cct | College Dublin
Computing • IT • Business

Machine Learning for Data Analysis
MSc in Data Analytics
CCT College Dublin

Hierarchical Clustering
Week 12

Lecturer: Dr. Muhammad Iqbal*

Email: miqbal@cct.ie

Agenda

- Unsupervised Learning: Clustering Task
- Clustering Task: Similarity Measures
- Hierarchical Clustering Methods
- Divisive Methods
- Agglomerative Methods
- Single-Linkage Clustering
- Complete-Linkage Clustering
- Comparison of both Linkage methods
- Hierarchical Clustering: Dendrogram
- Single-Linkage Clustering: Example
- Cluster Merging using Dendrograms: Stopping Criterion

Unsupervised Learning

Clustering Task

- *Clustering* refers to grouping records, observations, or tasks into classes of similar objects
- Cluster is collection records similar to one another
- Records in one cluster dissimilar to records in other clusters
- Clustering is unsupervised data mining task
- **Therefore, no target variable specified**
- Clustering algorithms segment records and maximize homogeneity in subgroups
- Similarity to records outside cluster minimized
- *Applying cluster analysis to enormous databases helpful*
- *Reduces search space for downstream algorithms*

Clustering Task

Clustering Tasks in Business and Research

- Target marketing for niche product, without large marketing budget
- Accounting auditing: Segment behavior into benign and suspicious categories
- As a dimension-reduction tool when data set has hundreds of attributes
- Gene expression clustering, where genes exhibit similar characteristics
- Clustering often performed as preliminary step in data mining process
- Clustering results used as input to other data mining techniques
    - *Cluster analysis addresses similar issues encountered in classification*
    - *Similarity measurement*
    - *Recoding categorical variables*
    - *Standardizing and normalizing variables*
    - *Number of clusters*

Clustering Task

Measuring Similarity

- Euclidean Distance measures distance between records

where

$$d_{\text{Euclidean}}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

$$\mathbf{x} = x_1, x_2, \ldots, x_m \text{ and } \mathbf{y} = y_1, y_2, \ldots, y_m$$

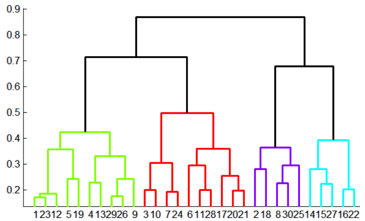
represent the m attribute value of two records

- Other distance measurements include City-Block Distance and Minkowski Distance

$$d_{\text{City-Block}}(\mathbf{x},\mathbf{y}) = \sum_i \left| x_i - y_i \right|$$

$$d_{\text{Minkowski}}(\mathbf{x},\mathbf{y}) = \sum_i \left| x_i - y_i \right|^q$$

Hierarchical Clustering Methods



- Clustering algorithms either ***<u>Hierarchical</u>*** or ***<u>Non-Hierarchical</u>***

- **Hierarchical**

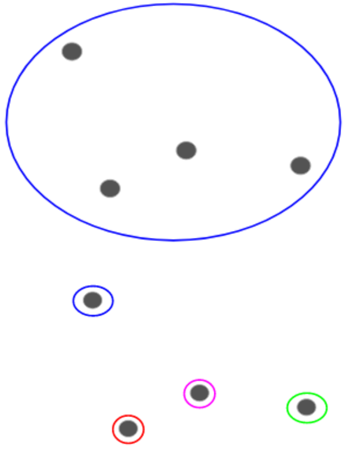
    - Treelike cluster structure (dendogram) created through recursive partitioning (Divisive Methods) or combining (Agglomerative Methods) existing clusters

    - **Divisive Methods**

    - All records initialized into single cluster

    - At each iteration, most dissimilar record split off into separate cluster

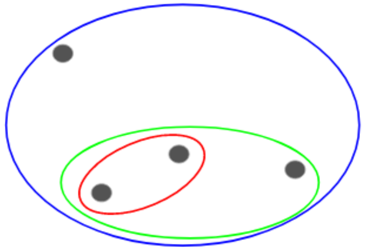
    - Continues until each record represents single cluster



Divisive Clustering

Hierarchical Clustering Methods

- **Agglomerative Methods**
  - Each observation initialized to become own cluster
  - At each iteration two closest clusters aggregated together
  - Number of clusters reduced by one, each step
  - Eventually, all records combined into single cluster
  - Agglomerative more popular hierarchical method
  - Therefore, focus remains on this approach
  - Measuring distance between records straightforward once recoding and normalization applied
  - However, ***how is distance between clusters determined?***



Agglomerative Clustering

Hierarchical Clustering Methods

- **Distance Between Clusters**
  - Several criteria examined to determine distance between clusters, **A** and **B**
  - **Single Linkage**
  - Known as ***Nearest-Neighbor Approach***
  - Minimum distance between any record in cluster **A**, and any record in cluster **B**
  - Cluster similarity based on most similar records from each cluster
  - Tends to form long, slender clusters
  - Sometime heterogeneous records clustered together

Hierarchical Clustering Methods

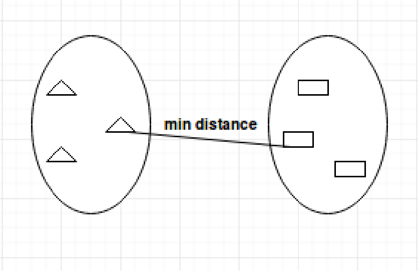
- Measure is average distance of records in cluster **A**, from records in cluster **B**
- Resulting clusters have approximately equal within-cluster variability

