Cluster Analysis I

An Unsupervised Learning Technique

Presented by Klaus Ackermann

What is Cluster Analysis

Cluster: collection of data objects

- Grouping similar objects in a cluster
- No pre-defined classes → unsupervised classification
- For analysis of data

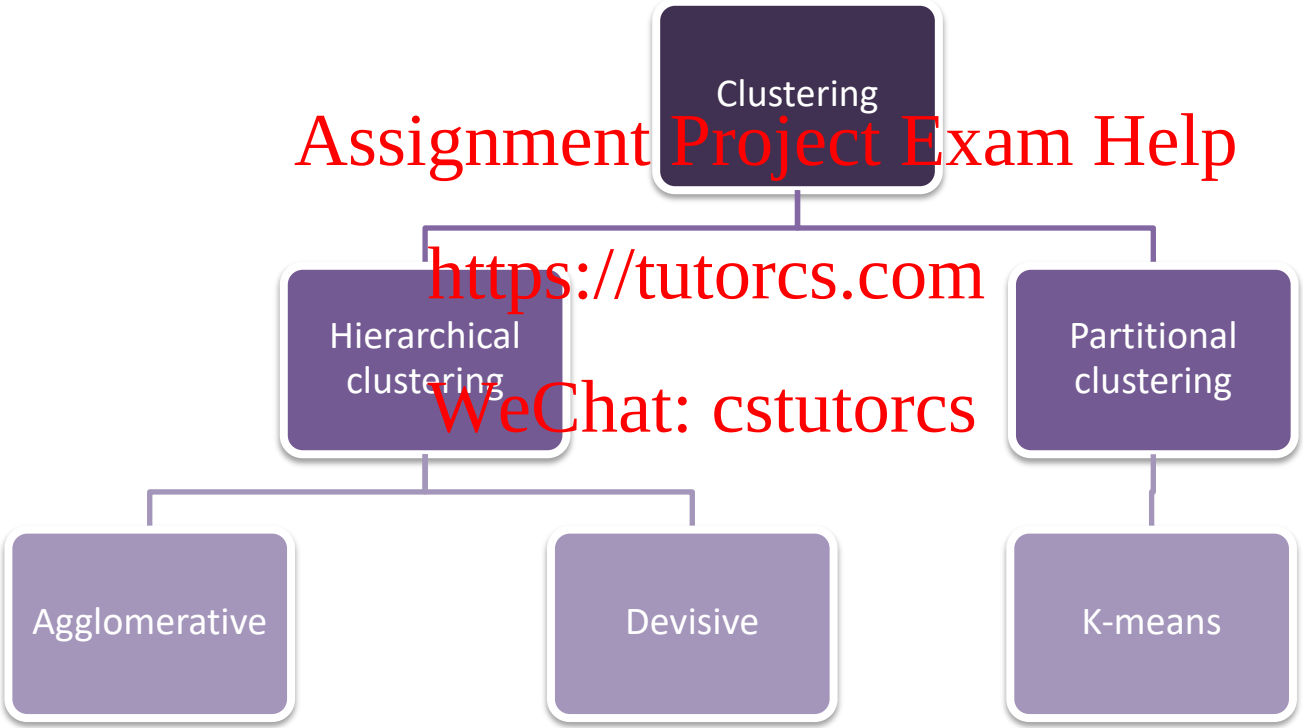
- Cluster analysis – grouping a set of data objects into cluster

- In what contexts would you want to complete a cluster analysis?

Cluster Analysis Applications

- Market segmentation – targeted marketing
- Market structure analysis – identify groups of similar products according to competitive measures of similarity
- Finance – develop a balanced portfolio
  - eg. Volatility, beta, vega, returns over different periods
- Land use – identify are of similar land use in Earth observation database
- Medical
  - eg. Types of tumour

Clustering Methods



Clustering Methods

Agglomerative

1. Group each observation into cluster of one
2. Merge most similar clusters
3. Merge those a but further apart
4. Sequence of nested clusters

