Ve572 Lecture 9

Manuel and Jing

UM-SJTU Joint Institute

June 14, 2018

- The process of assigning objects that are similar in some sense with observed

$$\text{features:} \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{ij} \quad \dots \quad x_{ip} \quad i = 1, 2, \dots, n$$

into clusters without a clear definition or knowledge of those clusters is called

clustering

Q: Can you identify the 2 major differences between

classification and clustering

- Clustering is usually the first step, e.g.

1. Cambridge Analytica, people in the same cluster tends to react in similar ways, so having different set of messages and videos for each cluster proves to be highly efficient.
2. Speech recognition, learning and prediction can be done more efficiently within each cluster.

- Given a dataset of n observations with p features:

$$\text{features:} \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{ij} \quad \dots \quad x_{ip} \quad i = 1, 2, \dots, n$$

- Let C_ℓ denotes the set containing the observations in the ℓth cluster.
- K-means clustering is a type of hard clustering that has exactly K clusters:

1. Clusters are non-overlapping

$$C_\ell \cap C_{\ell^*} = \varnothing \qquad \text{for all} \quad \ell \neq \ell^*$$

2. Each observation can belong to exactly one cluster

$$i \in C_\ell \implies i \notin C_{\ell^*} \qquad \text{for all} \quad \ell \neq \ell^*$$

3. The whole dataset is partitioned into the K clusters

$$C_1 \cup C_2 \cup \cdots \cup C_K = \{1, 2, \dots, n\}$$

- The idea behind K-means is that a good clustering is one for which the

  within-cluster variation

  is as small as possible

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{\ell=1}^{K} W\left(C_\ell\right) \right\}$$

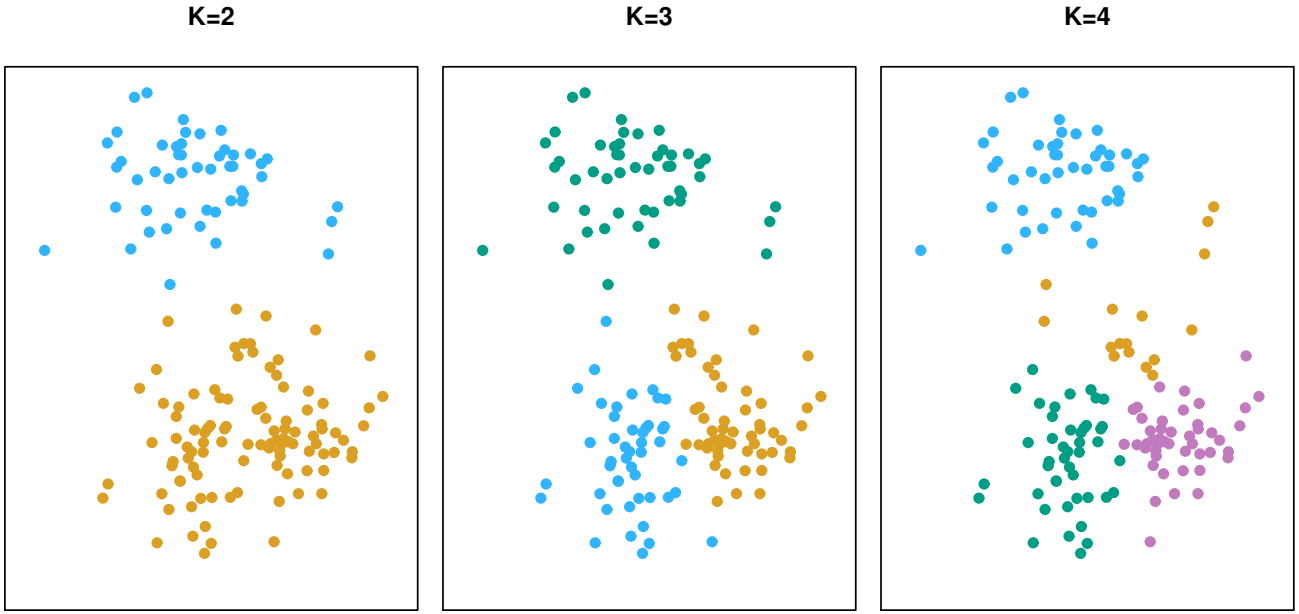
  where $W\left(C_\ell\right)$ is a measure of variation within C_ℓ.