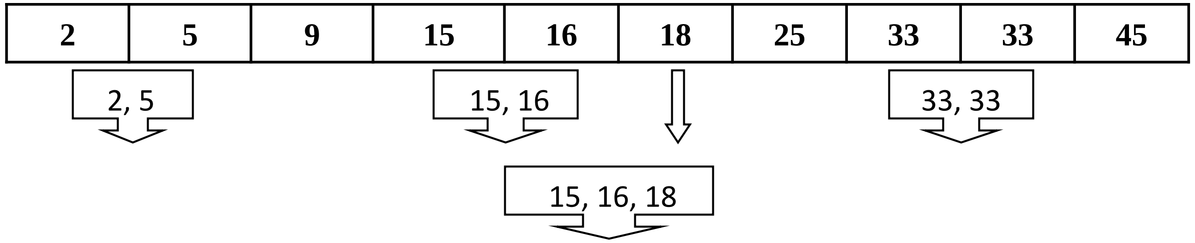
- Next, linkage methods examined using small data set

2 5 9 15 16 18 25 33 33 45

Single-Linkage Clustering

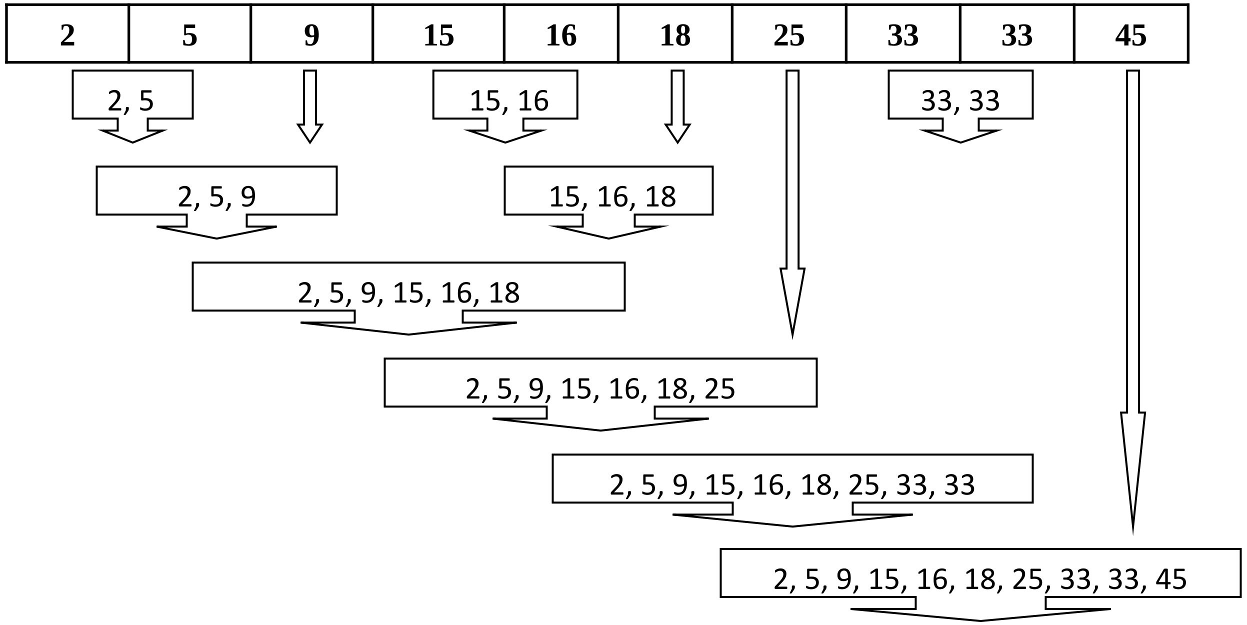
- To begin, each record assigned to own cluster
- Single-linkage seeks **minimum distance** between any two records, in separate clusters
- **Step 1:** Minimum cluster distance is between clusters **{33}** and **{33}**. Distance = 0, clusters combined
- **Step 2:** Clusters **{15}** and **{16}** combined, where distance = 1
- **Step 3:** Cluster **{15, 16}** combined with cluster **{18}**
- **Step 4:** Clusters **{2}** and **{5}** combined





2	5	9	15	16	18	25	33	33	45

Single-Linkage Clustering



2 5 9 15 16 18 25 33 33 45

2, 5

15, 16

33, 33

2, 5, 9

15, 16, 18

2, 5, 9, 15, 16, 18

2, 5, 9, 15, 16, 18, 25

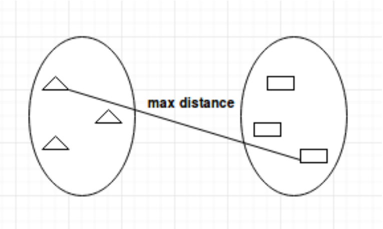
2, 5, 9, 15, 16, 18, 25, 33, 33

2, 5, 9, 15, 16, 18, 25, 33, 33, 45

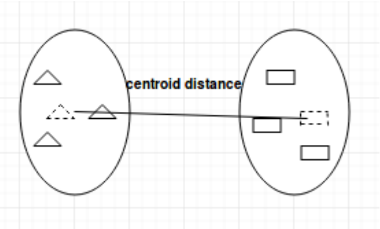
- **Agglomeration** continues similarly Steps 4 – 9
- Above, last cluster {2, 5, 9, 15, 16, 18, 25, 33, 33, 45} contains all records in data set

Hierarchical Clustering Methods

- **Complete Linkage**
- Known as ***Farthest-Neighbor Approach***
- Maximum distance between any record in cluster **A**, and any record in cluster **B**
- Cluster similarity based on most dissimilar records from each cluster
- Compact, sphere-like clusters formed
- All records in cluster within given diameter of other records



- **Centroid-linkage**
- It finds centroid of cluster **A** and centroid of cluster **B**, and then calculates the distance between the two before merging.