or

Devisive

1. Observation in one cluster
2. Recursively split into smaller clusters

Clustering Methods

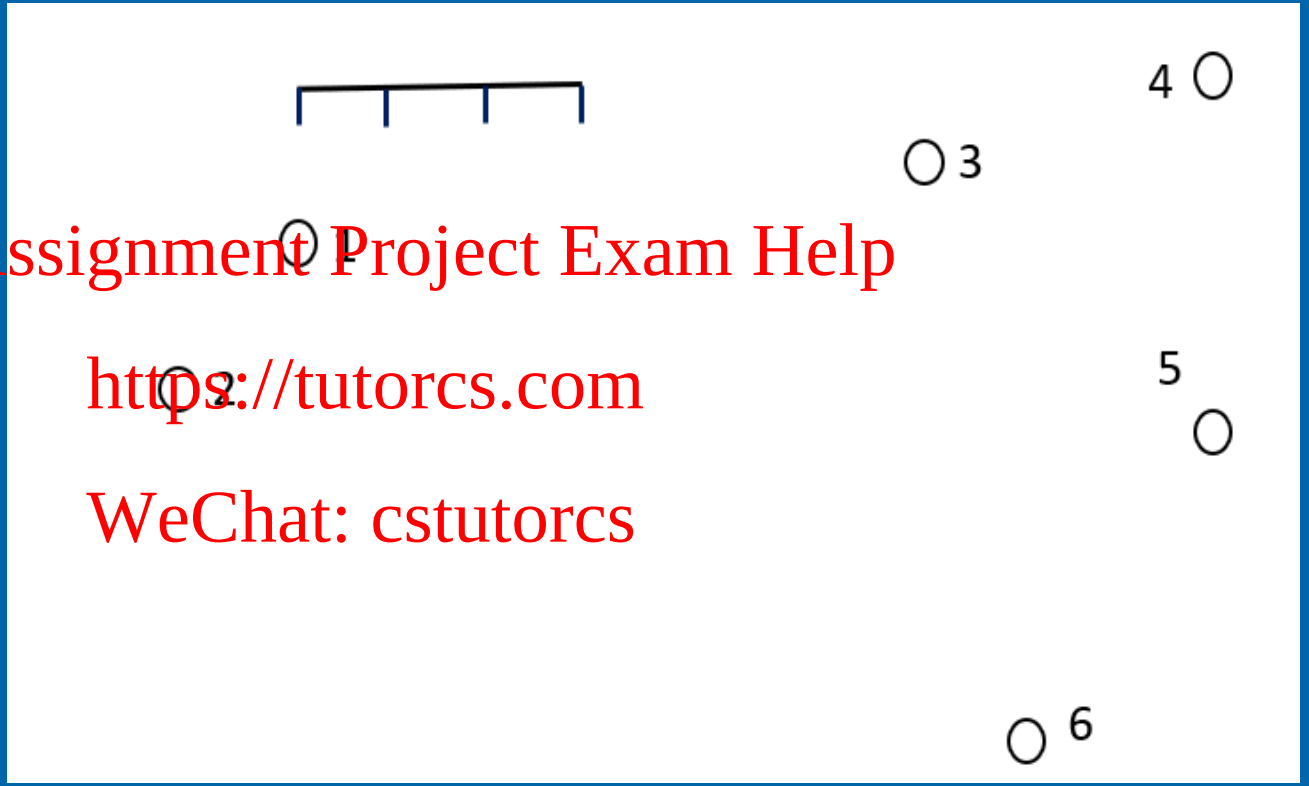
Partitional clustering

K-Means

1. Random assignment of points in the midst of cluster
2. Repeatedly re-assign points cluster
3. Minimise within-cluster distances
4. Maximise between-cluster distances

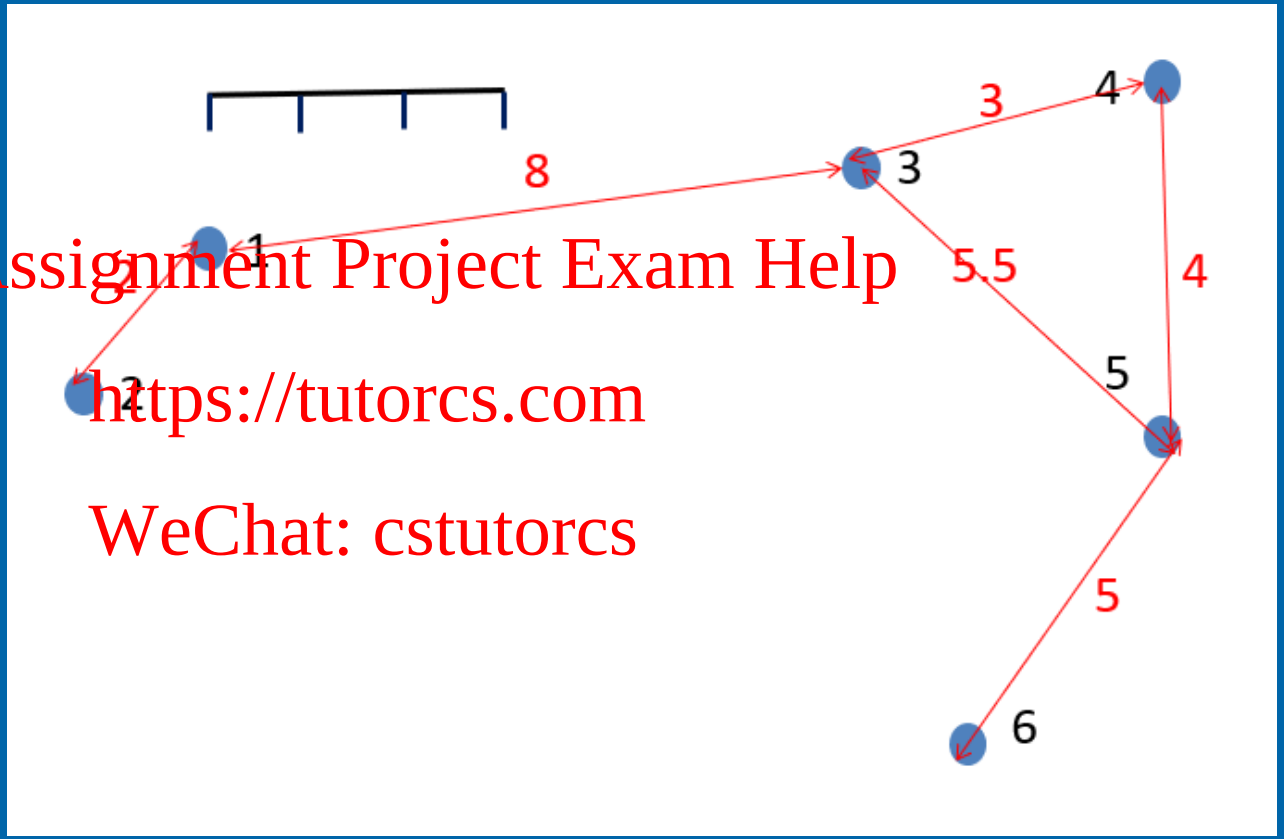
MONASH University

Distances between points in red; Names of points in black



MONASH University

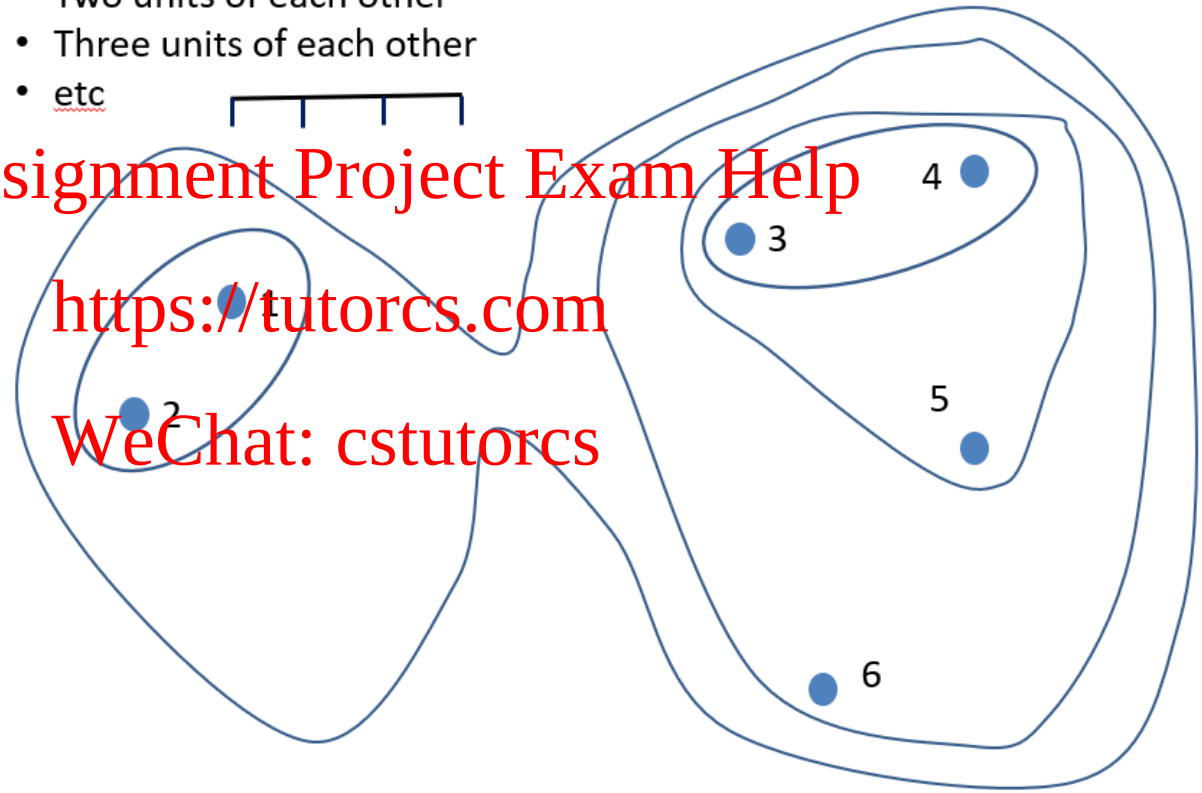
Distances between points in red; Names of points in black



