5149 week 12

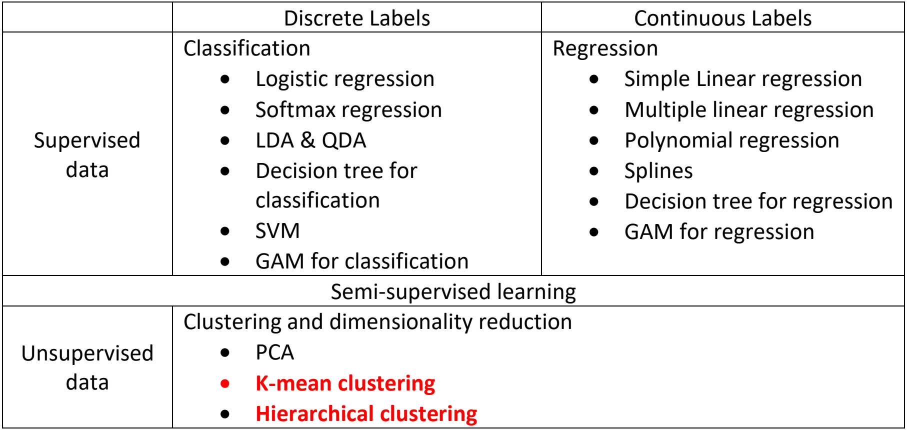
Outline

1. K-Means Clustering

2. Hierarchical Clustering

3. Summary

Where does clustering methods sit in the Unit?



	Discrete Labels	Continuous Labels
Supervised data	Classification • Logistic regression • Softmax regression • LDA & QDA • Decision tree for classification • SVM • GAM for classification	Regression • Simple Linear regression • Multiple linear regression • Polynomial regression • Splines • Decision tree for regression • GAM for regression
Semi-supervised learning		
Unsupervised data	Clustering and dimensionality reduction • PCA • **K-mean clustering** • **Hierarchical clustering**	

Figure: Major topics covered in FIT5149

Learning Outcomes

- Weekly learning outcomes
  - Implement various clustering methods
  - Differentiate between PCA and clustering methods.
  - Differentiate between supervised learning and unsupervised learning
  - understand the basic idea of topic modelling
- Unit learning outcomes
  - Evaluate the limitations, appropriateness and benefits of data analytics methods for given tasks;
  - Design solutions to real world problems with data analytics techniques;
  - Assess the results of an analysis;

Unsupervised Learning vs Supervised learning

- Supervised learning
  - What is observed:
    - A set of features (or predictors): $X_1, X_2, \ldots, X_p$ for each observed object
    - The corresponding response variable (e.g., class labels): Y
  - The goal: learn a model that can predict Y using $X_1, X_2, \ldots, X_p$.
- Unsupervised learning
  - What is observed:
    - Only a set of features $X_1, X_2, \ldots, X_p$ for each observed object
  - Discover interesting things about the measurements:
    - Visualize data
    - Cluster the variables or the observations into subgroups that share some pattens.

Unsupervised Learning: the Challenge

- Unsupervised learning is more subjective than supervised learning.
  - No simple goal for analysis
  - The computer have to learn how to do something that we don't tell it how to do.
- Some issues, for example
  - The number of subgroups (clusters)
  - The different results via K-means with different random initialisations
  - How to assess the performance of the unsupervised learning methods?
- The learning (or inference) procedure is hard!

Unsupervised Learning: advantage

- Unsupervised learning methods are of growing importance in a number of fields:
  - subgroups of breast cancer patients grouped by their gene expression measurements,
  - groups of shoppers characterised by their browsing and purchase histories,
  - movies grouped by the ratings assigned by movie viewers.
  - topic modelling
- Easier to obtain unlabelled data than labelled data
- Methods to be discussed:
  - Clustering methods: discovering unknown subgroups in data.
    - K-Means clustering
    - Hierarchical clustering

Multiple-choice question

Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to?