- A common measure of variation is simple total square Euclidean distance

$$W\left(C_\ell\right) = \frac{1}{\|C_\ell\|} \sum_{i,i^* \in C_\ell} \sum_{j=1}^{p} \left(x_{ij} - x_{i^*j}\right)^2$$

  where $\|C_\ell\|$ denotes the number of observations in the ℓth cluster.

- Notice the within-cluster distance is as small as possible



Q: How to solve this optimization problem?

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{\ell=1}^{K} \frac{1}{\|C_\ell\|} \sum_{i,i^* \in C_\ell} \sum_{j=1}^{p} (x_{ij} - x_{i^*j})^2 \right\}$$

- Notice the following identity is true

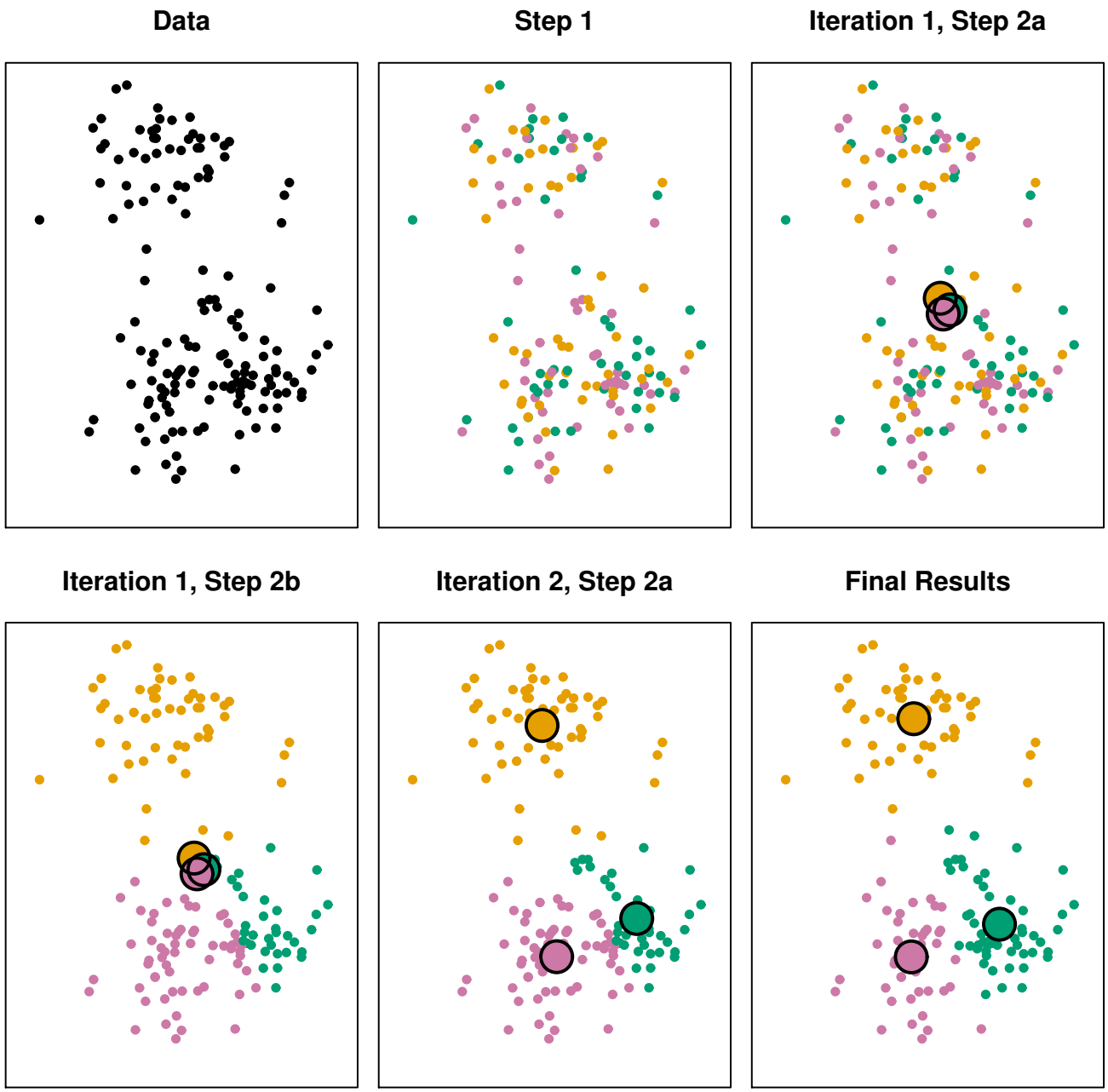
$$\frac{1}{\|C_\ell\|}\sum_{i,i^*\in C_\ell}\sum_{j=1}^{p}(x_{ij}-x_{i^*j})^2 = \frac{2}{\|C_\ell\|}\sum_{i\in C_\ell}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{\ell j})^2$$

