



## **BC2406 Business Analytics I: Predictive Techniques**

# **Seminar 9** **Clustering**

**Instructor: Prof. Lee Gun-woong**

*Nanyang Business School*

# Today's Outline

- Overview of Clustering
- Measures in Clustering
  - Similarity and Distance
- Preparation for Clustering
  - Data and Sample Selection
  - Variable Standardization
- Types of Clustering
  - Hierarchical Clustering
  - K-means Clustering
- Assessing Clustering Results

# Overview of Clustering

# Clustering

- Clusters: groups of *similar* items/objects
- Clustering is an **unsupervised data-mining task** that divides the data into clusters.
  - Clustering is not driven by some pre-specified target characteristics.
  - Clustering is used for knowledge discovery rather than prediction
- Clustering is guided by the principle that the objects **within a cluster should be very similar to each other**, but the objects in different clusters are not so similar.
  - Group the data so that the related elements are placed together!
- Example: Do our customers naturally fall into different groups?
  - Do we understand who our customers are?
  - Can we develop better products, better marketing campaigns, better sales methods, or better customer service by understanding the **natural grouping**?

# Clustering and Business Problems

- Clustering is searching for *patterns* in complex data
  - Patterns can lead to business decisions
- Business Problems with Clustering
  - Who will buy full-price hardcovers?
  - What food court promotions are most effective at certain times of day?
  - What sorts of complaints are most common for different call centers?
- Clustering is useful whenever diverse and varied data can be exemplified by a much **smaller number of groups**.

# More Applications

- **Bank/Internet Security**: fraud/spam pattern discovery
- **Biology**: taxonomy of living things such as kingdom, phylum, class, order, family, genus and species
- **City-planning**: Identifying groups of houses according to their house type, value, and geographical location
- **Climate change**: understanding earth climate, find patterns of atmospheric and ocean
- **Finance**: stock clustering analysis to uncover correlation underlying shares
- **Image Compression/segmentation**: coherent pixels grouped
- **Information retrieval/organisation**: Google search, topic-based news
- **Land use**: Identification of areas of similar land use in an earth observation database
- **Marketing**: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Social network mining**: special interest group automatic discovery

# Classify the following:

- Take a piece of paper and classify the following bags into **x** number of clusters, where **x** can be any number you wish. Explain how you classify



# Archetypes and Clustering

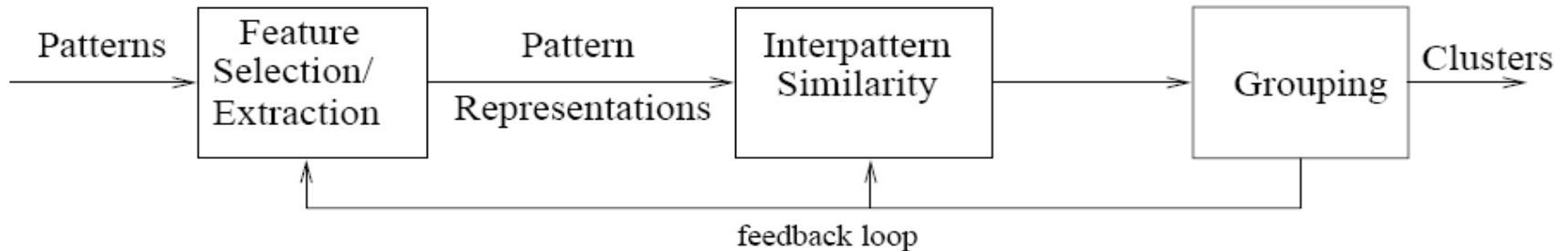
- Believe that lower level observations can be clustered into higher orders
  - Types of customer segments, e.g. airlines
  - Types of companies e.g. prospector, defender, analyzer, reactor
  - Types of living organisms
- What can we do with this clustering?
  - Influence business strategy
  - Possible predicting future behavior



# Clustering with Machines

- Without advance knowledge of what comprises a cluster, how can a computer possibly know where one group ends and another begins?
  - Clustering is guided by the principle that objects inside a cluster should be very similar to each other, but very different from those outside.
  - Your computer identifies a set of clusters in an unsupervised manner.
    - No information is provided to the algorithm on which data points belong to which clusters.

# Key Steps in Clustering



- **Feature Selection**
  - Identifying the most effective subset of the original features to use in clustering
- **Feature Extraction**
  - Transformations of the input features to produce new salient features.
- **Inter-pattern Similarity**
  - Measured by a distance function defined on pairs of patterns.
- **Grouping**
  - Methods to group similar patterns in the same cluster

# Measures in Clustering

# Similarity

- Clustering groups the data so that the related elements are placed together (i.e., similarity)
- What is similarity?
- To illustrate the difficulties involved in judging similarity, consider your answer to the following question:
  - Which is more similar to a duck; a crow or a penguin?



Am I similar to



or



?



NANYANG  
TECHNOLOGICAL  
UNIVERSITY

# Similarity and Distance

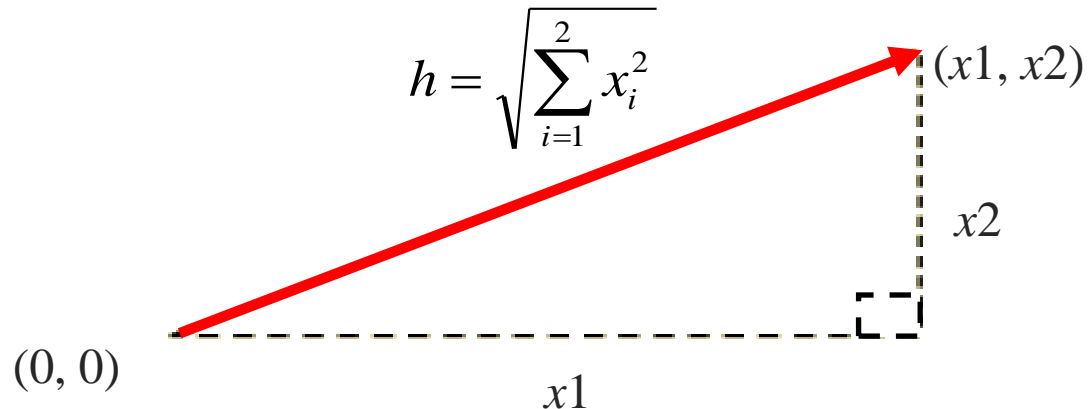
- In Data Science
  - Similarity between objects = Distance between objects
  - The closer two objects are in the space defined by the features, the more similar they are
- Can we measure the similarity or distance between objects directly?
- Consider two instances from credit application data

Attribute	Person A	Person B
Age	23	40
Years at current address	2	10
Residential Status(1=Owner, 2=Renter)	2	1
Gender (1=Female, 2=Male)	1	1

# Euclidean Distance

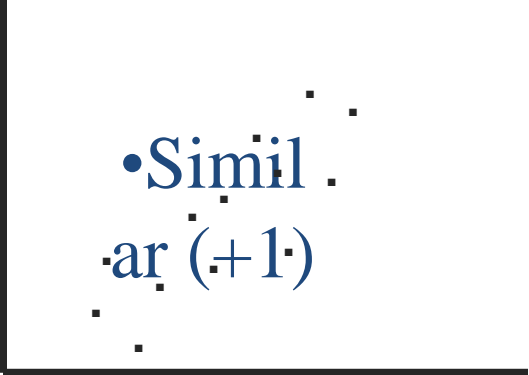
$$D_E = \|\mathbf{x} - \mathbf{w}\| = \sqrt{\sum_{i=1}^k (x_i - w_i)^2}$$

- Euclidean distance gives the linear distance between any two points in  $n$ -dimensional space.
- It is a generalization of the Pythagorean theorem.