MONASH University

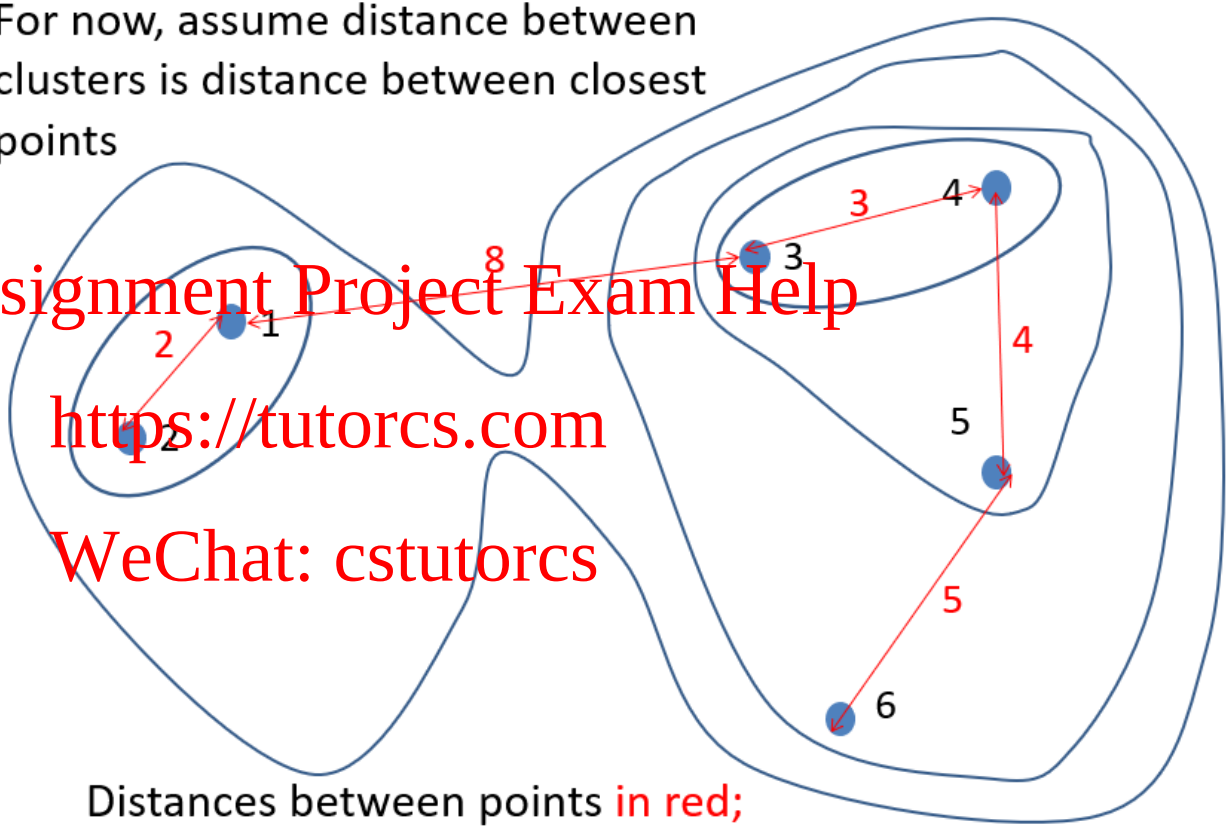
Hierarchical clustering: Suppose in order to be in a cluster, points have to be within:

- Two units of each other
- Three units of each other
- etc



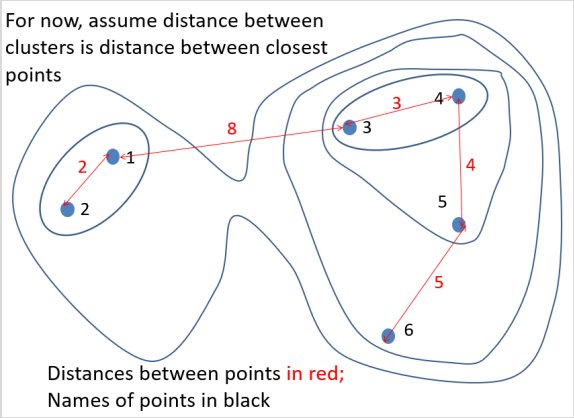
MONASH University

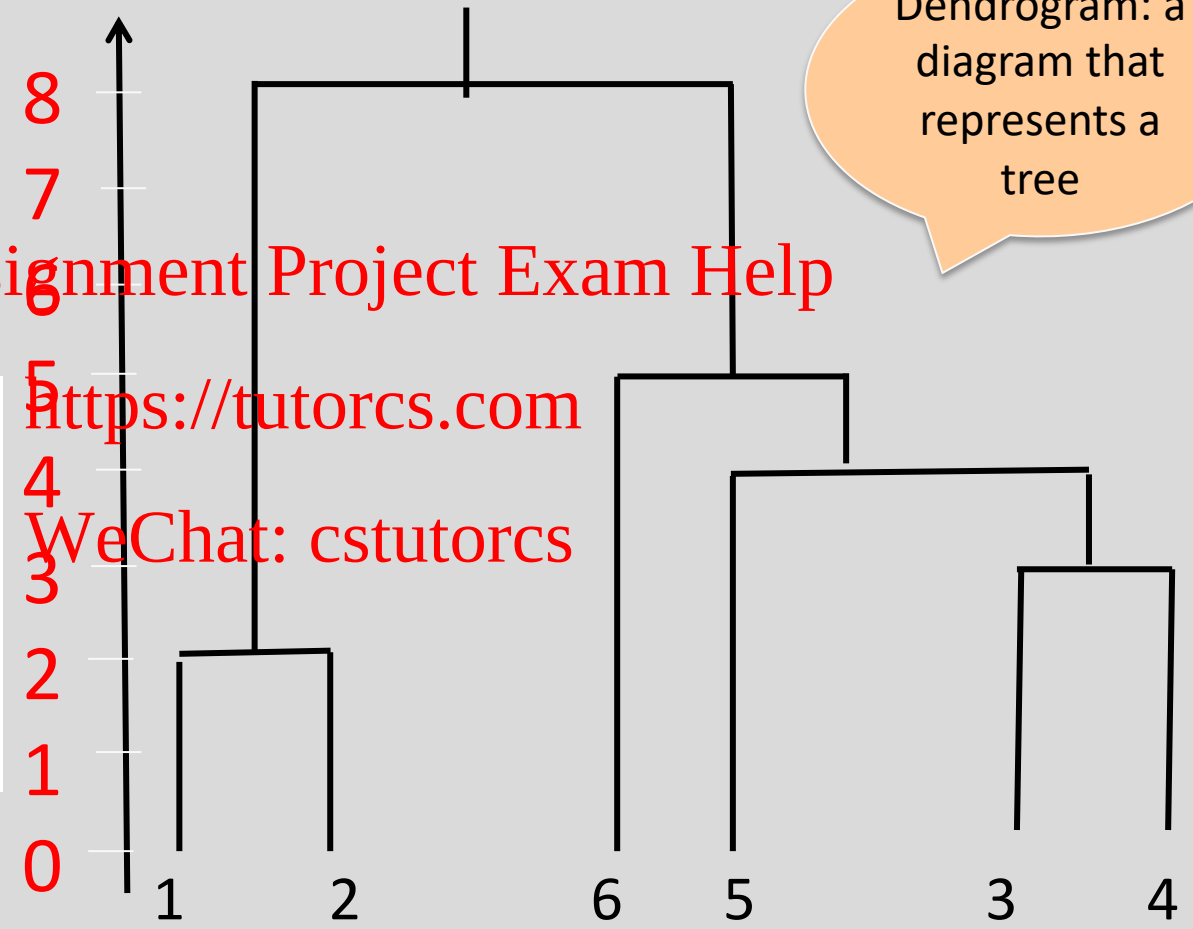
For now, assume distance between clusters is distance between closest points



MONASH University

Dendrogram: a diagram that represents a tree

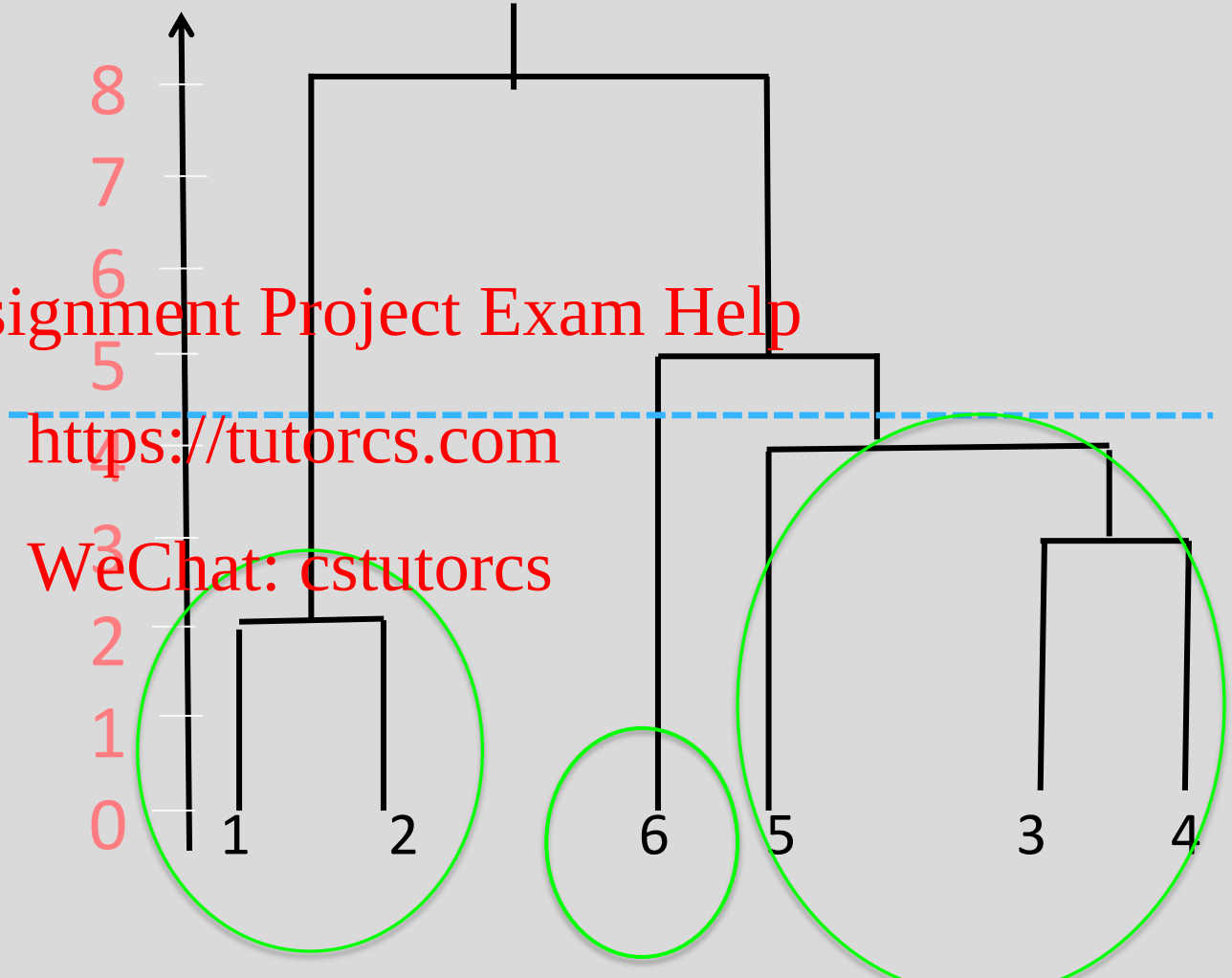




MONASH University

If this distance is just above 4, we create three clusters.