A. Given a large set of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments.

B. Given a patient's blood test and x-ray results, predict how likely the patient has breast cancer.

C. Having a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

D. Given data on how 1000 medical patients respond to an experimental drug (such as effectiveness of the treatment, side effects, etc.), discover whether there are different categories of patients in terms of how they respond to the drug, and if so what these categories are.

Outline

1. K-Means Clustering
2. Hierarchical Clustering
3. Summary

Clustering

- A broad set of techniques for finding subgroups or clusterings in a data set
  - Partition data into distinct groups so that the observations within each group are quite similar to each other,
  - Define similarity/dissimilarity measure
  - Take into account domain-specific knowledge of the data be studied
- Clustering vs PCA
  - 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.
- Applications of clustering algorithms
  - Cluster text: groups documents with similar topics together
  - Monitor the progress of students' academic performance: group students into different clusters based their scores, each cluster denotes the different level of performance.
  - Perform market segmentation: identify subgroups of people who might be more receptive to a particular form of advertising, or more likely to purchase a particular product.

K-Means Clustering

- K-Means Clustering: partition a data set into K distinct, **non-overlapping** clusterings
- Two general steps:
  - Specify the desired number of clusters K
  - Assign each observation to exactly one of the K clusters

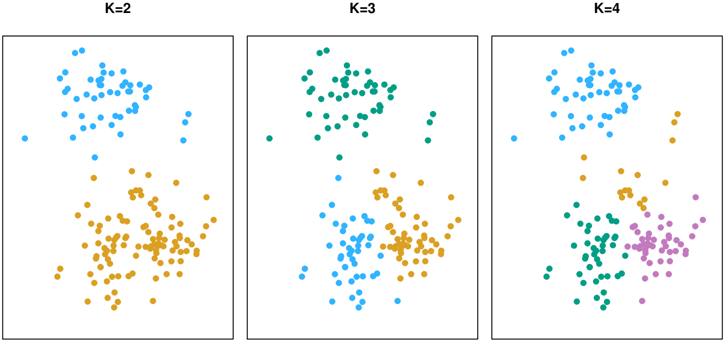


Figure: A simulated data set with 150 observations in 2-dimensional space.

K-Means Clustering

- K-Means Clustering: partition a data set into K distinct, **non-overlapping** clusterings
- Two import properties:

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

  1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.
  2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

  For instance, if the ith observation is in the kth cluster, then $i \in C_k$.

K-Means Clustering: Details

- The basic idea: A good clustering is one for which the within-cluster variation is as small as possible.
- The within-cluster variation for cluster C_k: a measure $W(C_k)$ of the amount by which the observations within a cluster differ from each other.
- The problem to be solved (the target function):

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

- Typically we use squared Euclidean distance

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

- The optimisation problem is now

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

K-Means Clustering: the algorithms

Algorithm 10.1 *K-Means Clustering*

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:

   (a) 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.

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

K-Means Clustering: Visualisation [1]

- Running example of the clustering algorithm

[1] The visualisation is from `https://github.com/karanveerm/kmeans`

K-Means Clustering: Visualisation [1]

- The problem of local optimum

[1] The visualisation is from `https://github.com/karanveerm/kmeans`

K-Means Clustering: Visualisation [1]

- Download the Javascript from https://github.com/karanveerm/kmeans, and play it with yourself.