Complete-Linkage Clustering

- Complete-linkage explored using sample data

```
2   5     9     15     16     18     25     33     33     45
```

- Distance among records in two clusters farthest from each other minimized
- **Step 1:** Each cluster contains single record

  No difference between single and Complete-linkage

  Clusters **{33}** and **{33}** combined
- **Step 2:** Clusters **{15}** and **{16}** combined

  No difference between single and Complete-linkage

Complete-Linkage Clustering

- **Step 3:** Complete-linkage diverges from Single-linkage

  Farthest neighbors between **{15, 16}** and **{18}** are 15 and 18, distance = 3

  Clusters **{2}** and **{5}** also have distance = 3

  Algorithm silent regarding ties

  Result, **{2, 5}** arbitrarily chosen

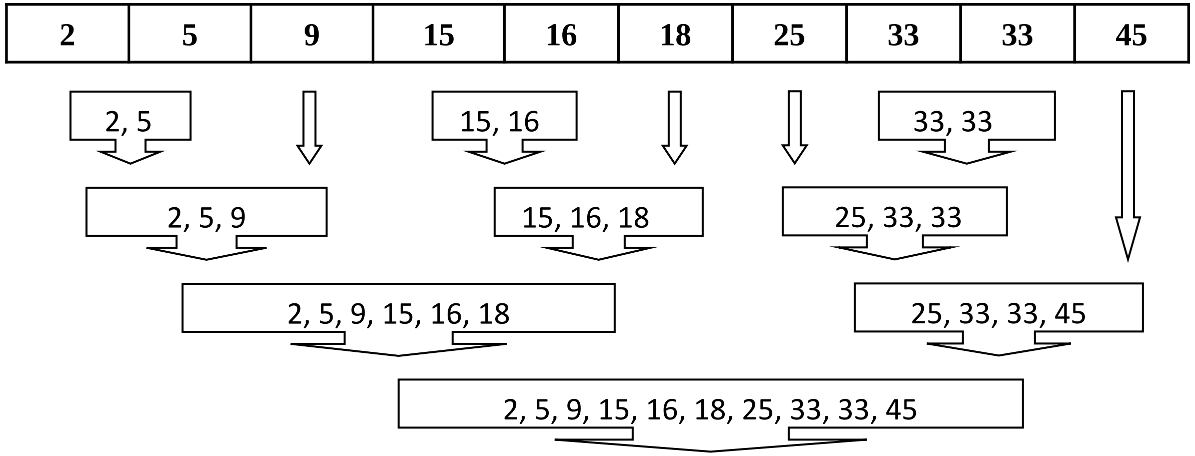
- Complete-linkage procedure continues for Steps 4 – 9, until all records contained in same cluster