Name the members of the clusters



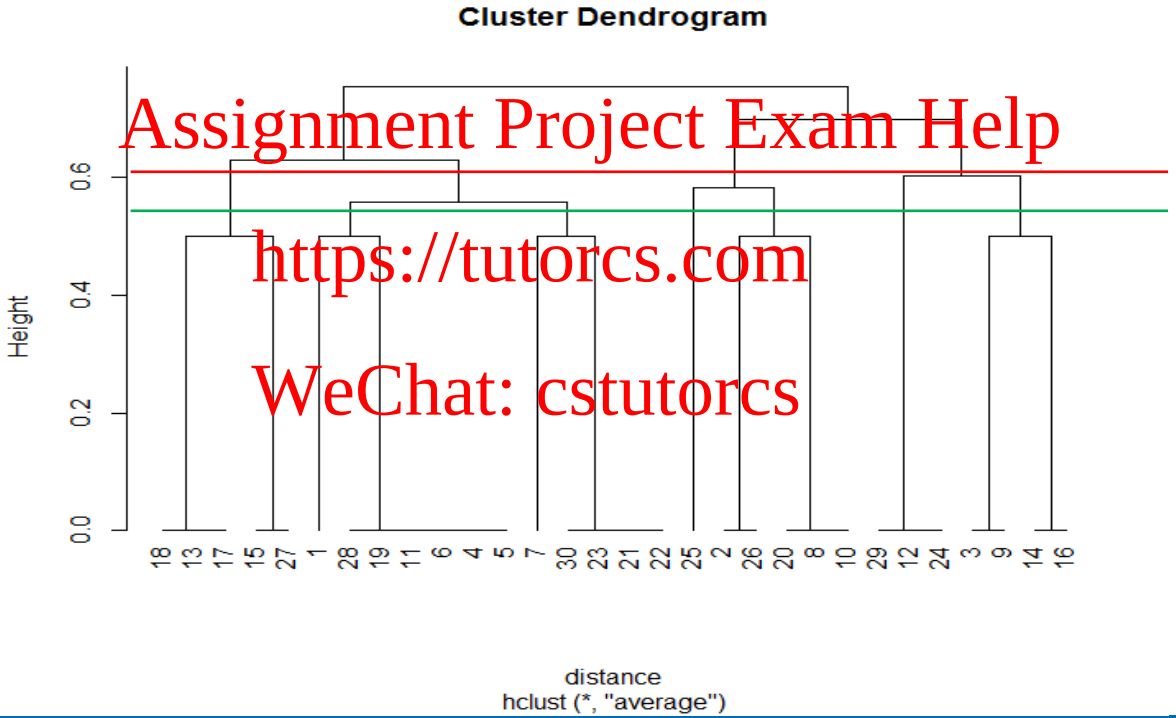
Dendrogram

- Visually, see what the possible numbers of clusters are by introducing a horizontal line

- **Exercise**: Based on the dendrogram on the next slide, if there are to be 7 clusters, list the ID numbers corresponding to the cluster containing the ID 10

- In the Hierarchical clustering workshop, we use R to select the clusters and then investigate cluster properties

Dendrogram 4 and 7 clusters



Distance

We need to decide

- How to define the distance between two data points
  - Recall that the data may be of different types

- How to define the distance between two clusters
  - This will depend on the distance between two points, but once that distance is decided there are still many alternative ways of defining distance between clusters

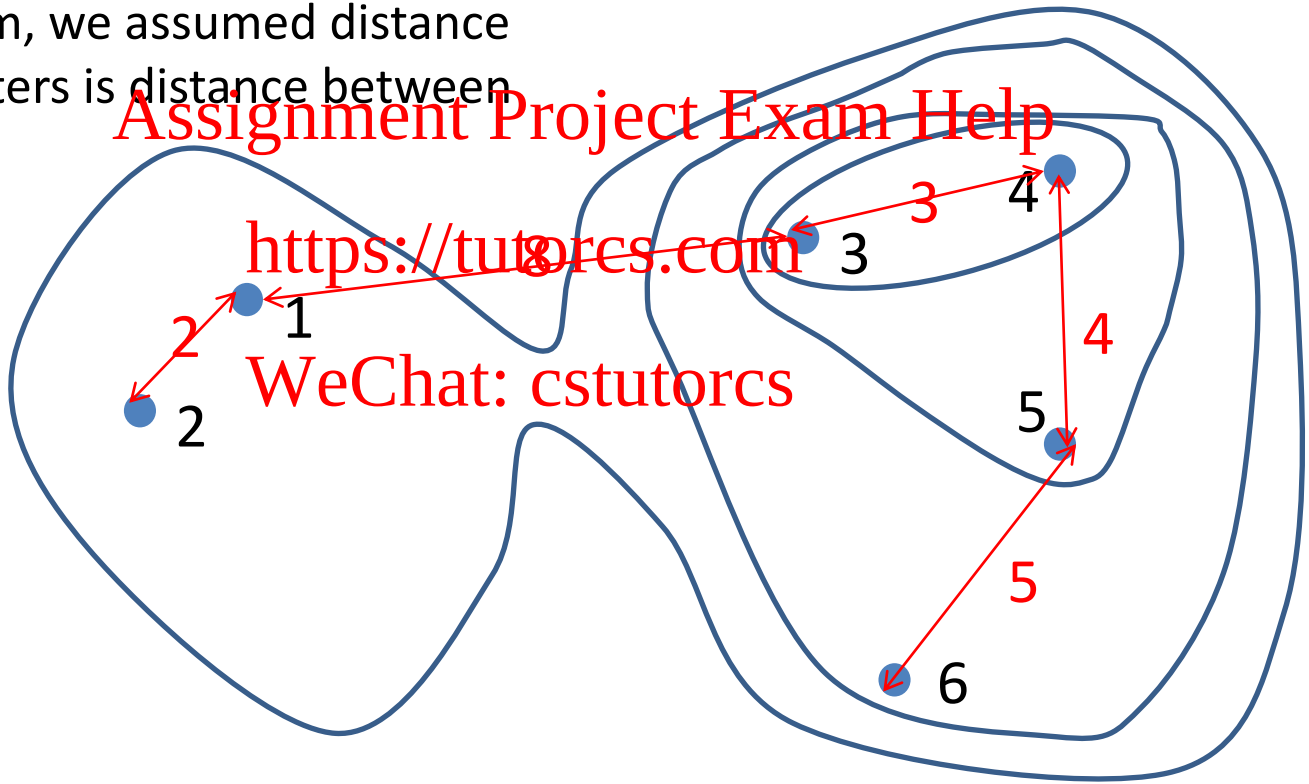
Distance

- In the example, we used
  - Euclidean distance (the usual shortest distance between two points)
  - between-cluster distance was taken to be the distance between the two points, one in each cluster, closest together

- When you have a data frame whose columns represent many different kinds of variable, defining distance between data points takes some thought.