[1] The visualisation is from https://github.com/karanveerm/kmeans

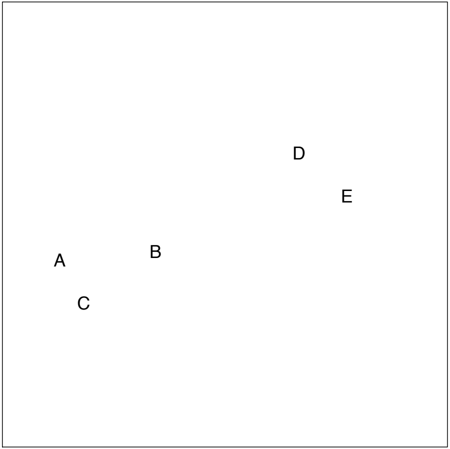
Outline

1. K-Means Clustering
2. Hierarchical Clustering
3. Summary

Hierarchical Clustering: basic idea

Hierarchical Clustering: the idea

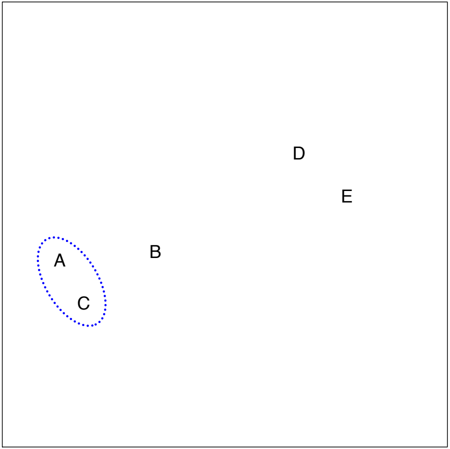
Builds a hierarchy in a “bottom-up” fashion...



Hierarchical Clustering: basic idea

Hierarchical Clustering: the idea

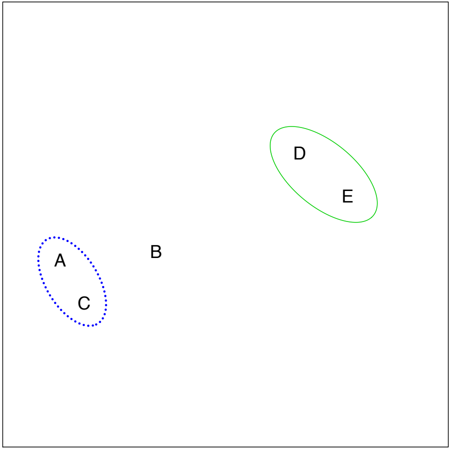
Builds a hierarchy in a “bottom-up” fashion...



Hierarchical Clustering: basic idea

Hierarchical Clustering: the idea

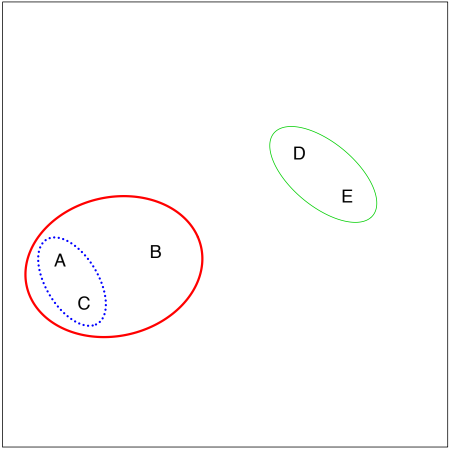
Builds a hierarchy in a “bottom-up” fashion...



Hierarchical Clustering: basic idea

Hierarchical Clustering: the idea

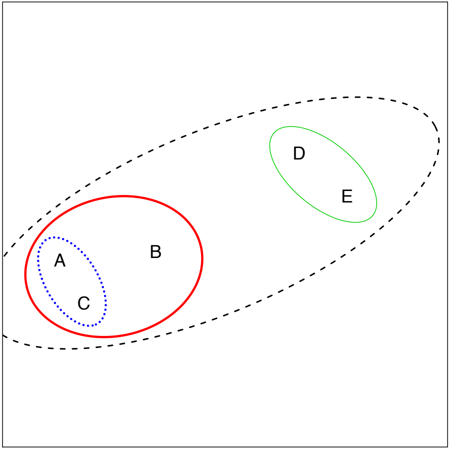
Builds a hierarchy in a “bottom-up” fashion...