# Pearson Correlation

• Similar (+1)



No association (0)



Dissimilar (-1)



# The Problem with Correlation

Attribute	Person A	Person B
x1	5	51
x2	4	42
x3	3	33
x4	2	24
x5	1	15

- The correlation between Person A and Person B is **1.0** which implies perfect association.
  - They are indistinguishable from one another looking at the correlation coefficient.
- But are the customers really similar?



# Preparation for Clustering

# Data and Sample Selection

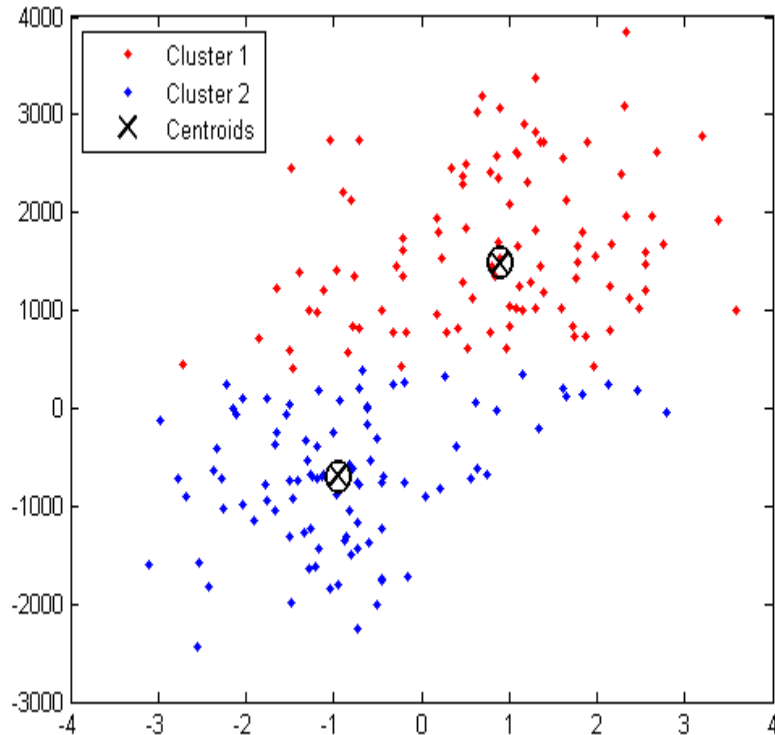
- It is not necessary to cluster a large population if you use clustering techniques that lend themselves to scoring.
  - e.g., Instance A's probability of being included in Cluster 1.
- In these cases, it is useful to take a random sample for clustering and score the remainder of the larger population.

# Variable Standardization

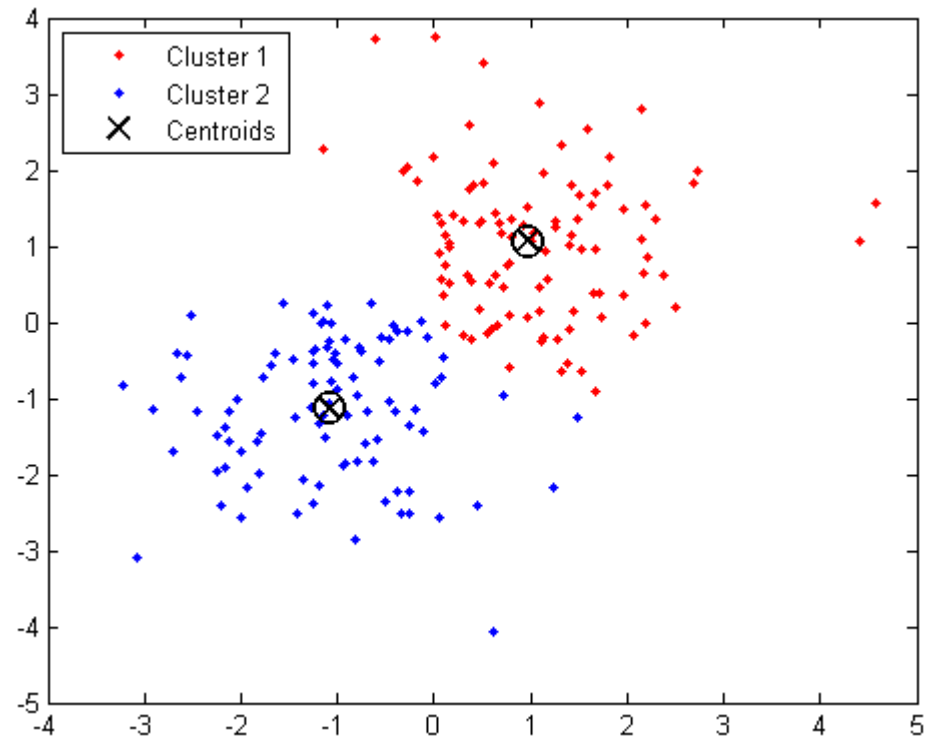
- Raw distance measures are highly influenced by scale of measurements.
- A common practice employed prior to any analysis using distance calculations is to normalize or z-score standardize the variables so that each utilizes the same range.
  - To avoid a problem in which some variables come to dominate solely because they have a large range of values than the other
  - **Note:** *Cluster analysis requires a data frame containing only numeric variables!*

# Example: Variable Standardization

Before Standardization



After Standardization



# Z-score Standardization

- Rescale the variables so that they have a mean of zero and a standard deviation of one
  - Z-score Standardization:  $Z = \frac{x - \text{mean}(X)}{SD(X)}$
  - After Standardization: *Variables*  $\sim N(0,1)$
- Example
  - Variables: Sales of items in 20 countries (**SG: \$9,077**)
    - Mean(Sales) = 8,914
    - SD(Sales) = 3,550
    - Z-score for SG Sales =  $(9,077 - 8,914) / 3,550 = 0.046$

# Standardization in R

- The *scale* () function allows us to standardize the variables.