where the bar notation denotes the mean for feature j in the ℓth cluster

$$\bar{x}_{\ell j} = \frac{1}{\|C_\ell\|}\sum_{i\in C_\ell} x_{ij}$$

Q: How can we use this identity to derive an optimisation scheme?

1. Randomly assign each of the observation into a cluster out of the K cluster.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster centroid, which is the vector of the p means, $\bar{x}_{\ell j}$, for observations in the ℓth cluster.
   (b) Reassign each observation to the cluster whose centroid is closest according to the Euclidean distance.


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

- Note there is no reason to expect that the K-means will manage to find the global optimum, so it is important to run the algorithm multiple times from different random initial configuration, and select the best solution.

- K-means requires a predetermined K, which is not always available.
- If the goal is to investigate the number of possible clusters, it does not give any information on the possible number of clusters in terms of dissimilarity.

Q: What is a natural way to do clustering when K is not given?

- Imagine we have a bunch of Chinese medicines, and some features for each

`Bitterness`

`Market Price`

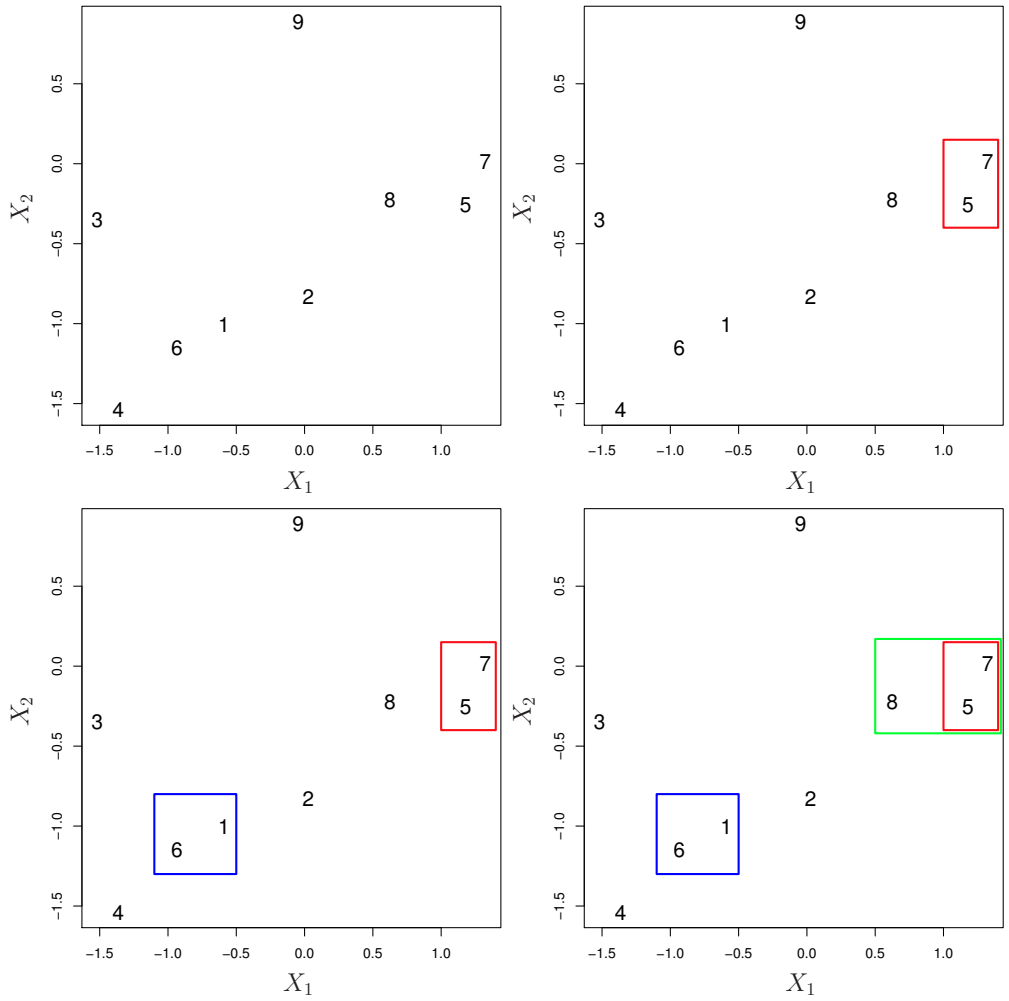
`Density`

`Size`

`RGB Color Spectrum`

Q: How to gradually form clusters to reflect their dissimilarity?

- Given a measure of dissimilarity, we can start with pairwise comparison,



Q: How can we form clusters with more than two observations?

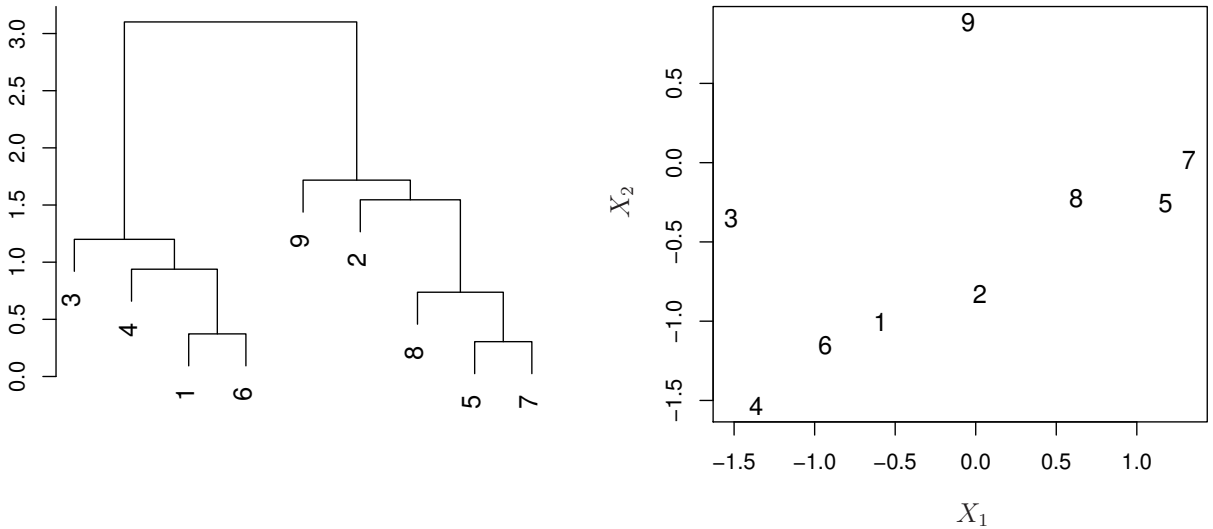
- The definition of dissimilarity can be extended to clusters using linkage.
- Common linkage are given below

Linkage	Description
Complete	Compute all pairwise dissimilarity measures between the observations in cluster A and the observations in cluster B, use the largest value as the inter-cluster dissimilarity.
Single	Compute all pairwise dissimilarity measures between the observations in cluster A and the observations in cluster B, use the smallest value as the inter-cluster dissimilarity.
Average	Compute all pairwise dissimilarity measures between the observations in cluster A and the observations in cluster B, use the average value as the inter-cluster dissimilarity.
Centroid	Compute the centroid of cluster A and cluster B, use the centroid dissimilarity as the inter-cluster dissimilarity.