Complete and Single-Linkage Clustering

Comparison

- Figure illustrates Complete-linkage agglomerative clustering on sample data

Complete-linkage



2	5	9	15	16	18	25	33	33	45

2, 5

15, 16

33, 33

2, 5, 9

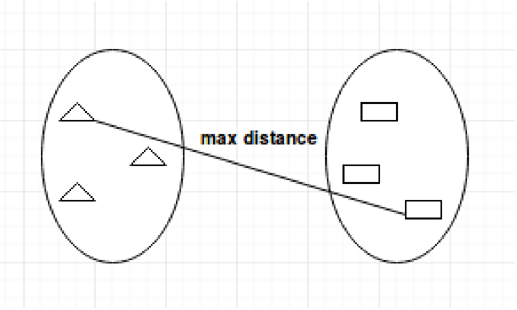
15, 16, 18

25, 33, 33

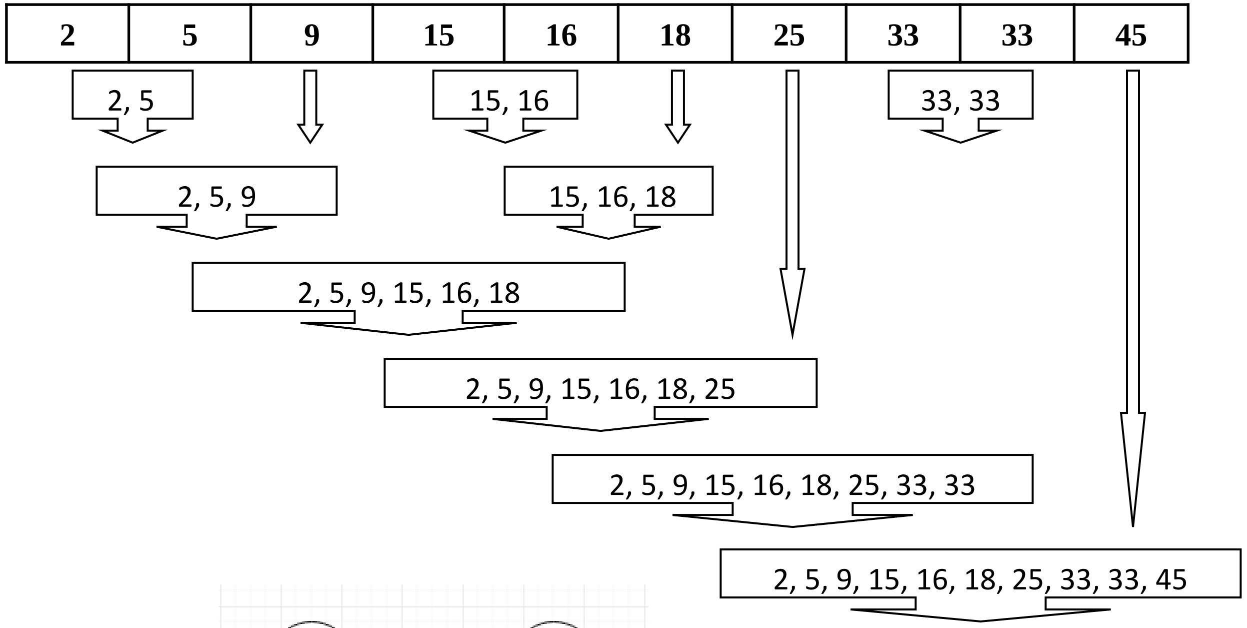
2, 5, 9, 15, 16, 18

25, 33, 33, 45

2, 5, 9, 15, 16, 18, 25, 33, 33, 45



Single-linkage



2	5	9	15	16	18	25	33	33	45

2, 5

15, 16

33, 33

2, 5, 9

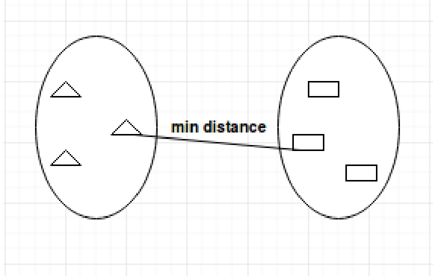
15, 16, 18

2, 5, 9, 15, 16, 18

2, 5, 9, 15, 16, 18, 25

2, 5, 9, 15, 16, 18, 25, 33, 33

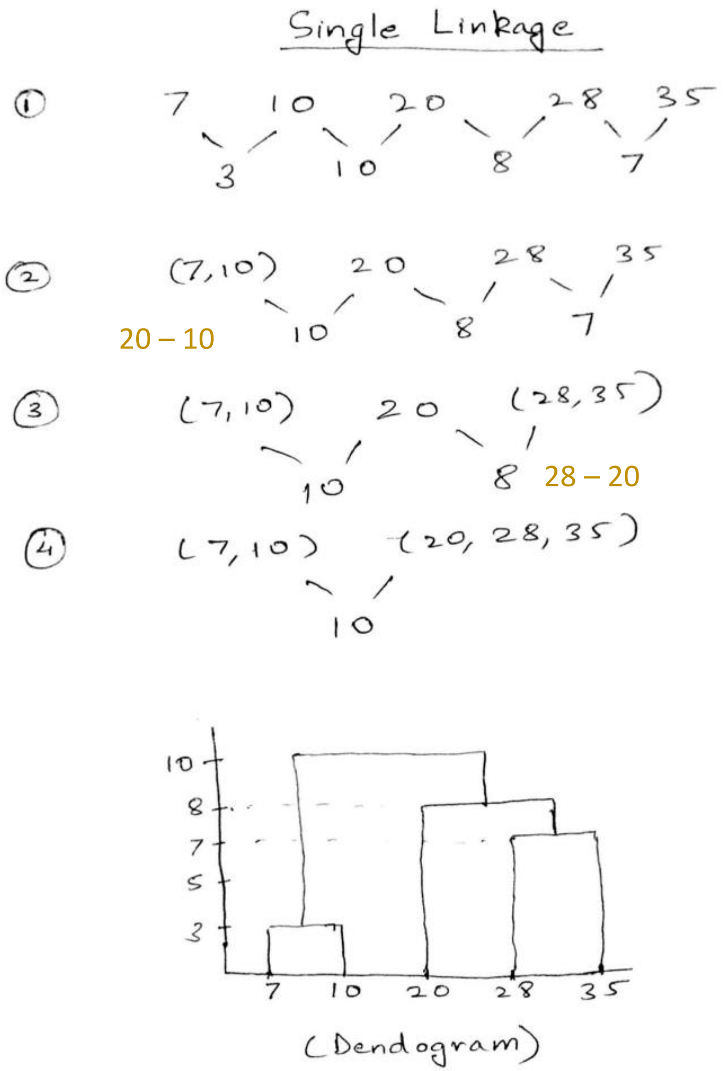
2, 5, 9, 15, 16, 18, 25, 33, 33, 45



Hierarchical Clustering

Dendrogram

- **For one dimensional data set {7, 10, 20, 28, 35}, perform hierarchical clustering and plot the dendrogram to visualize it.**
- We can solve this problem by hand using both the types of agglomerative hierarchical clustering (Single or Complete).
- **Single Linkage:** In single link hierarchical clustering, we merge two clusters in each step, whose two closest members have the smallest distance.
- Using single linkage, two clusters are formed
  - **Cluster 1:** (7, 10)
  - **Cluster 2:** (20, 28, 35)



Single Link Clustering

Problem: Assume that the database **D** is given by the table below. Follow single link technique to find clusters in **D**. Use Euclidean distance to measure.

Distance matrix

	x	y
p1	0.402	0.530
p2	0.220	0.380
p3	0.350	0.315
p4	0.260	0.189
p5	0.080	0.410
p6	0.450	0.300

Distance matrix

p1	0					
p2	0.24	0				
p3	0.22	0.15	0			
p4	0.37	0.20	0.15	0		
p5	0.34	0.14	0.28	0.29	0	
p6	0.23	0.24	0.11	0.22	0.39	0
	p1	p2	p3	p4	p5	p6

$d(p_1, p_2) = 0.235 = 0.24$ (rounded value)

$$d(i, j) = \text{sqrt}\,(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)$$