– Example: Wine Data in the tutorial

```
> WineData_Stand <- scale(WineData)
```

## Before Standardization

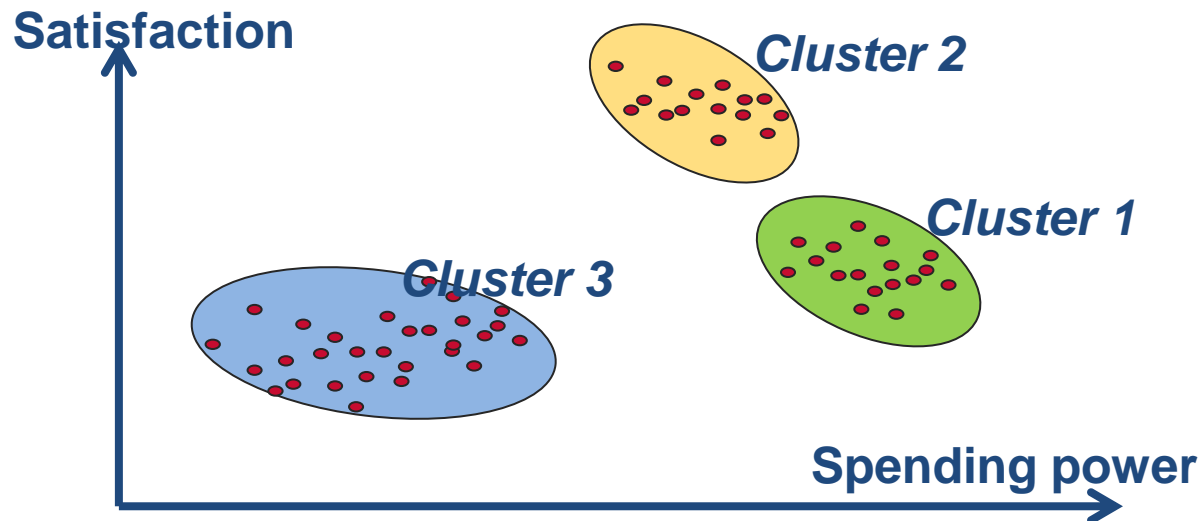
	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proanthocyanins	Color	Hue	Dilution	Proline
1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
2	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
3	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
4	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
5	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735
6	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450

## After Standardization

	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proanthocyanins	Color	Hue	Dilution	Proline
[1,]	1.5143408	-0.56066822	0.2313998	-1.1663032	1.90852151	0.8067217	1.0319081	-0.6577078	1.2214385	0.2510088	0.3611585	1.8427215	1.01015939
[2,]	0.2455968	-0.49800856	-0.8256672	-2.4838405	0.01809398	0.5670481	0.7315653	-0.8184106	-0.5431887	-0.2924962	0.4049085	1.1103172	0.96252635
[3,]	0.1963252	0.02117152	1.1062139	-0.2679823	0.08810981	0.8067217	1.2121137	-0.4970050	2.1299594	0.2682629	0.3174085	0.7863692	1.39122370
[4,]	1.6867914	-0.34583508	0.4865539	-0.8069748	0.92829983	2.4844372	1.4623994	-0.9791134	1.0292513	1.1827317	-0.4263410	1.1807407	2.32800680
[5,]	0.2948684	0.22705328	1.8352256	0.4506745	1.27837900	0.8067217	0.6614853	0.2261576	0.4002753	-0.3183774	0.3611585	0.4483365	-0.03776747
[6,]	1.4773871	-0.51591132	0.3043010	-1.2860793	0.85828399	1.5576991	1.3622851	-0.1755994	0.6623487	0.7298108	0.4049085	0.3356589	2.23274072

# Approximate the Number of Clusters

- Plotting can help to determine such key things as
  - the shape of the clusters
  - relative cluster dispersion (variation)
  - the approximate number of clusters in the data.



# Types of Clustering



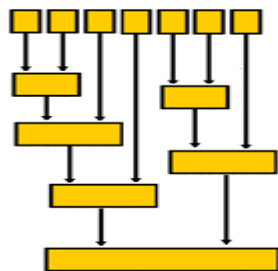
# Clustering Algorithms

- Popular Clustering Algorithms
  - Hierarchical
  - k-means
  - k-medoids
  - Fuzzy c-means
- Different ways of measuring similarity

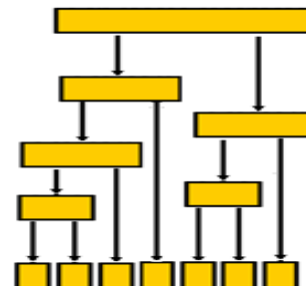
# Hierarchical Methods

- Given the input set  $S$ , the goal is to produce a hierarchy in which nodes represent subsets of  $S$ .
- **Agglomerative Methods**
  - Begin with  $n$ -clusters (each record its own cluster)
  - Keep joining records into clusters until one cluster is left (the entire data set)
  - **Most popular**
- **Divisive Methods**
  - Start with one all-inclusive cluster
  - Repeatedly divide into smaller clusters