Hierarchical Clustering: basic idea

Hierarchical Clustering: the idea

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

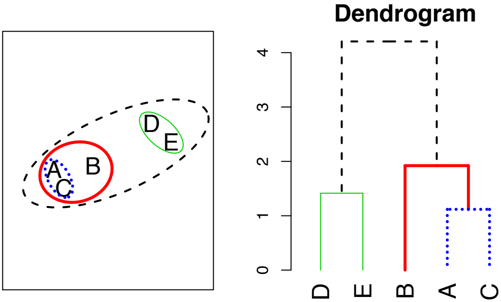


Hierarchical Clustering: basic idea

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.



Hierarchical Clustering: Algorithm

Algorithm 10.2 *Hierarchical Clustering*

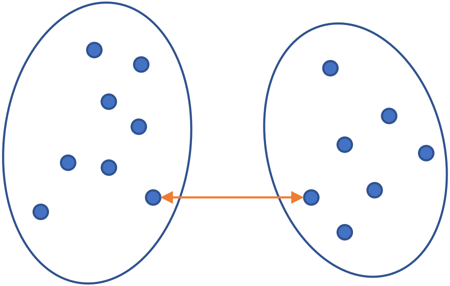
1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

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 (that is, most similar). Fuse 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.

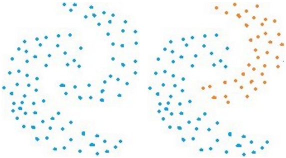
Linkage function: Single linkage



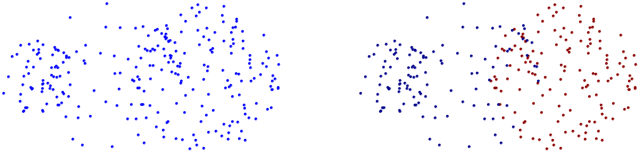
$$L(r, s) = \min(D(x_{r,i}, x_{s,i}))$$

Linkage function: Single linkage[2]

- pros: It can separate non-elliptical shapes as long as the gap between two clusters is not small.

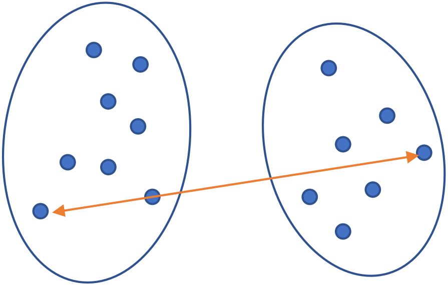


- cons: It cannot separate clusters properly if there is noise between clusters.



[2] See https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

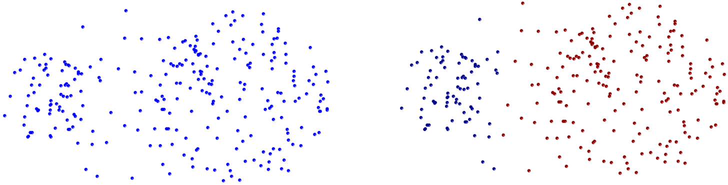
Linkage function: Complete linkage



$$L(r, s) = \max(D(x_{r,i}, x_{s,i}))$$

Linkage function: Complete linkage[3]

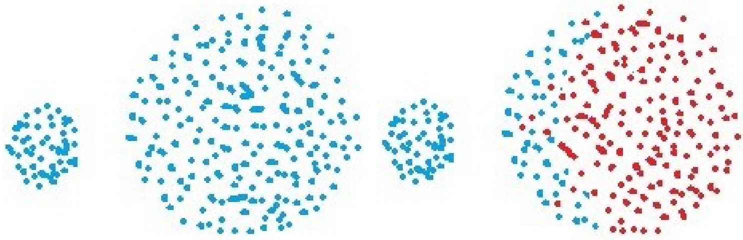
- pros: less susceptible to noise and outliers.