- This is known as the Hierarchical Clustering. Given n observations:

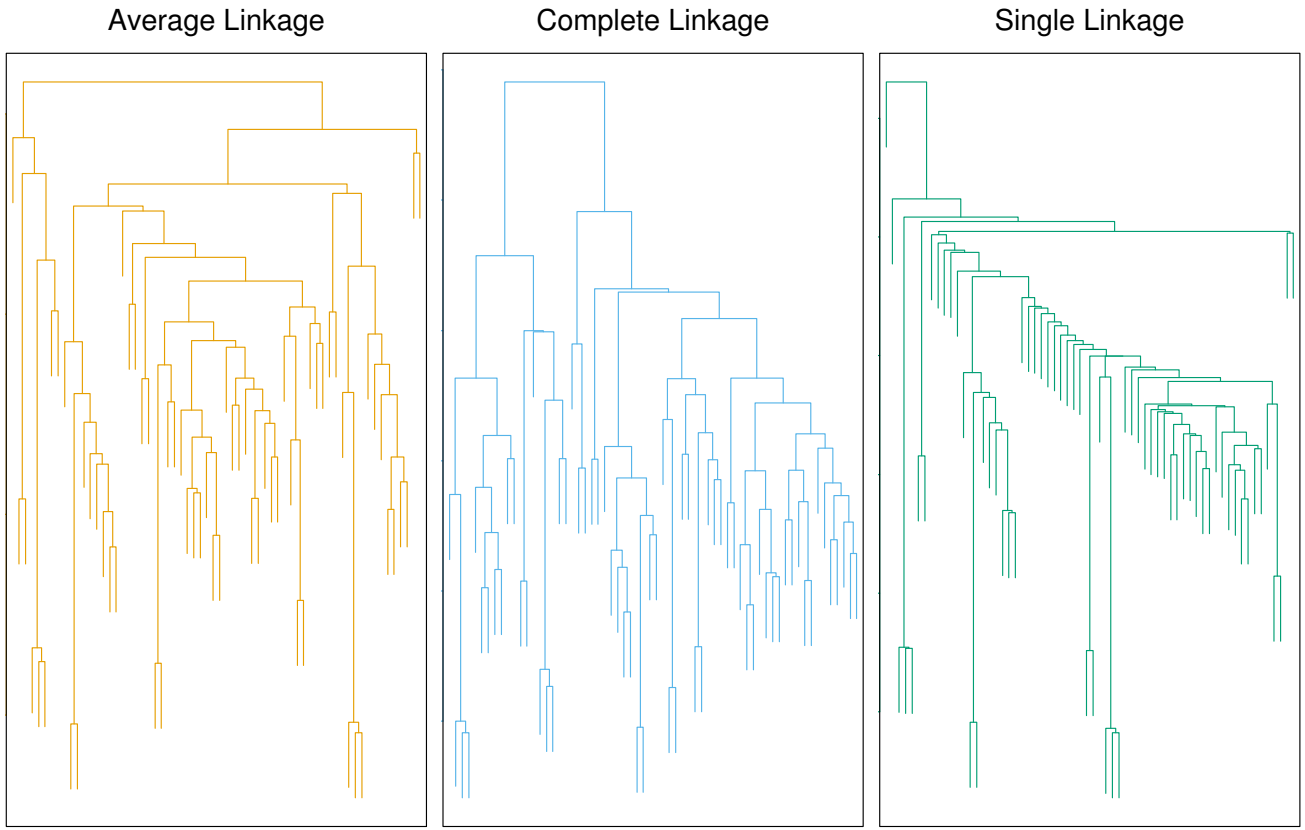
1. Compute the pairwise dissimilarity measure between every 2 observations. Treat each observation as its own cluster.
2. For $i = n, n-1, \ldots, 2$:
   (a) Examine all pairwise inter-cluster dissimilarities among those i clusters.
   (b) Identify the pair of clusters that are least dissimilar, i.e. most similar.
   (c) Fuse these two clusters that are identified.