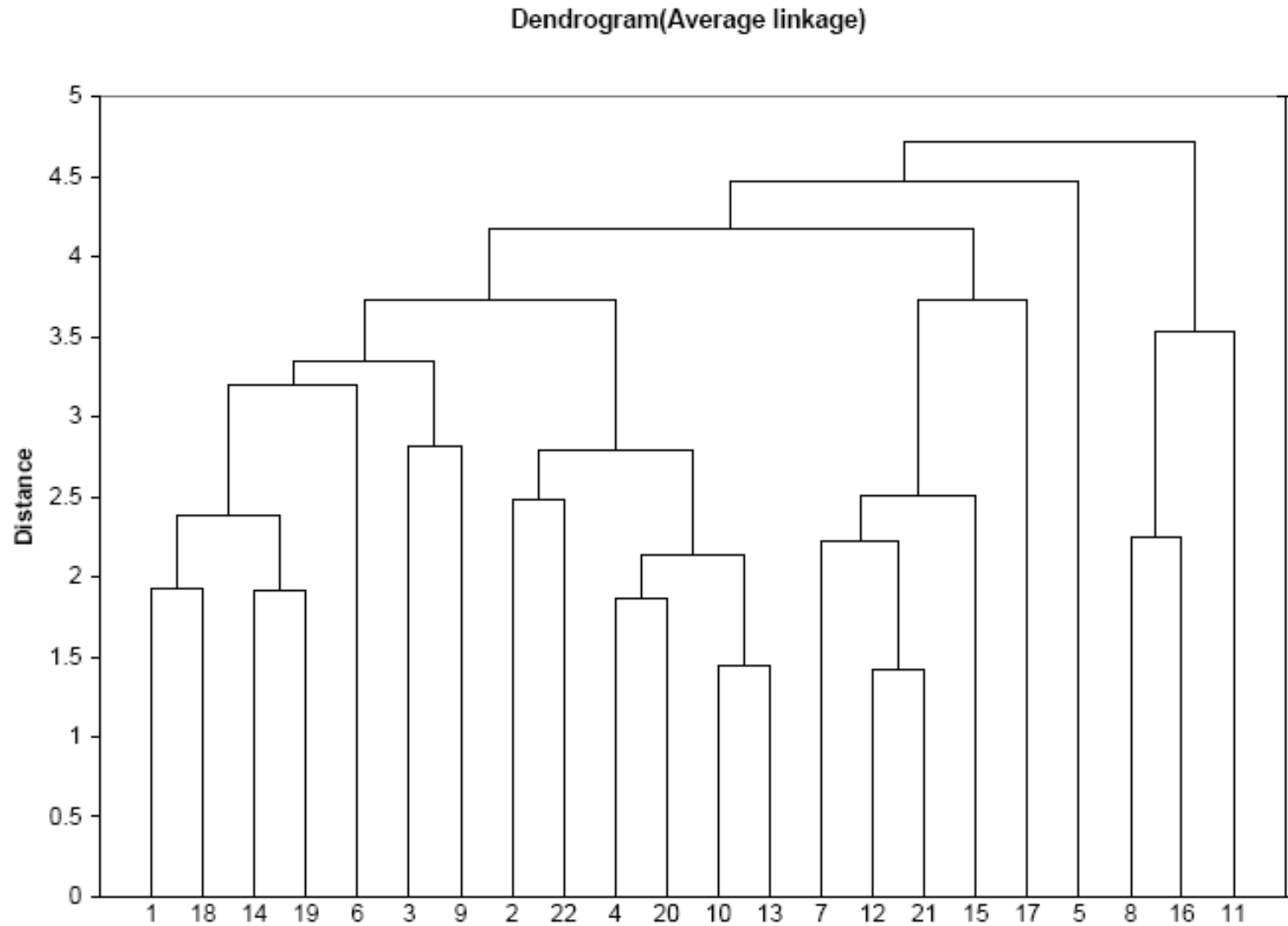
Agglomerative



Divisive



## A Dendrogram shows the cluster hierarchy



# Measuring Distance in Hierarchical Clustering

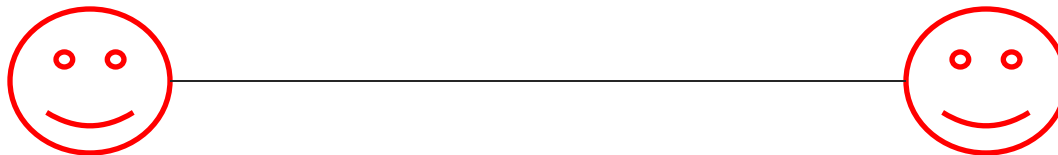
Between instances/records

Between clusters

# Distance Between Two Instances

**Euclidean Distance** is most popular:

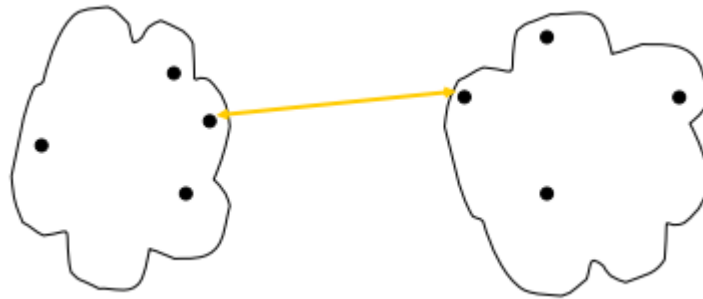
$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$



# Measuring Distance Between Clusters

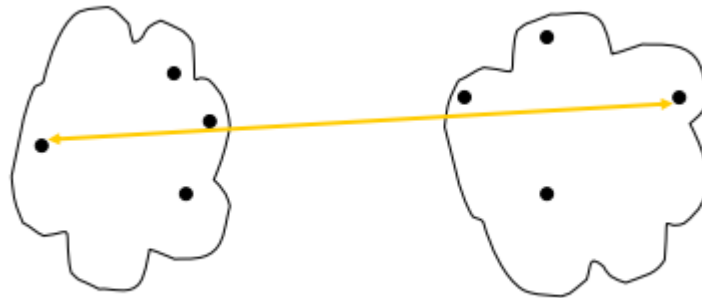
# Minimum Distance (Cluster A to Cluster B)

- Also called **single linkage** (nearest neighbour)
- Distance between two clusters is the distance between the pair of records  $A_i$  and  $B_j$  that are closest



# Maximum Distance (Cluster A to Cluster B)

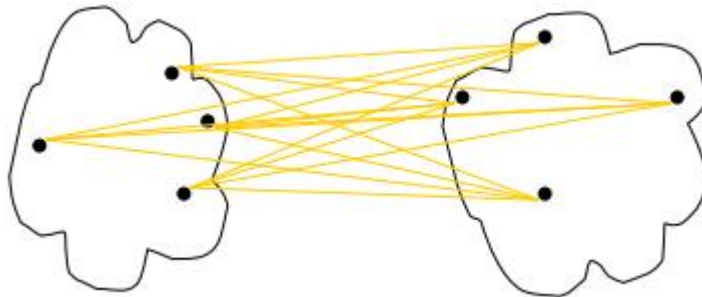
- Also called **complete linkage** (furthest neighbour)
- Distance between two clusters is the distance between the pair of records  $A_i$  and  $B_j$  that are farthest from each other