Distance

In this diagram, we assumed distance between clusters is distance between closest points



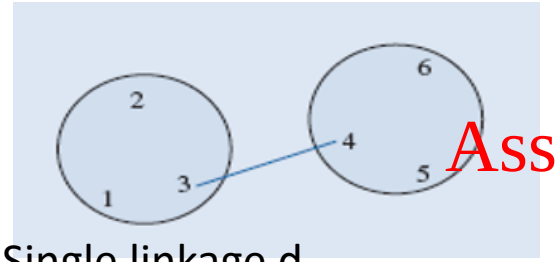
Distance

- Need to decide:
  - What are “distances” between points (=cases=rows)?
- And for the hierarchical method, in addition
  - What are “distances” between clusters?

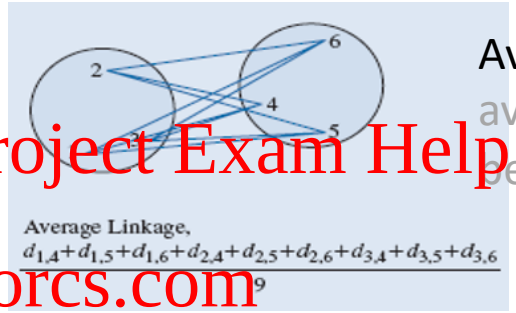
- *Distance* is a dissimilarity measure (larger distance, less similar)
- For categorical variables, see how many attributes match: *Matching* is a similarity measure (more matching, more similar)
- Where a similarity measure s has been used, a corresponding “distance” d is defined to feed into the cluster algorithm:

$$d = \sqrt{1 - s}$$

Distance between Clusters



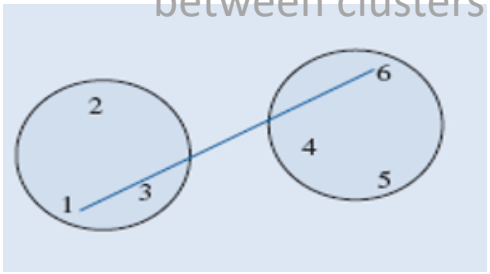
Single linkage $d_{3,4}$
shortest distance
between clusters



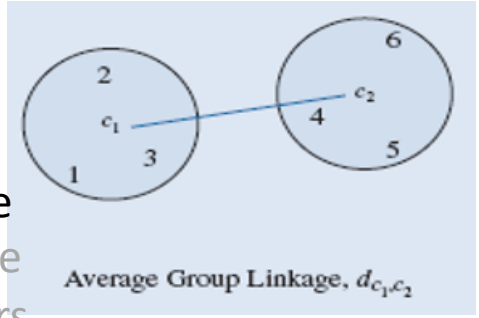
Average linkage
average distance
between all linkage

$$d = \sqrt{1-s}$$

Complete linkage $d_{1,6}$
longest distance
between clusters



Centroid linkage
average distance
between clusters



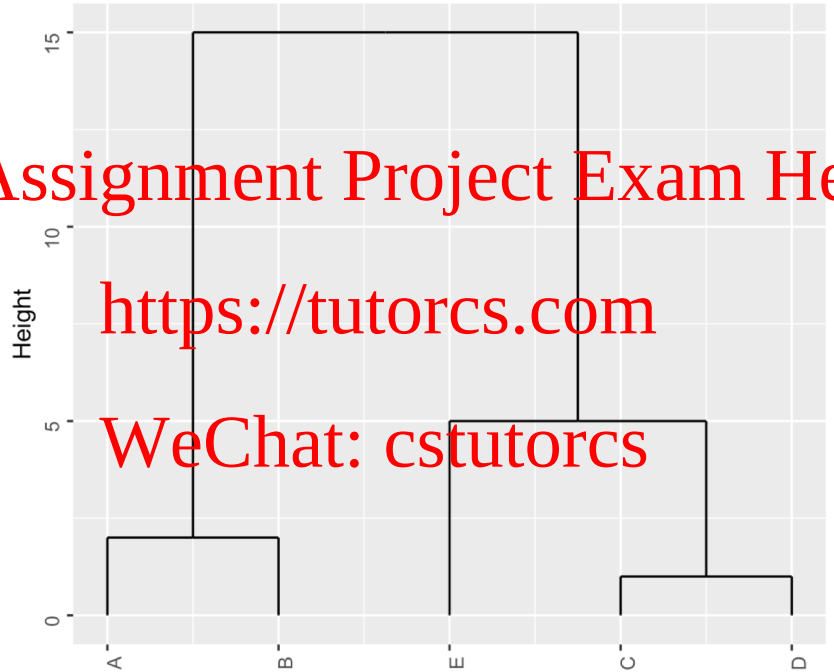
Distance Matrix to Dendrogram

	A	B	C	D	E
A	0	2	8	10	13
B	2	0	10	15	14
C	8	10	0	1	5
D	10	15	1	0	4
E	13	14	5	4	0

Given these distances, create a dendrogram for

1. complete linkage
2. single linkage

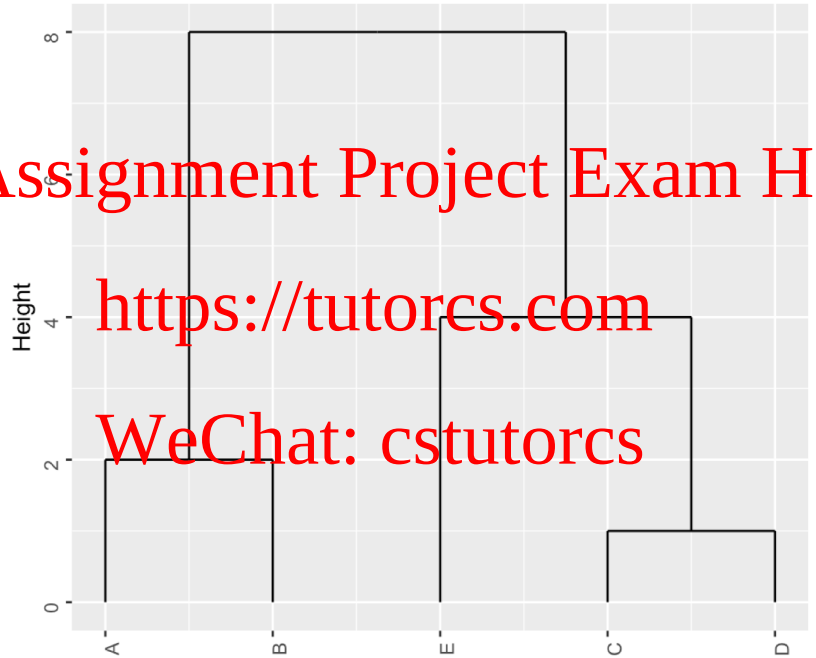
Complete Linkage



Complete Linkage – Breaking Down Steps

- Let's start with the smallest distance:
  1. C-D (dist 1): This is first cluster, recompute the similarities by considering C and D as one observation = 1
  2. A-B (dist 2): is the second cluster, recompute the similarities = 2
- For the next step, look for the larger distance between D-E or C-E. Which is longer? (D-E = 4), (C-D = 5). For complete linkage, look for maximum distance between cluster
  3. A new cluster C-D-E = 5
- Now look for maximum distance between A-B and C-D-E
  4. A-B to C-D-E = 15