- A common dissimilarity measure is simply the Euclidean distance, e.g.



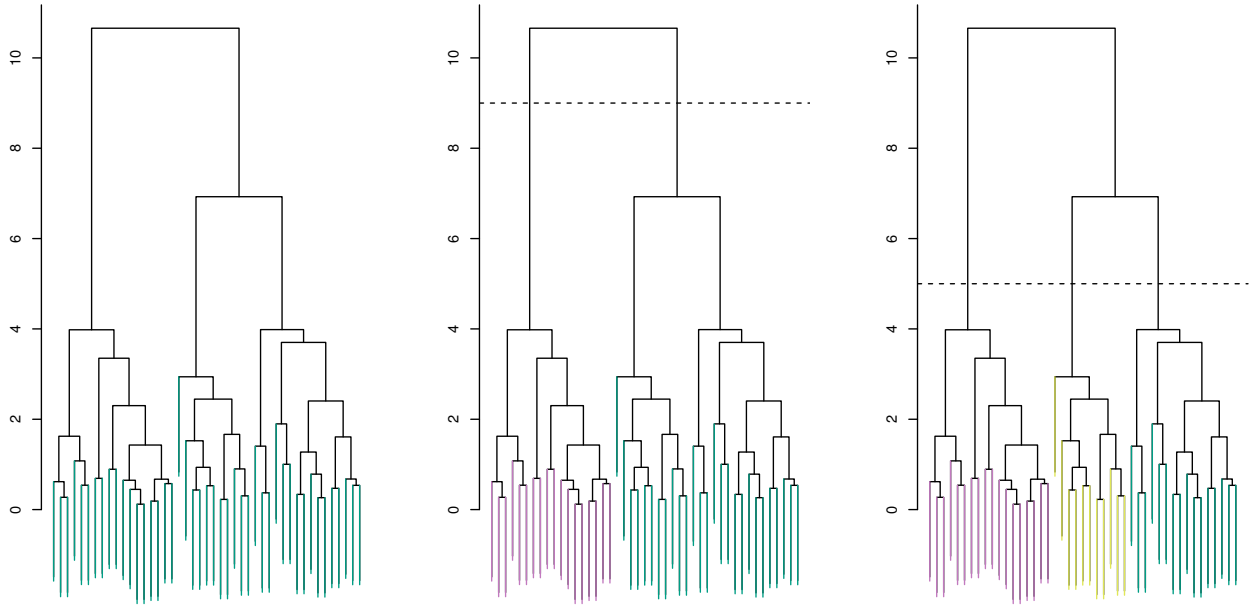
- The plot on the left is known as a dendrogram.

- Centroid linkage is slightly cheaper to use, and often used genomics, but it causes inversion, which is difficult to interpret and visualise in dendrogram.



- Complete and average are usually preferred for they give balanced results.
- Single linkage can result in extended, trailing clusters.

- A very attractive aspect of hierarchical clustering: one single dendrogram can be used to obtain any number of clusters.
- In practice, people often look at the dendrogram and select by eye a sensible number of clusters, based on the heights of the fusion and the number of clusters desired.



- 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.

- A major problem with hierarchical clustering is the fact this assumption of nested structure might be unrealistic.

Q: Can you imagine a dataset that this assumption is clearly not true?

- Football fans?

Russia Vs Saudi Arabia

- If we collect only two features,

`Country`	Russian, Saudis and Others
`Gender`	Female and Male

Q: How many natural clusters are there?

- Due to situations such as this imaginary dataset of football fans, hierarchical clustering can sometimes yield much worse results than K-means clustering for a given number of clusters.