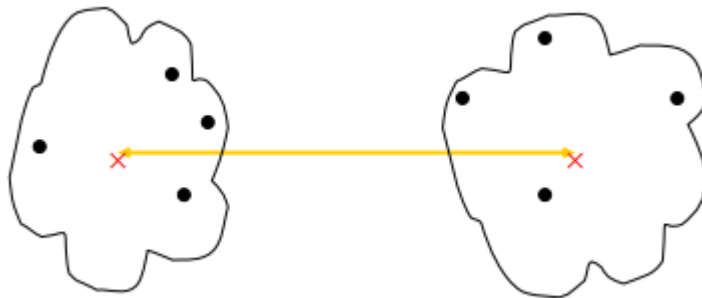
# Average Distance

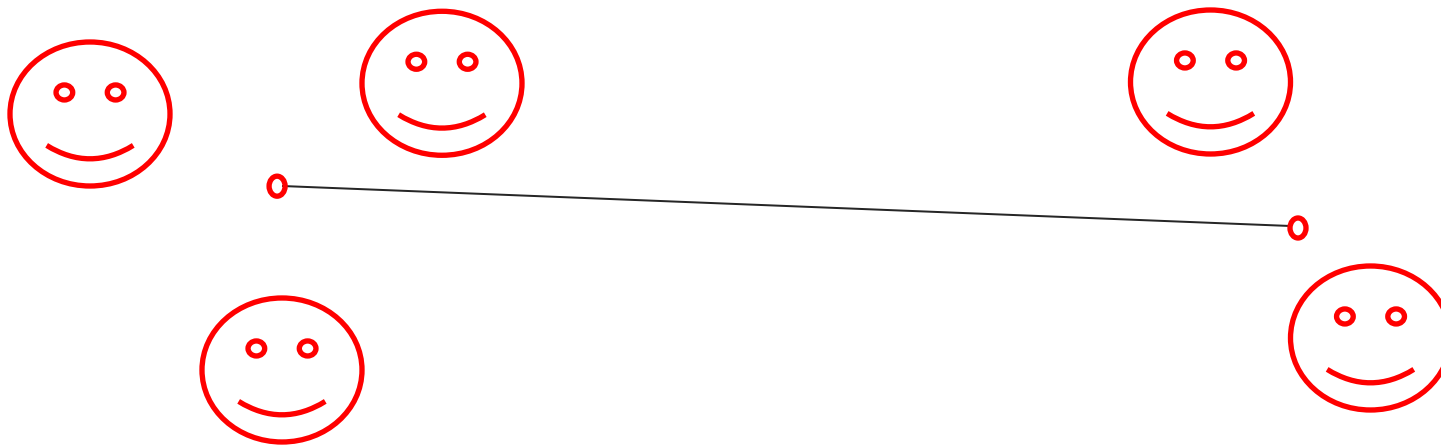
- Also called **average linkage**
- Distance between two clusters is the average of all possible pair-wise distances
- Most popular: as don't we just love averages



# Centroid Distance

- Distance between two clusters is the distance between the two cluster centroids.
- Centroid is the vector of variable averages for all records in a cluster (i.e., center of a cluster)
- Also popular, more for non-hierarchical





Distance between the centers of each cluster

# Ward's Method

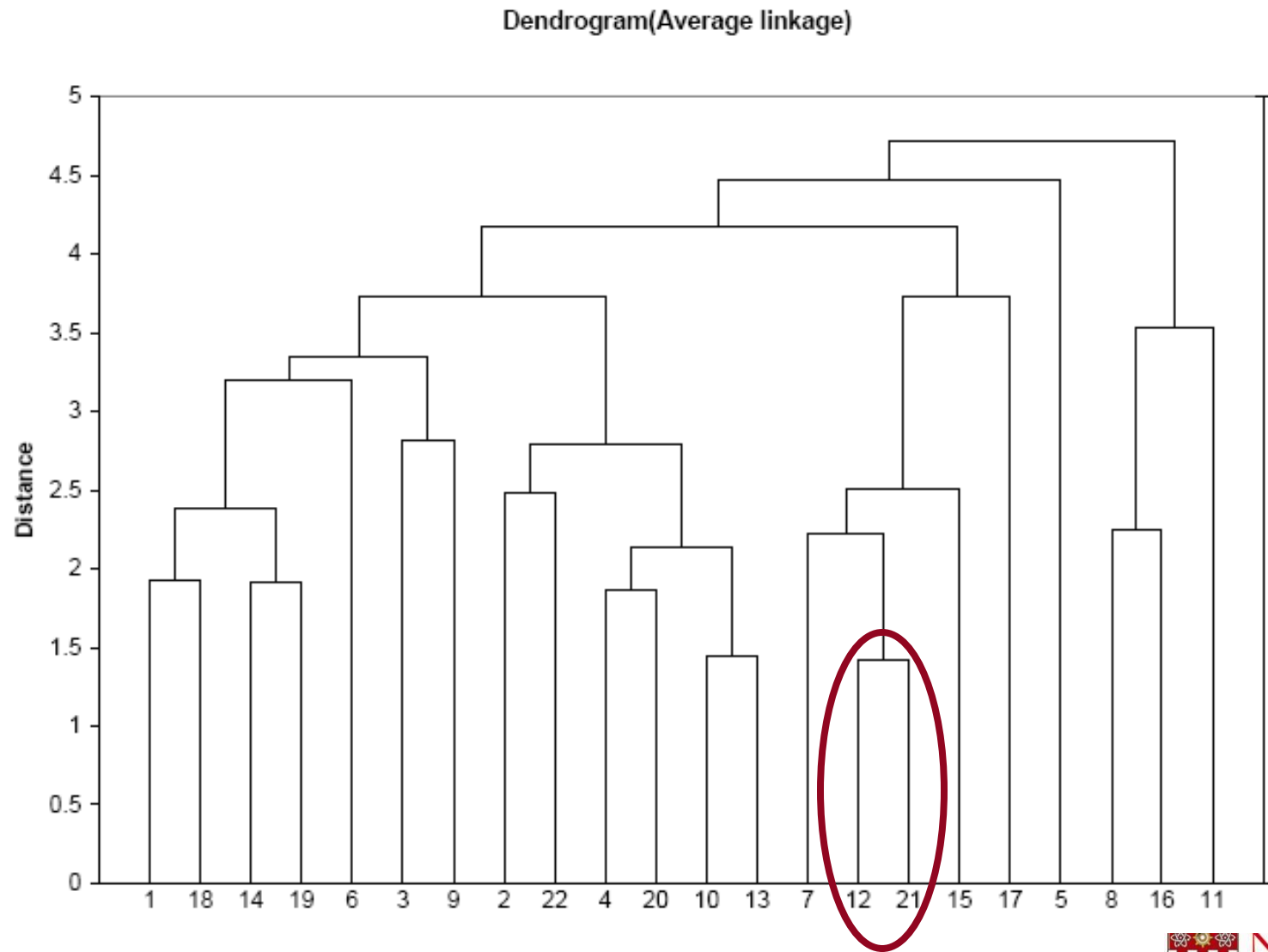
- Looks at cluster analysis as an **analysis of variance** problem, instead of using distance metrics.
  - This method is used for an agglomerative clustering algorithm.
- Measures how much the sum of squares will increase when  $A_i$  and  $B_j$  are merged
  - Similarity of two clusters is based on the increase in squared error when two clusters are merged.
- Less susceptible to noise and outliers

# The Hierarchical Clustering Steps (Using Agglomerative Method)

1. Start with  $n$  clusters (each record is its own cluster)
2. Merge two closest records into one cluster
3. At each successive step, the two clusters closest to each other are merged