Single Link Clustering

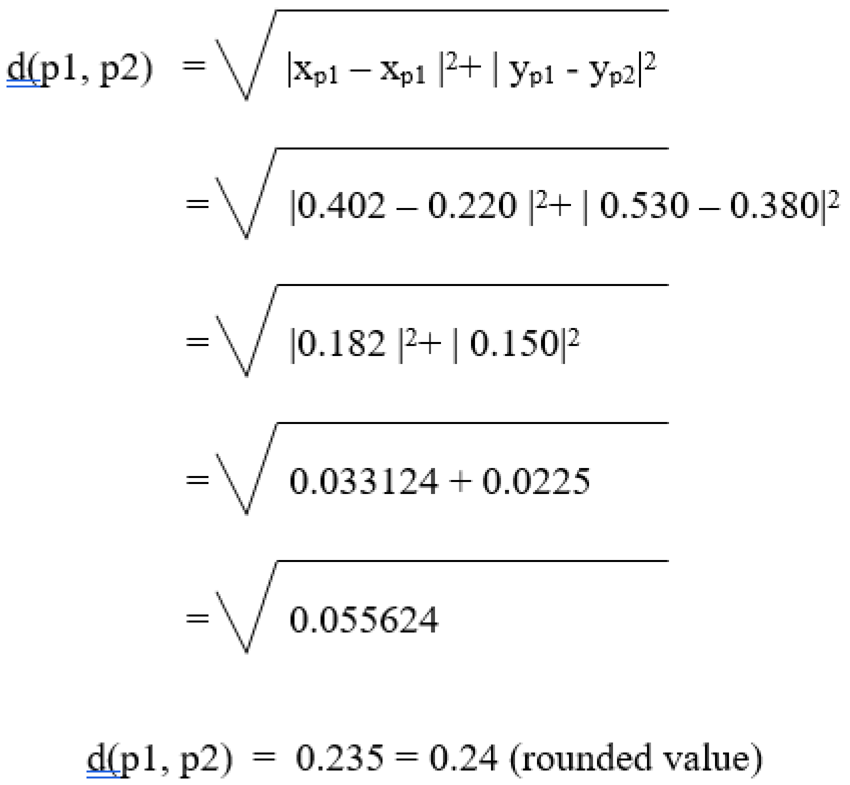
- Calculate the distance from each object (point) to all other points, using Euclidean distance measure, and place the numbers in a distance matrix.

$$d(i, j) = \text{sqrt}\,(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)$$

Distance matrix

p1	0					
p2	0.24	0				
p3	0.22	0.15	0			
p4	0.37	0.20	0.15	0		
p5	0.34	0.14	0.28	0.29	0	
p6	0.23	0.24	0.11	0.22	0.39	0
	p1	p2	p3	p4	p5	p6

$d(p_1, p_2) = 0.235 = 0.24$ (rounded value)



$$d(p1, p2) = \sqrt{|x_{p1} - x_{p1}|^2 + |y_{p1} - y_{p2}|^2}$$

$$= \sqrt{|0.402 - 0.220|^2 + |0.530 - 0.380|^2}$$

$$= \sqrt{|0.182|^2 + |0.150|^2}$$

$$= \sqrt{0.033124 + 0.0225}$$

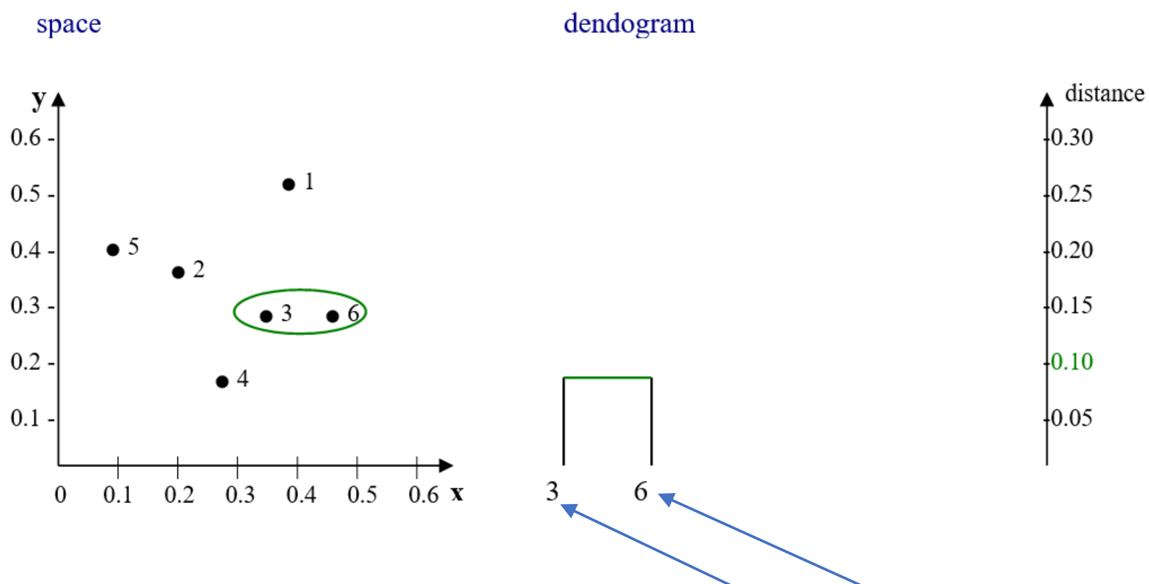
$$= \sqrt{0.055624}$$

$d(p1, p2) = 0.235 = 0.24$ (rounded value)

Single Link Clustering

p1	0					
p2	0.24	0				
p3	0.22	0.15	0			
p4	0.37	0.20	0.15	0		
p5	0.34	0.14	0.28	0.29	0	
p6	0.23	0.24	0.11	0.22	0.39	0
	p1	p2	p3	p4	p5	p6

p1	0				
p2	0.24	0			
(p3, p6)	0.22	0.15	0		
p4	0.37	0.20	0.15	0	
p5	0.34	0.14	0.28	0.29	0
	p1	p2	(p3, p6)	p4	p5