Single Linkage



Single Linkage – Breaking Down Steps

- Let's start with the smallest distance:
  1. C-D (dist 1): This is first cluster, recompute the similarities by considering C and D as one observation = 1
  2. A-B (dist 2): is the second cluster, recompute the similarities = 2
- For the next step, look for the shorter distance between D-E or C-E. Which is shorter? (D-E = 4), (C-D = 5). For single linkage, look for minimum distance between cluster
  3. A new cluster C-D-E = 4
- Now look for minimum distance between A-B and C-D-E
  4. A-B to C-D-E = 8

Centroid

- A centroid is the average observation of a cluster

- That is, the average value for each variable across the points in the cluster

- The centroid doesn't have to be an actual observation

Hierarchical Cluster: Pros & Cons

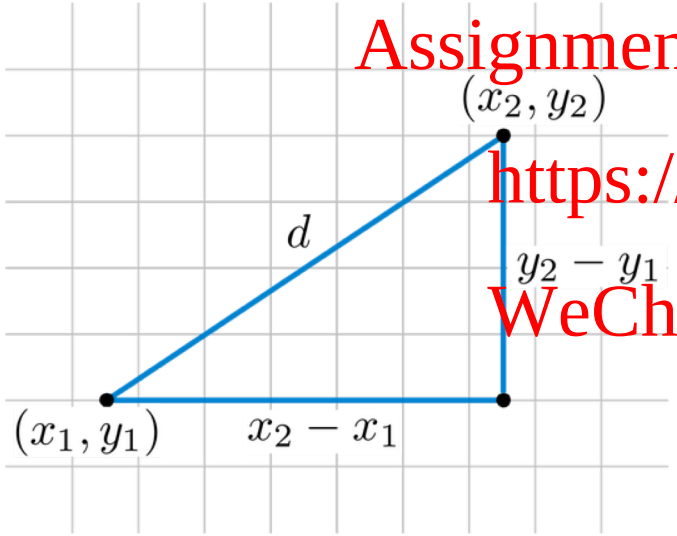
- Hierarchical clustering methods differ in the method of representing similarity between clusters, each with advantages and disadvantages

Single linkge	Complete linkage	Average linkage	Centroid linkage
• Most versatile algorithm • Poorly delineated cluster structures within data can produce unacceptable snakelike “chairs” for clusters	• Eliminates the chaining problem • Only considers outermost observations in a cluster → impacted by outliers	• Based on average similarity of all individuals in a cluster and tends to generate clusters with small within-cluster variation • less affected by outliers	• Measures distance between cluster centroids • Less affected by outliers

Numerical Distance

- For numerical variables, dissimilarity is measured by:
  - Euclidean distance
  - Manhattan distance
  - Many others

Numerical Distance



Euclidean distance becomes smaller as observations become more similar with respect to their variable values. Same is true for Manhattan distance

Euclidean distance

$$d_E = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Manhattan distance

$$d_M = |x_2 - x_1| + |y_2 - y_1|$$

Euclidean Distance

Euclidean distance: Most common method to measure **distance** between observations, when observations include continuous variables.
Let observations $u = (Age_u, Income_u, Kids_u)$ and

$v = (Age_v, Income_v, Kids_v)$

each comprise measurements of 3 variables.

The Euclidean distance between observations u and v is

$$d_{u,v} = \sqrt{(Age_u - Age_v)^2 + (Income_u - Income_v)^2 + (Kids_u - Kids_v)^2}$$

Manhattan Distance

Manhattan distance: **Distance** between observations that include continuous variables.

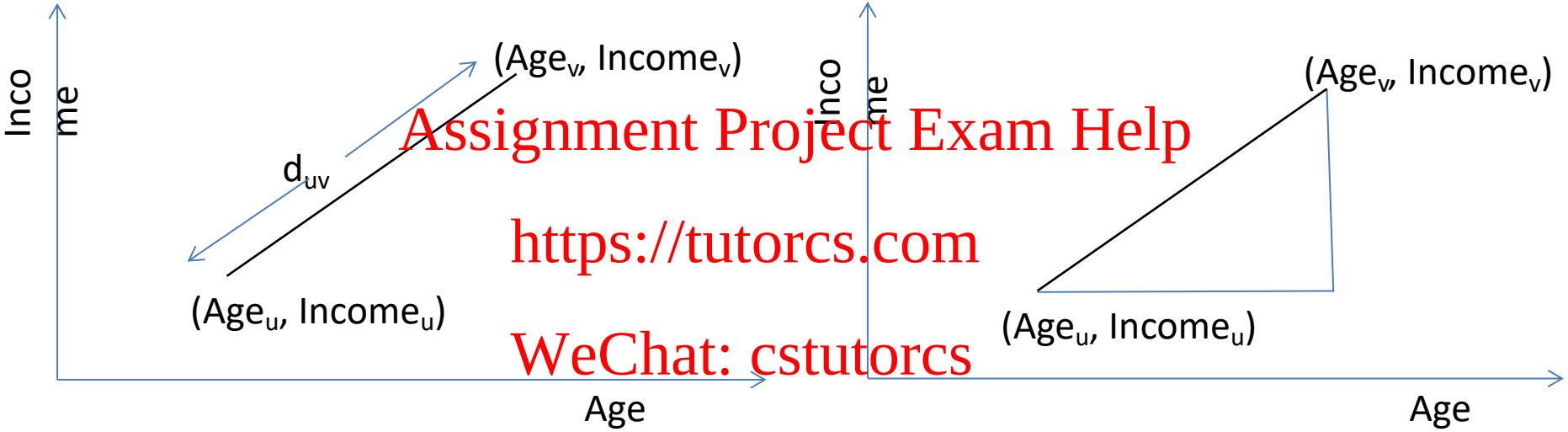
- Let observations *u* and *v* each comprise measurements of the variables *Age*, *Income*, and *Kids*.
- The Manhattan distance between observations *u* and *v* is

$$d_{u,v} = |Age_u - Age_v| + |Income_u - Income_v| + \cdots + |Kids_u - Kids_v|$$