Dendrogram, from bottom up, illustrates the process

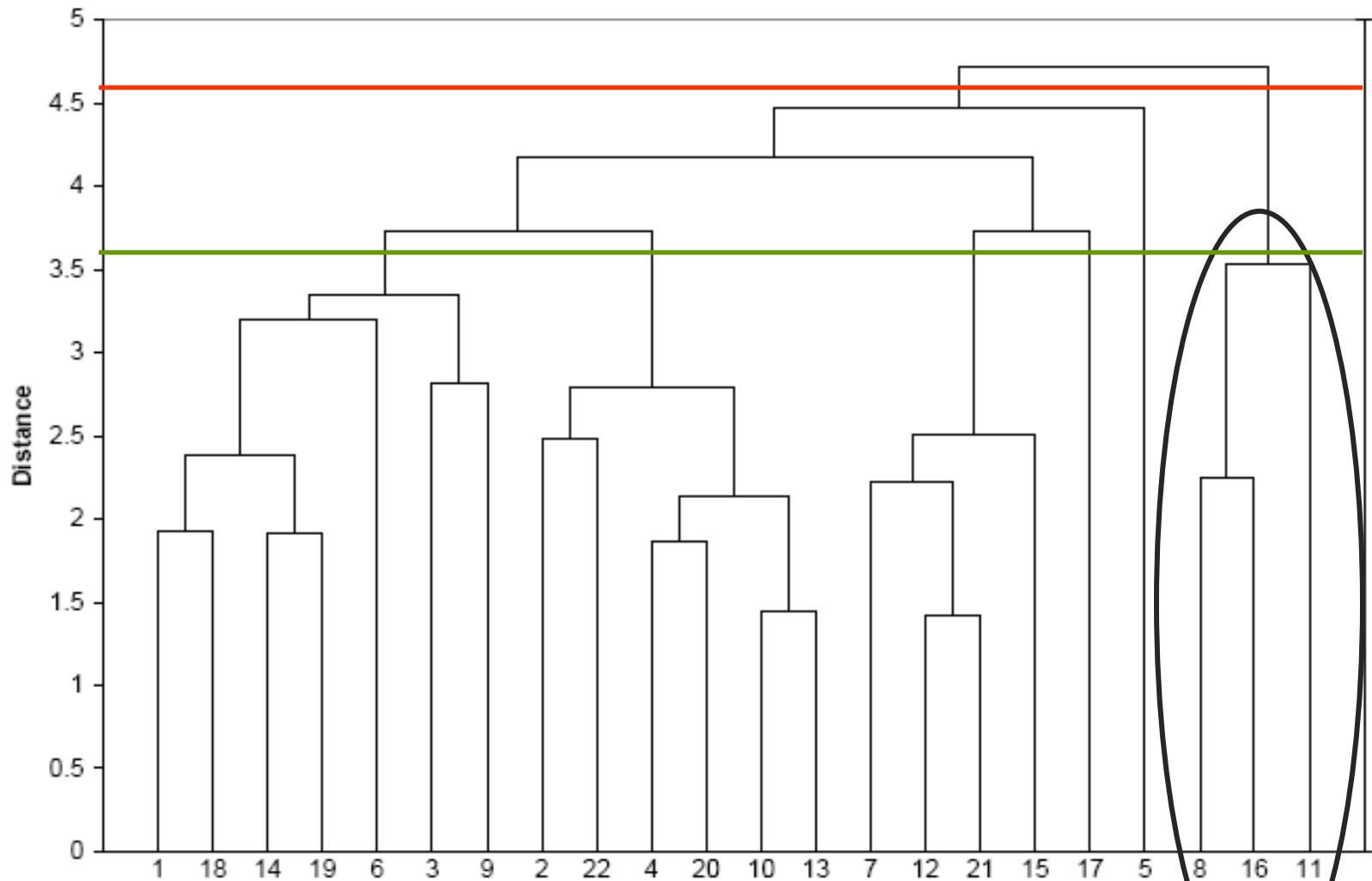
Records 12 & 21 are closest & form first cluster



# Reading the Dendrogram

- **See process of clustering:** Lines connected lower down are merged earlier
  - 10 and 13 will be merged next, after 12 & 21
- **Determining number of clusters:** For a given “distance between clusters”, a horizontal line intersects the clusters that are that far apart, to create clusters
  - E.g., at distance of 4.6 (**red line** in next slide), data can be reduced to 2 clusters -- The smaller of the two is circled
  - At distance of 3.6 (**green line**) data can be reduced to 6 clusters, including the circled cluster

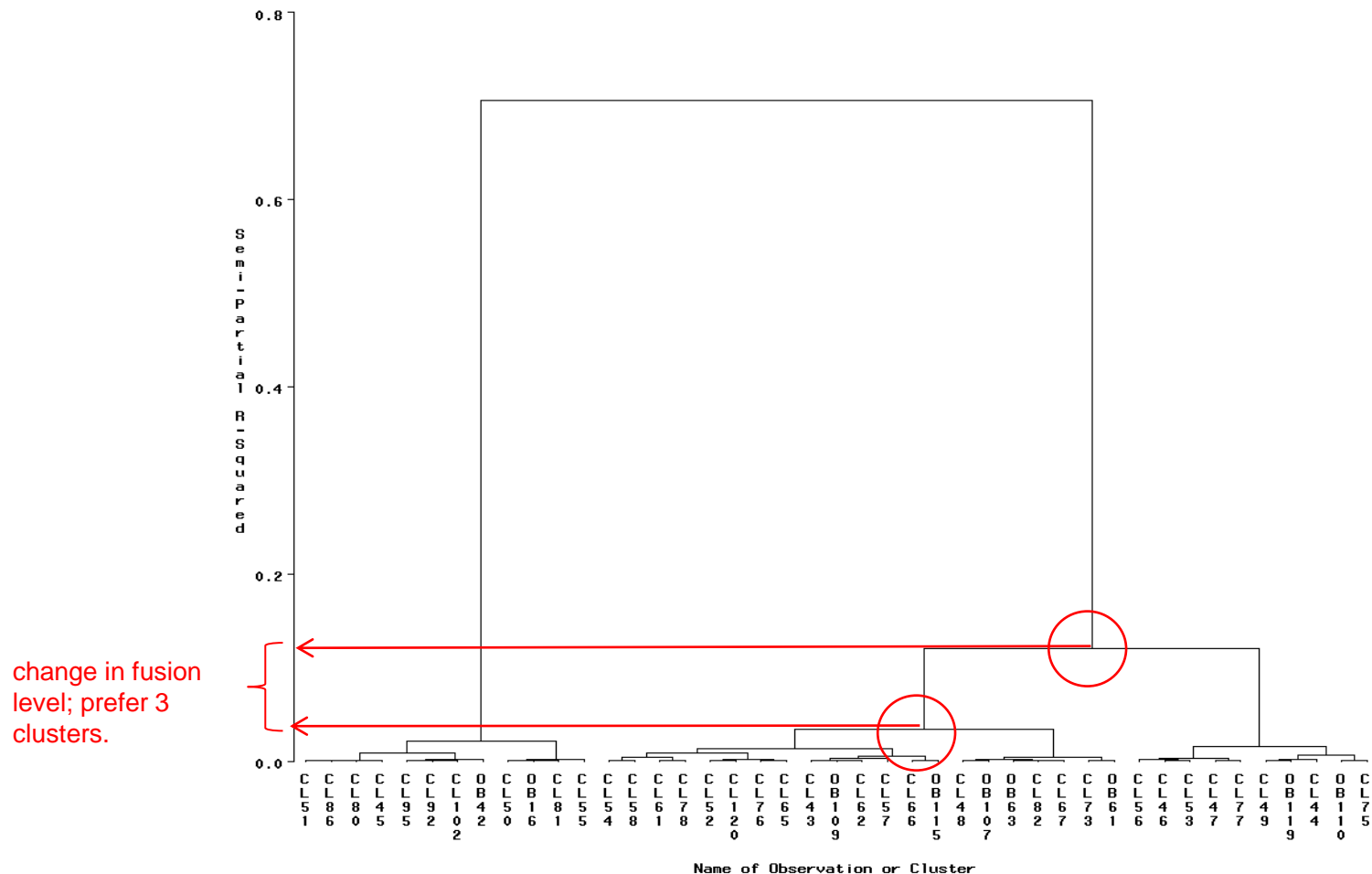
Dendrogram(Average linkage)