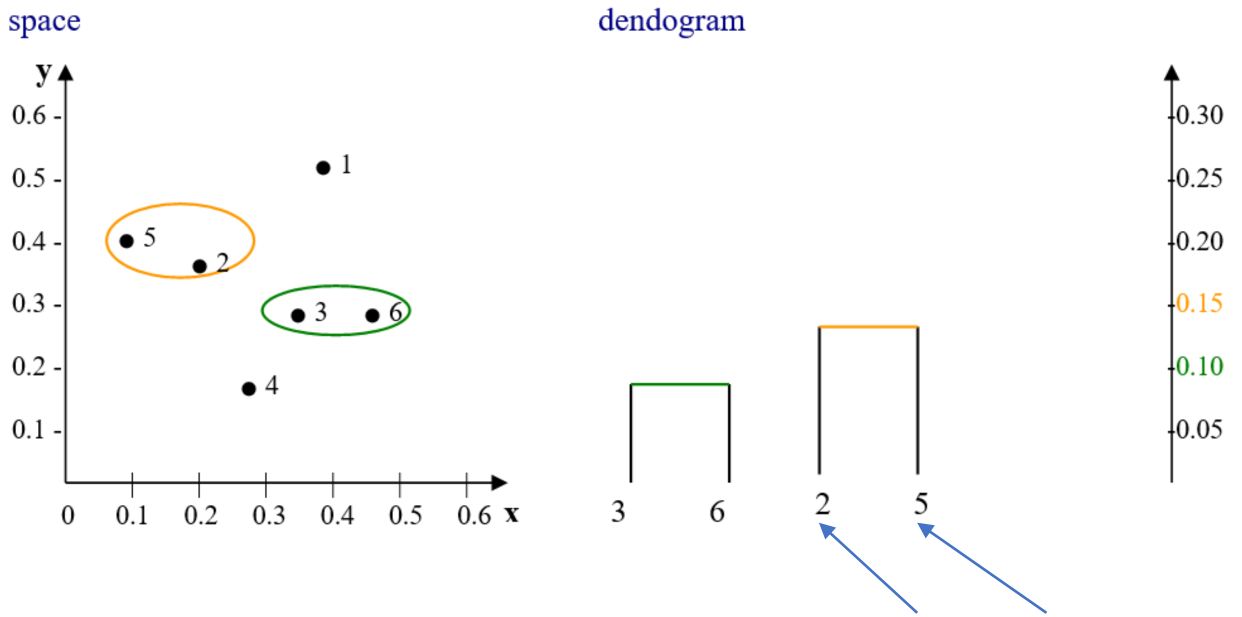


- The minimum distance between p3 and p6 is 0.11.
- First, merge these data points.

Single Link Clustering

p2	0.24	0			
(p3, p6)	0.22	0.15	0		
p4	0.37	0.20	0.15	0	
p5	0.34	0.14	0.28	0.29	0
	p1	**p2**	(p3, p6)	p4	**p5**

p1	0			
(p2, p5)	0.24	0		
(p3, p6)	0.22	0.15	0	
p4	0.37	0.20	0.15	0
	p1	**(p2, p5)**	**(p3, p6)**	p4

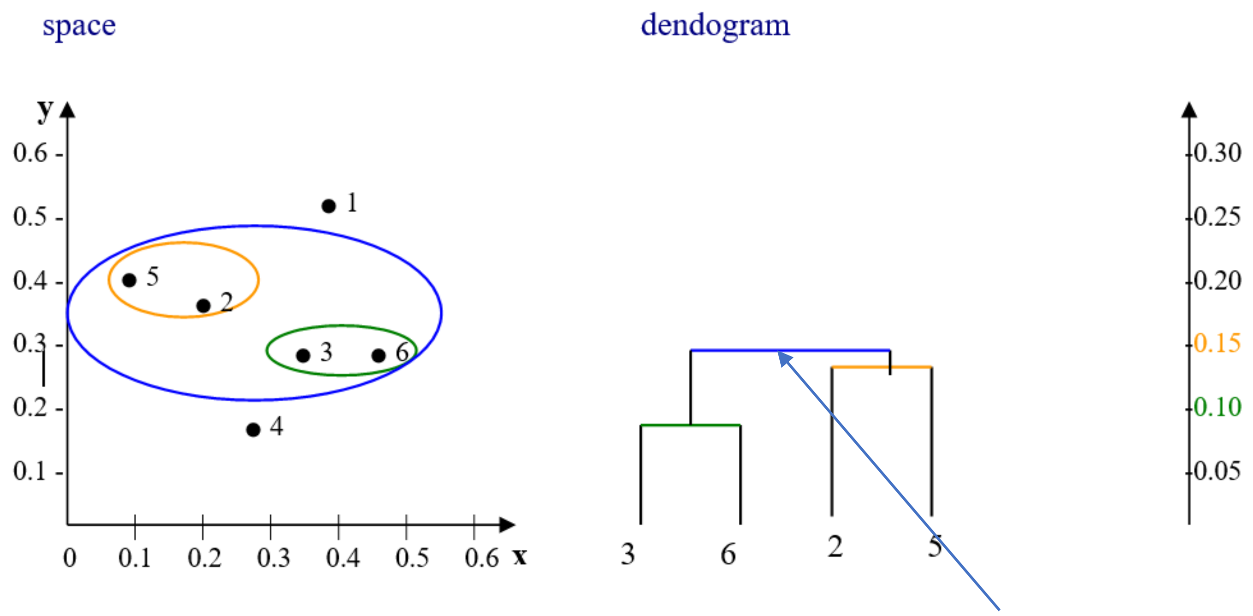


- Next, the minimum distance between p2 and p5 is 0.14.
- Second, merge these data points.