[3] See `https://cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_hierarchical.pdf`

Linkage function: Complete linkage[3]

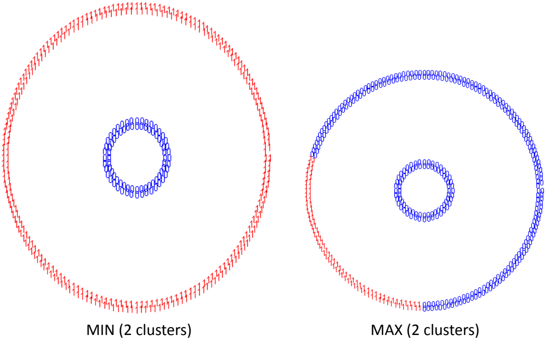
- cons: tends to break large clusters.



[3] See https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

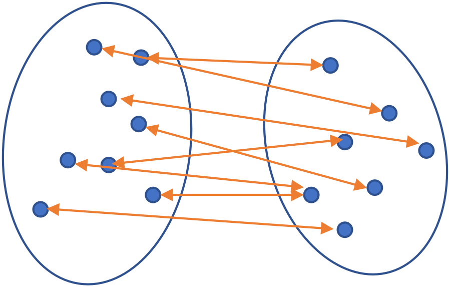
Linkage function: Complete linkage[3]

- cons: biased towards globular clusters.



[3] See https://cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_hierarchical.pdf

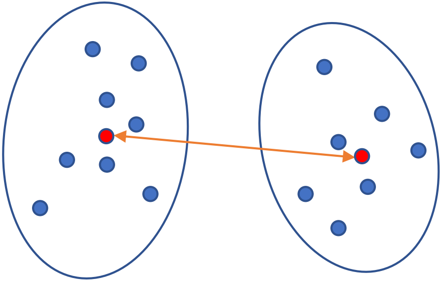
Linkage function: Average linkage



$$L(r,s) = \frac{1}{n_s n_r} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{r,i}, x_{s,i})$$

- Pros: Less susceptible to noise and outliers
- Cons: Biased towards globular clusters

Linkage function: Centroid linkage



$$L(r,s) = D(\hat{x}_r, \hat{x}_s)$$

Linkage function: Ward's Method

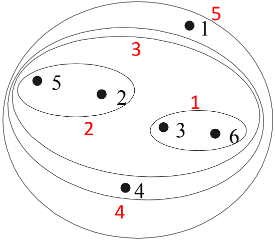
Ward's method says that the distance between two clusters, A and B, is how much the sum of squares will increase when we merge them:

$$\begin{aligned}\Delta(A,B) &= \sum_{i \in A \cup B} \|\vec{x}_i - \vec{m}_{A \cup B}\|^2 - \sum_{i \in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i \in B} \|\vec{x}_i - \vec{m}_B\|^2 \\ &= \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2\end{aligned}$$

where $\vec{m}_j$ is the center of cluster j, and n_j is the number of points in it. Δ is called the merging cost of combining the clusters A and B.

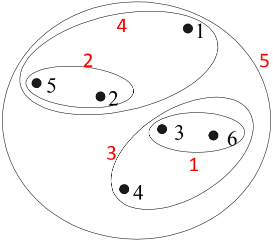
- Pros: Less susceptible to noise and outliers
- Cons: Biased towards globular clusters

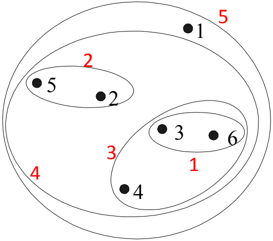
Linkage function: comparison[4]



MIN

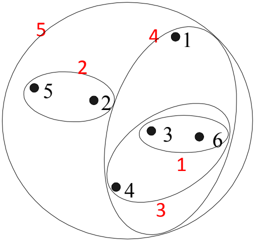
MAX





Group Average

Ward's Method



[4] See https://cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_hierarchical.pdf