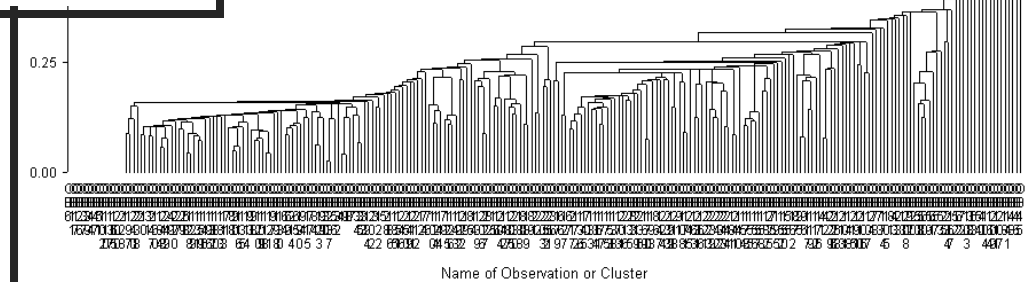
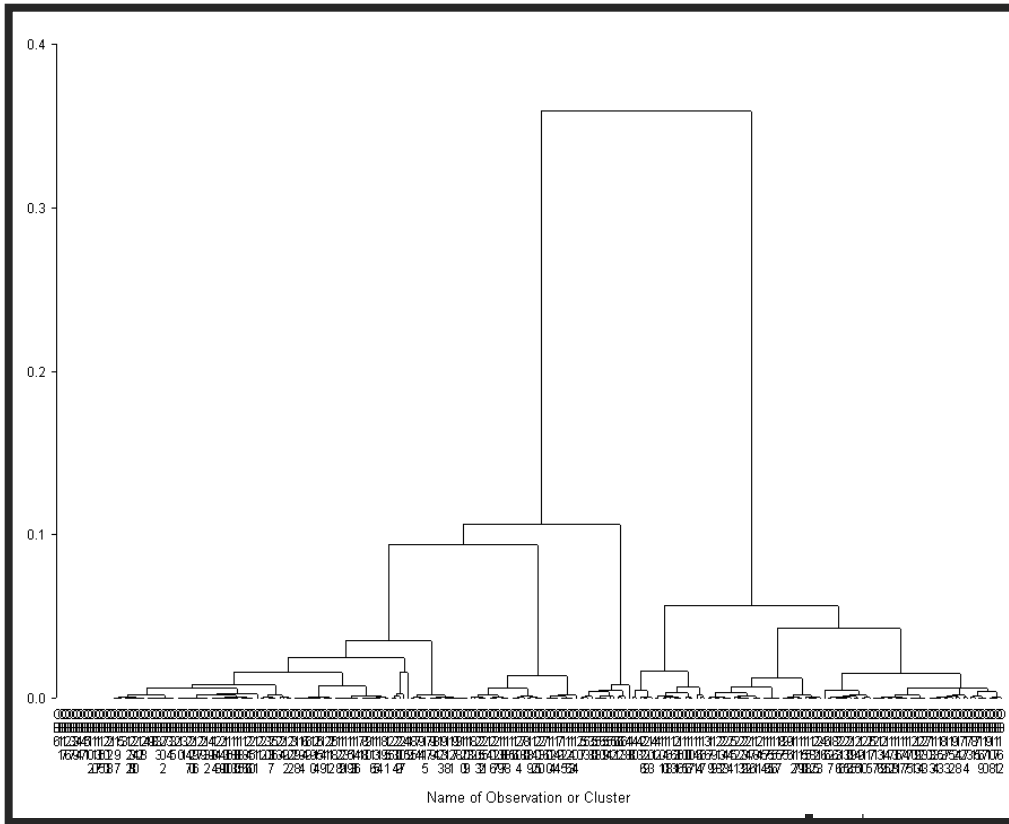


# Interpreting Dendrograms

- For interpreting **any** hierarchical clustering method



# Examples of Dendrograms



# Pros and Cons

- Pros:
  - Easy to understand (and well-accepted)
  - Not required to determine the number of clusters beforehand
    - The desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level
- Cons:
  - Sensitive to distance metrics
  - Slow
    - It has to make several merge/split decisions

# Partitive Clustering: **k-means Clustering**

# K-Means Clustering Algorithm

1. Determine the number of  $k$  (i.e., number of clusters)
2. Randomly select  $K$  centroids (i.e. centers of clusters)
3. Assign each data point to its closest centroid
4. Recalculate the centroids as the average of all data points in a cluster
5. Assign data points to their closest centroids
6. Repeat Step 4 and 5 until the observations are not reassigned or the maximum number of iterations is reached

# K-means Algorithm: Choosing $k$ and Initial Partitioning

Choose  $k$  based on the how results will be used

e.g., “How many market segments do we want?”

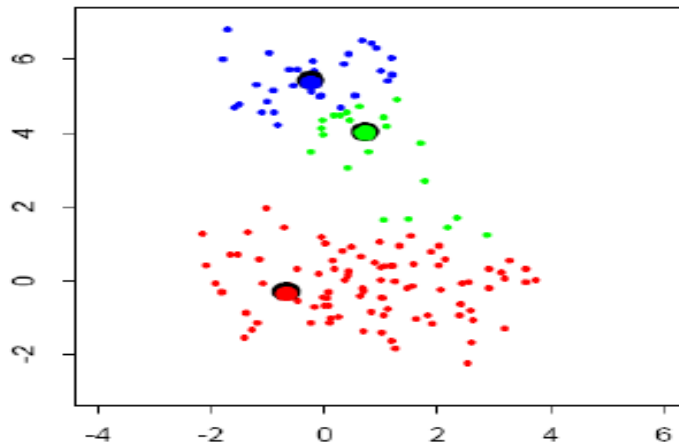
Also experiment with slightly different  $k$ 's