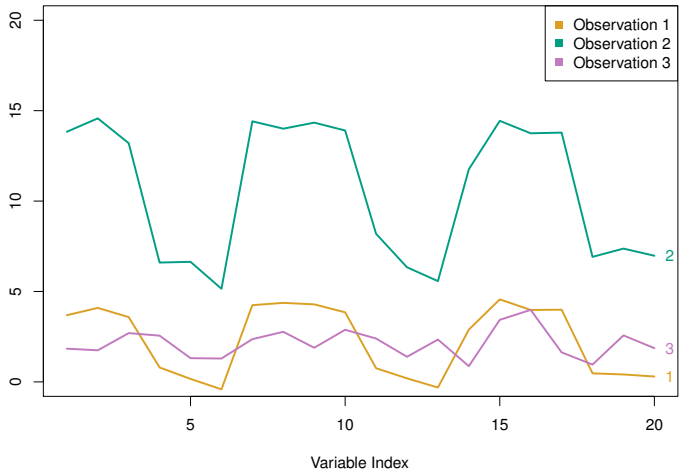
- Choice of dissimilarity measure depends on the type of data.
- Euclidean distance

$$\begin{aligned} \mathrm{d}\left(\mathbf{x}_i, \mathbf{x}_{i^*}\right) &= \left\|\mathbf{x}_i - \mathbf{x}_{i^*}\right\|_2 \\ &= \sqrt{\left(x_{i1} - x_{i^*1}\right)^2 + \left(x_{i2} - x_{i^*2}\right)^2 + \cdots + \left(x_{ip} - x_{i^*p}\right)^2} \end{aligned}$$

- Correlation-based distance

$$\mathrm{d}\left(\mathbf{x}_i, \mathbf{x}_{i^*}\right) = 1 - r \qquad \text{or} \qquad \mathrm{d}\left(\mathbf{x}_i, \mathbf{x}_{i^*}\right) = 1 - |r|$$

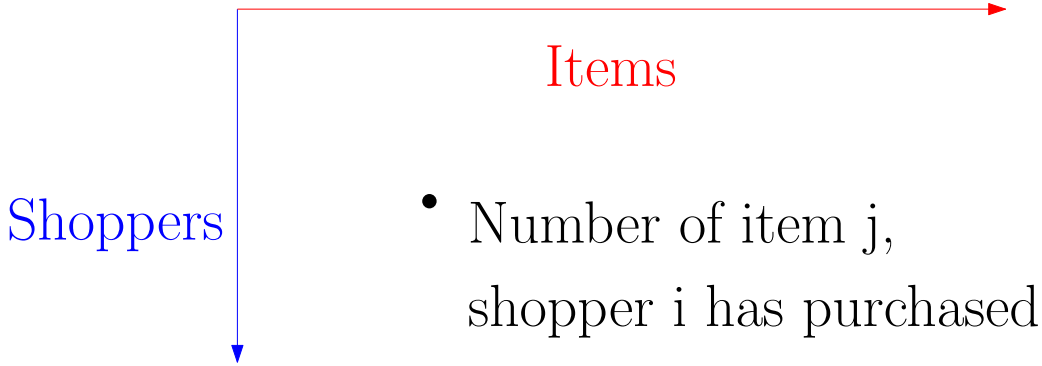
where r is the Pearson or Spearman correlation coefficient.