Single Link Clustering

p1	0			
(p2, p5)	0.24	0		
(p3, p6)	0.22	0.15	0	
p4	0.37	0.20	0.15	0
	p1	(p2, p5)	(p3, p6)	p4

p1	0		
(p2, p5, p3, p6)	0.22	0	
p4	0.37	0.15	0
	p1	(p2, p5, p3, p6)	p4

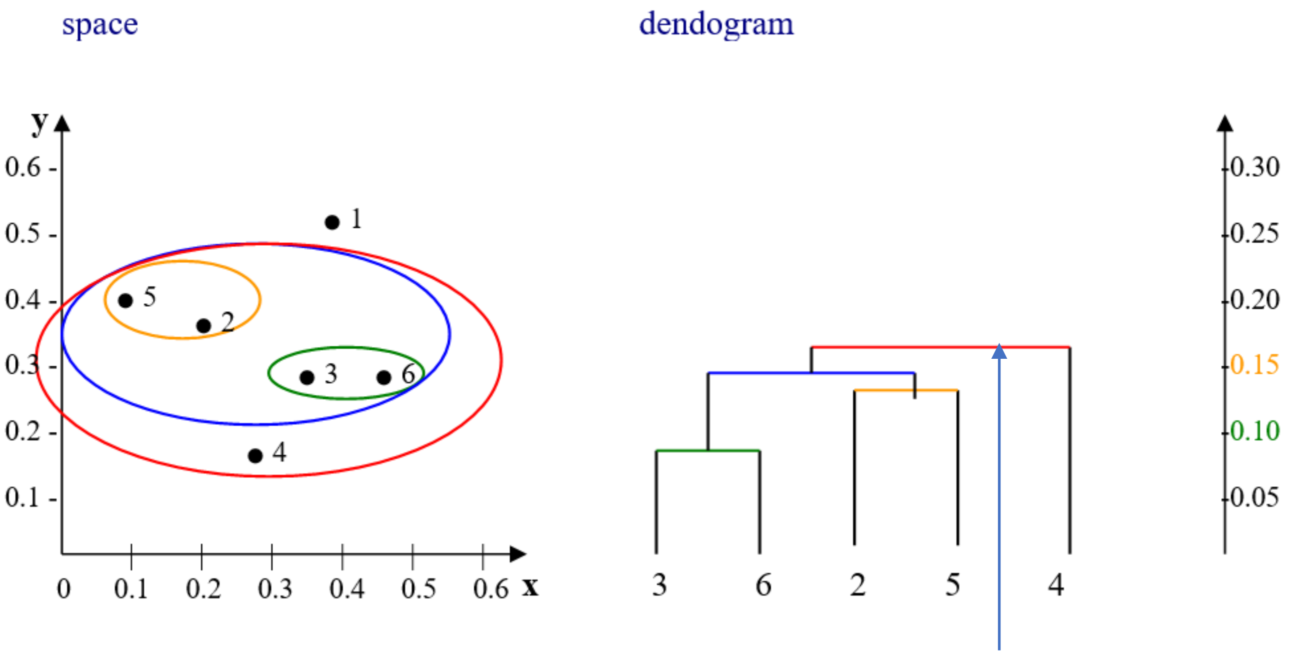


- Next, the minimum distance between (p2, p5) and (p3, p6) is 0.15.
- Third, merge these data points.

Single Link Clustering

p1	0		
(p2, p5, p3, p6)	0.22	0	
p4	0.37	0.15	0
	p1	(p2, p5, p3, p6)	p4

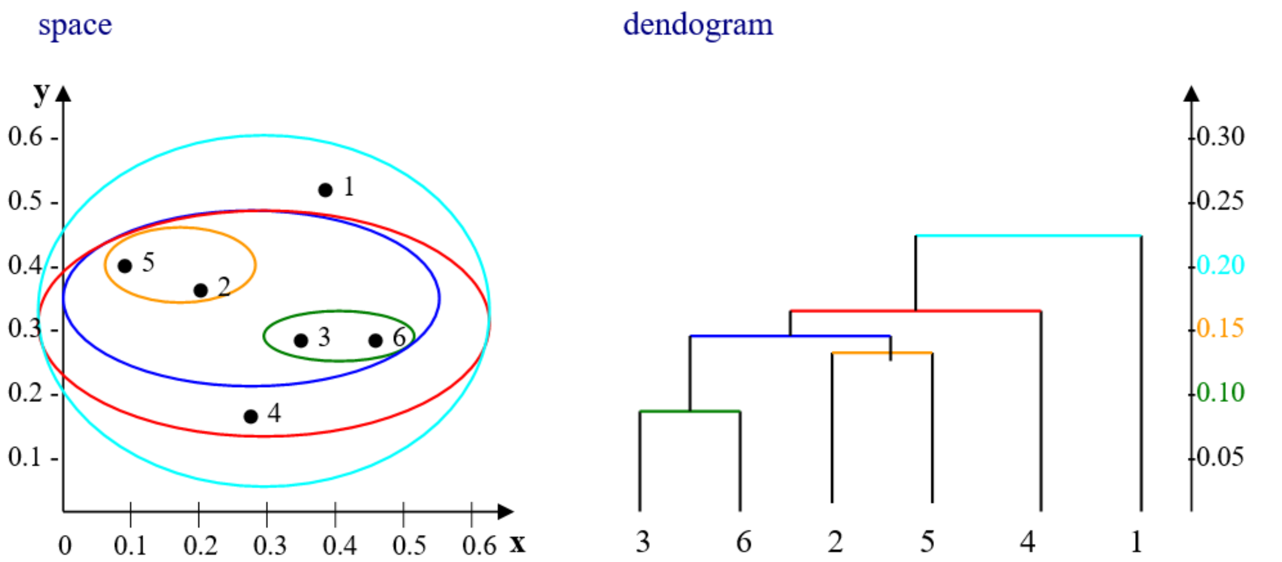
p1	0	
(p2, p5, p3, p6, p4)	0.22	0
	p1	(p2, p5, p3, p6, p4)



- Next, the minimum distance between (p2, p5, p3, p6) and p4 is 0.15.
- Fourth, merge these data points.