Initial partition into clusters can be random, or based on domain knowledge

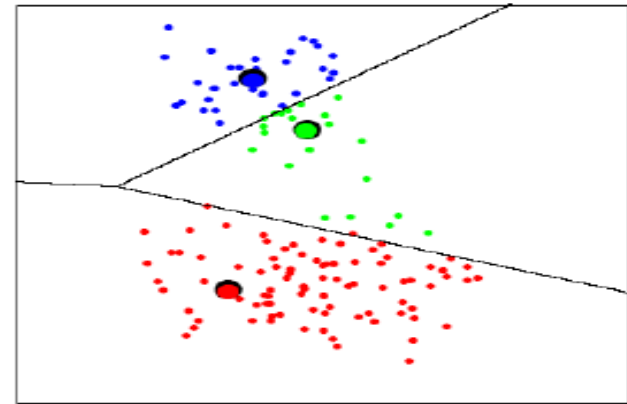
If random partition, repeat the process with different random partitions

# K-Means Clustering Example

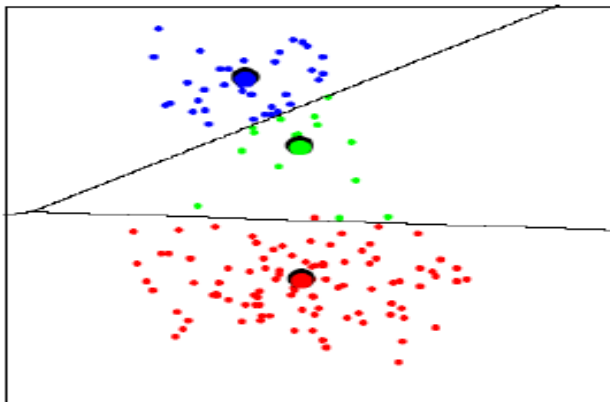
Initial Centroids



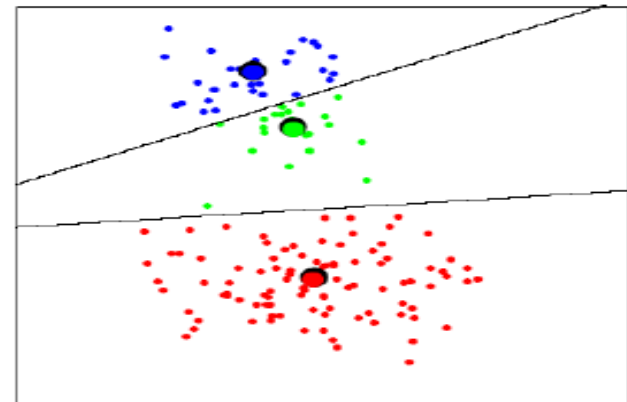
Initial Partition



Iteration Number 2



Iteration Number 20



# Distance Between Two Instances

## Euclidean Distance

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$





# Within-clusters Sum of Squares

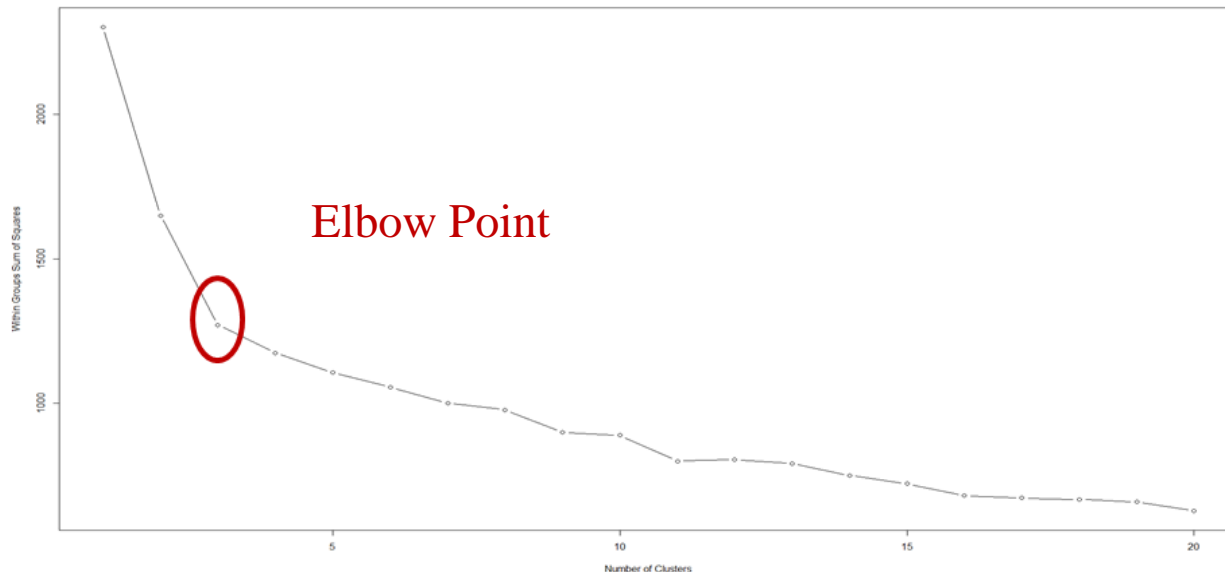
- Minimise the within-groups/clusters sum of squares (WSS)

$$WSS(k) = \sum_{i=1}^n \sum_{j=0}^p (x_{ij} - \text{mean}(x_{kj}))^2$$

where,  $k$  is the cluster,  $x_{ij}$  is the value of the  $j^{th}$  variable for the  $i^{th}$  observation, and  $\text{mean}(x_{kj})$  is the mean of the  $j^{th}$  variable for the  $k^{th}$  cluster.

# Choosing the Number of Clusters

- Elbow method
  - Gauge how the heterogeneity within clusters changes for various of  $k$ .
    - The heterogeneity within clusters is expected to decrease with more clusters.
    - The heterogeneity is measured by within-clusters/groups sum of squares (**WSS**)



# Pros and Cons

- Pros:
  - Low complexity
  - Performs well enough under many real-world use cases
- Cons:
  - Necessity of specifying  $k$
  - Clusters are sensitive to initial assignment of centroids
    - It is not guaranteed to find the optimal set of clusters
    - Clusters can be inconsistent from one run to another

# Validating Clusters

# Interpretation

**Goal:** obtain meaningful and useful clusters

**Caveats:**

(1) Random chance can often produce apparent clusters

(2) Different cluster methods produce different results

**Solutions:**

- Obtain summary statistics
- Also review clusters in terms of variables **not** used in clustering
- Label the cluster (e.g. “Gold Customers” ,” Silver Customers”, and “Bronze Customers”)

# Desirable Cluster Features

**Stability** – are clusters and cluster assignments sensitive to slight changes in inputs? Are cluster assignments in partition B similar to partition A?

**Separation** – check ratio of between-cluster variation to within-cluster variation (higher is better)

# Summary

- Cluster analysis is an exploratory tool. Useful only when it produces **meaningful** clusters
- **Hierarchical** clustering gives visual representation of different levels of clustering
  - On other hand, due to non-iterative nature, it can be unstable, can vary highly depending on settings, and is computationally expensive
- **K-means** clustering is computationally cheap and more stable; requires user to define  $k$
- Can use both methods
- Be wary of chance results; data may not have definitive “real” clusters