- For instance, consider the online retailers cat and dog

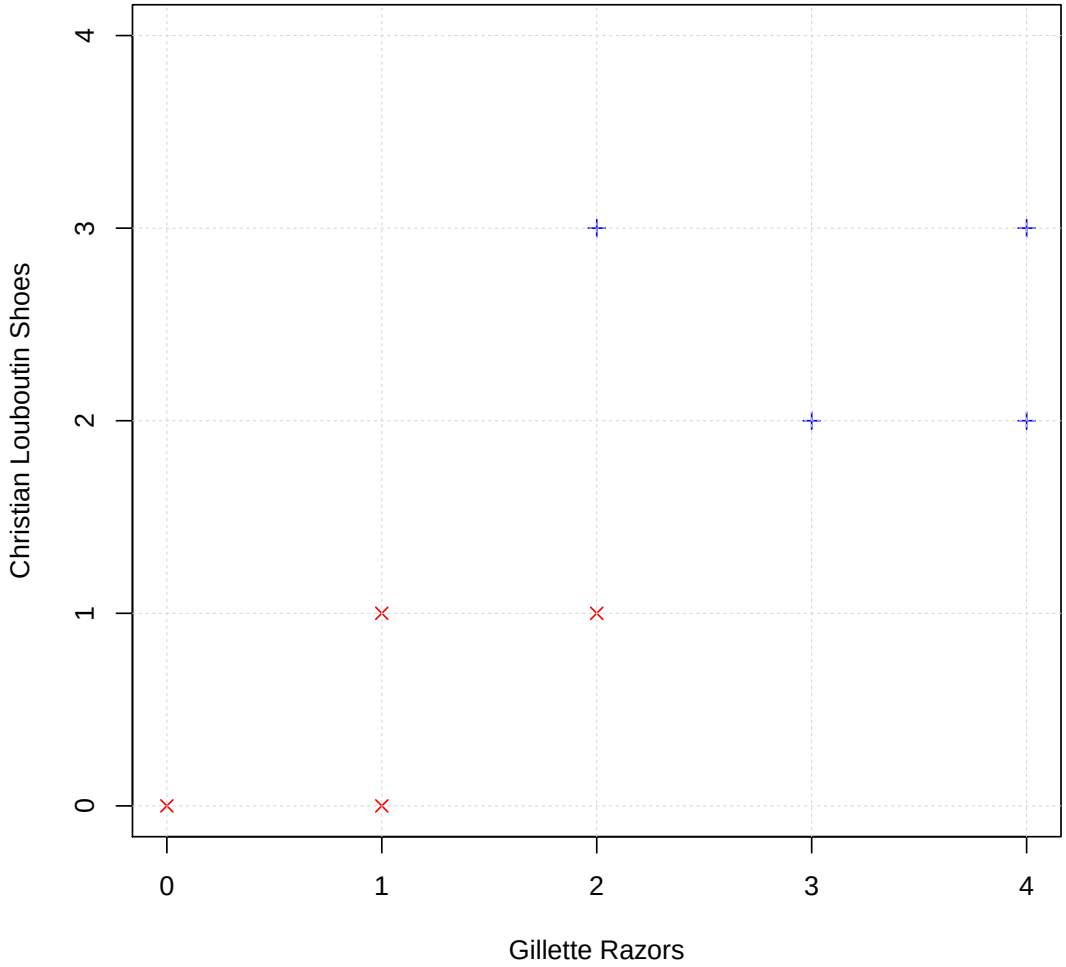
  interested in clustering shoppers based on their past shopping histories.

- Suppose the data takes the usual rectangular form



Q: What type of dissimilarity measure should be used to cluster the shoppers?

- If Euclidean distance is used, then



users will clustered overwhelmingly according to their usage.

- Consider the following dataset

	Gillette Razors	Christian Louboutin Shoes	Air Jordan Shoes
obs. 1	8	3	7
obs. 2	1	0	1
obs. 1	5	9	3

- If Euclidean distance is used, we will have following matrix of dissimilarity

$$\begin{bmatrix} & 9.69536 & 7.81025 \\ 9.69536 & & 10.04988 \\ 7.81025 & 10.04988 & \end{bmatrix}$$

- However, if a correlation-based distance is used, we will have

$$\begin{bmatrix} & 0.01801949 & 1.86602540 \\ 0.01801949 & & 1.94491118 \\ 1.86602540 & 1.94491118 & \end{bmatrix}$$

Q: Do you notice the difference and impact?

- Another approach is to standardising the data, which is done for each feature

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j}$$

where $\bar{x}_j$ and s_j are the mean and s.d. of the jth feature, respectively.

- After standardisation, we have the following instead

	Gillette Razors	Christian Louboutin Shoes	Air Jordan Shoes
obs. 1	0.9491580	-0.2182179	1.0910895
obs. 2	-1.0440738	-0.8728716	-0.8728716
obs. 1	0.0949158	1.0910895	-0.2182179

and the corresponding matrix of dissimilarity is given by

$$\begin{bmatrix} & 2.873793 & 2.039191 \\ 2.873793 & & 2.362840 \\ 2.039191 & 2.362840 & \end{bmatrix}$$

- Notice observation 3 is now closer to observation 2 than to observation 1.

- Scaling the data instead of standardising the data