Hierarchical Clustering: Interpretation of Dendrogram

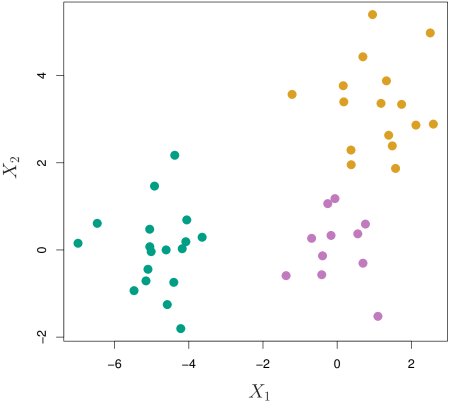
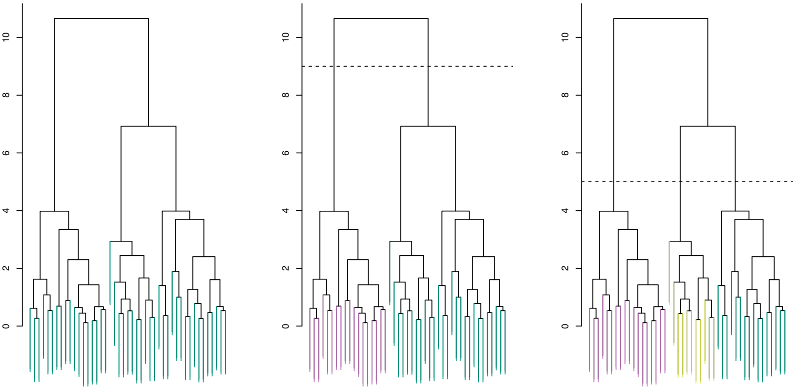


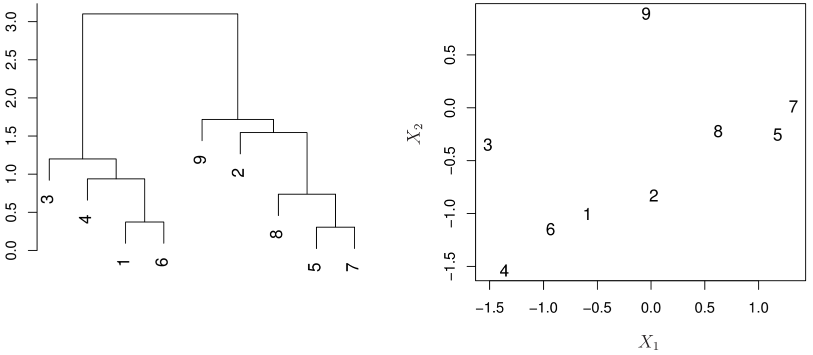
Figure: 45 observations generated in two-dimensional space. In reality there are three distinct classes, shown in separate colours. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.

Hierarchical Clustering: Interpretation of Dendrogram



- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other.
- The height of the fusion indicates how different the two observations are.

Hierarchical Clustering: Interpretation of Dendrogram



- The height of the fusion indicates how different the two observations are.
  - Incorrect: observations 9 and 2 are quite similar
  - We cannot draw conclusion about the similarity of two observations based on their proximity along the horizontal axis.

Hierarchical Clustering: Toy Example

Suppose that we have four observations (X_1, X_2, X_3, X_4), for which we compute a dissimilarity matrix, given by

	X_1	X_2	X_3	X_4
X_1		0.3	0.4	0.7
X_2	0.3		0.5	0.8
X_3	0.4	0.5		0.45
X_4	0.7	0.8	0.45	

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8. On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using **single linkage**.

Clustering: Practical Issues

- Should the observations or features first be standardised in some way? For instance, maybe the variables should be centred 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?
  - Nonparametric Bayesian Methods

Summary

- K-Means clustering
- Hierarchical clustering
- Reading materials:
  - "Unsupervised Learning", Chapter 10 of "Introduction to Statistical Learning", 6th edition
- Acknowledgement:
  - Figures in this presentation were taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
  - Some of the slides are reproduced based on the slides from T. Hastie and R. Tibshirani