Also need to standardise variables for Manhattan distance

Manhattan distance is more robust to outliers

Examples with Two Variables



Euclidean distance becomes smaller as observations become more similar with respect to their variable values. Same is true for Manhattan distance

Example: Know Thy Customer

- KTC is a Financial Advising Company
- Wishes to perform market segmentation of its clients. Thus each client is a case or "point"
- Variables: *Age, Gender, Annual Income, Marital Status, Number of Children, Car Loan or not, Home Mortgage or not*
  - *What type is each variable?*
  - *How should similarity between cases be measured in relation to values of*
    - *Numerical variables?*
    - *Categorical variables?*

Example: Know Thy Customer

Corresponding to each customer, KTC has a vector of measurements on seven customer variables: Age, Female, Income, Married, Children, Car Loan, Mortgage

Age	Female	Income	Married	Children	Car Loan	Mortgage
…	…	…	…	…	…	…
61	1	57881	1	2	0	1
27	0	55539	1	0	0	1

- 61-year-old female with an annual income of $57,881, married with two children, no car loan and but a mortgage.
- 27-year-old male with annual income of $55,539, married, no children, no car loan but mortgage

Start by considering only the binary variables: Female, Married, Car Loan, Mortgage

Measuring (Dis)Similarity Between *Observations*

- For binary variables, we discuss similarity measures by:
  - Matching coefficient
  - Jaccard's coefficient
- Many other measures are available

Similarity for Categorical Variables

Matching coefficient: Simplest overlap measure of **similarity** between two observations

- Find the proportion of variables with matching values
- Used when clustering observations solely on the basis of categorical variables encoded as 0–1
- Matching coefficient

$$s_M = \frac{\text{number of variables with matching value for observations } u \text{ and } v}{\text{total number of variables}}$$

Similarity for Categorical Variables

- Weakness of matching coefficient - If two observations both have a 0 entry for a categorical variable, this is counted as a sign of similarity between the two observations

- However, matching 0 entries do not necessarily imply similarity

  - If 1 means “Owns a Toyota Land Cruiser” and
    0 means “Does not own a Toyota Land Cruiser”
    then two observations sharing a zero does not indicate significant similarity

Similarity for Categorical Variables

Jaccard's coefficient: A similarity measure that does not count matching zero entries

$$s_J = \frac{\text{number of variables with matching } \textbf{\textit{nonzero}} \text{ value for observations } u \text{ and } v}{(\text{total number of variables}) - (\text{number of variables with matching zero values for observations } u \text{ and } v)}$$

Similarity for Categorical Variables

- For the following example, which coefficient should we use?
  - Matching?
  - Jaccard?

0 0 0 0 0 1 1 1 1 1

0 0 0 0 1 1 0 0 0 0

Similarity for Categorical Variables

	Female	Married	Car loan	Mortgage
U	1	1	0	1
V	0	1	0	1
W	1	1	1	1

Similarity between	Jaccard	Matching
U and V		
V and W		
W and U		

Similarity for Categorical Variables

- Binary variables in KTC data:
  - Gender
  - Married or not
  - Car loan or not
  - Home mortgage or not
- Given these variables, would you use Matching or Jaccard?

Example: Know Thy Customer

Again, corresponding to each customer, KTC has a vector of measurements on seven customer variables: Age, Female, Income, Married, Children, Car Loan, Mortgage

Age	Female	Income	Married	Children	Car Loan	Mortgage
…	…	…	…	…	…	…
61	1	57881	1	2	0	1
27	0	55539	1	0	0	1

Euclidean Distance

$$d_{1,2} = \sqrt{(61-27)^2 + (57881-55539)^2 + (2-0)^2}$$

$$= \sqrt{1156 + 5483964 + 4}$$

$$= \sqrt{5486124}$$

$$= 2342$$

What's the problem here?

ID	Age	Income	Children
	…	…	…
1	61	57881	2
2	27	55539	0

Manhattan Distance

$$d_{1,2(M)} = |61-27| + |57881-55539| + |2-0|$$
$$= 34 + 2342 + 2$$
$$= 2378$$

ID	Age	Income	Children
	…	…	…
1	61	57881	2
2	27	55539	0

The differences of scale are less, but the variables age and number of children are still swamped by Income

Scaling Distances

- Euclidean and Manhattan distance are highly influenced by the scale on which variables are measured.
  - Therefore, it is common to standardize the units of each variable of each observation *u*;
    - Example: $Income_u$, the value of variable Income in observation *u*, is replaced with its *z*-score.
- The conversion to *z*-scores also makes it easier to identify outlier measurements, which can distort the Euclidean distance between observations.

- Recall: What are z-scores?

Z-Scores

Consider values of a variable x $x_1, x_2, \quad , x_n$

Suppose the mean and standard deviation of this set of value is given by $\overline{x}$ and s

Then the z- score corresponding to x_i is given by $z_i = \dfrac{x_i - \overline{x}}{s}$

If certain variables are known to be more important, deliberately choose unequal weighting

Variables of Mixed Types

- A database may contain all types of variables
- A weighted formula is used to combine the effects of numerical and categorical variables
  - [See the *gower* metric in the command *daisy* in the package *cluster*]
- We focus on cluster analysis using just one type of variable

Limitations of Hierarchical Clustering

- Need to calculate and store an $n \times n$ distance matrix; infeasible for large datasets
- Low stability: change a few records and a very different result may be obtained
- Influenced by choice of metric, especially when using “average”
- Sensitive to outliers
- Only one pass through the data: misallocations can never be fixed

Interpreting, Profiling, & Validating Clusters

- The clusters centroid, a mean profile of the cluster on each clustering variable, is particularly useful in the interpretation stage
  - Interpretation involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters
  - If there is not substantial difference between clusters, then other cluster solutions should be examined
  - The cluster centroid should also be assessed for correspondence with the researcher's prior expectations based on theory or practical experience

Interpreting, Profiling, & Validating Clusters

- Validation is essential in cluster analysis since the clusters are descriptive of structure and require additional support for their relevance:
  - Cross-validation empirically validates a cluster solution by creating two sub-samples (randomly splitting the sample) and then comparing the two cluster solutions for consistency with respect to number of clusters and the cluster profiles.
  - Validation is also achieved by examining differences on variables not included in the cluster analysis but for which there is a theoretical and relevant reason to expect variation across the clusters

MONASH University

MONASH BUSINESS SCHOOL