Single Link Clustering

p1	0	
(p2, p5, p3, p6, p4)	0.22	0
	p1	(p2, p5, p3, p6, p4)



- Next, the minimum distance between (p2, p5, p3, p6, p4) and p1 is 0.22.
- Last, merge these data points.

Hierarchical Clustering Algorithm

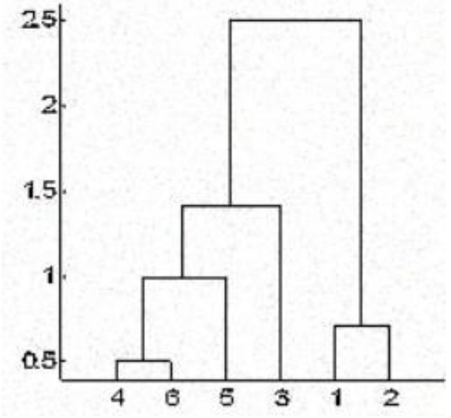
- The key operation in hierarchical agglomerative clustering is to repeatedly combine the two nearest clusters into a larger cluster. There are three key questions that need to be answered first
  - **How do we represent a cluster of more than one point?**
  - **How do we determine the "nearness" of clusters?**
  - **When do we stop combining clusters?**
- **Before applying hierarchical clustering, consider the following points**

1. It starts by calculating the distance between every pair of observation points and store it in a distance matrix.
2. **It puts every point in its own cluster.**
3. **It starts merging the closest pairs of points based on the distances from the distance matrix and as a result, the amount of clusters goes down by 1.**
4. It recomputes the distance between the new cluster and the old ones, and stores them in a new distance matrix.
5. Lastly it repeats the steps 2 and 3 until all the clusters are merged into one single cluster.

Cluster Merging using Dendrograms

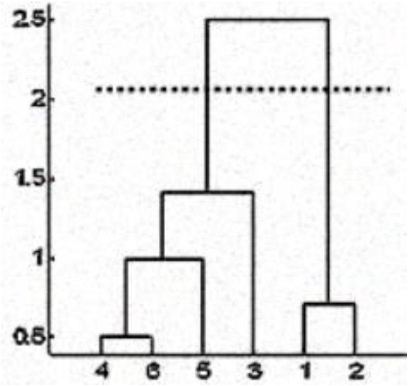
Stopping Criterion

– **How do you decide when to stop merging the clusters?**



- For example, if we are clustering football players on a field based on their positions on the field which will represent their coordinates for distance calculation, we already know that we should end with only 2 clusters as there can be only two teams playing a football match.

- If we don't have that information too. In such cases, we can leverage the results from the dendrogram to approximate the number of clusters.



- We cut the dendrogram tree with a horizontal line at a height where the line can traverse the maximum distance up and down without intersecting the merging point.

- In the above case, it would be between heights 1.5 and 2.5 as shown in the figure. If we make the cut as shown, we end up with only two clusters.

- **The stop of merging depends on the domain knowledge about the data.**

Linkage Methods of Clustering

College Dublin
Computing • IT • Business

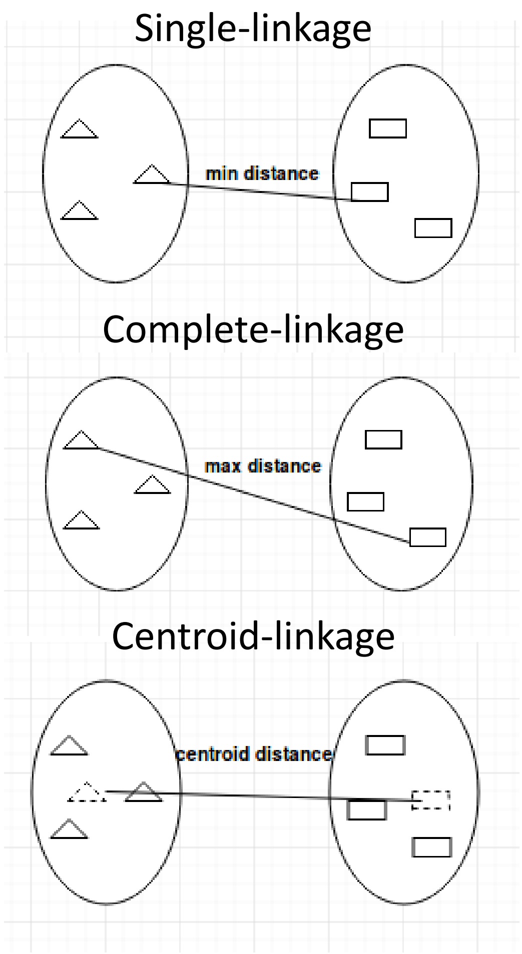
- There are several ways to measure the distance between clusters in order to decide the rules for clustering, and they are called **Linkage Methods**.

- **Single-linkage:** calculates the minimum distance between the clusters before merging. This linkage may be used to detect high values in your dataset which may be outliers as they will be merged at the end.

- **Complete-linkage:** calculates the maximum distance between clusters before merging.

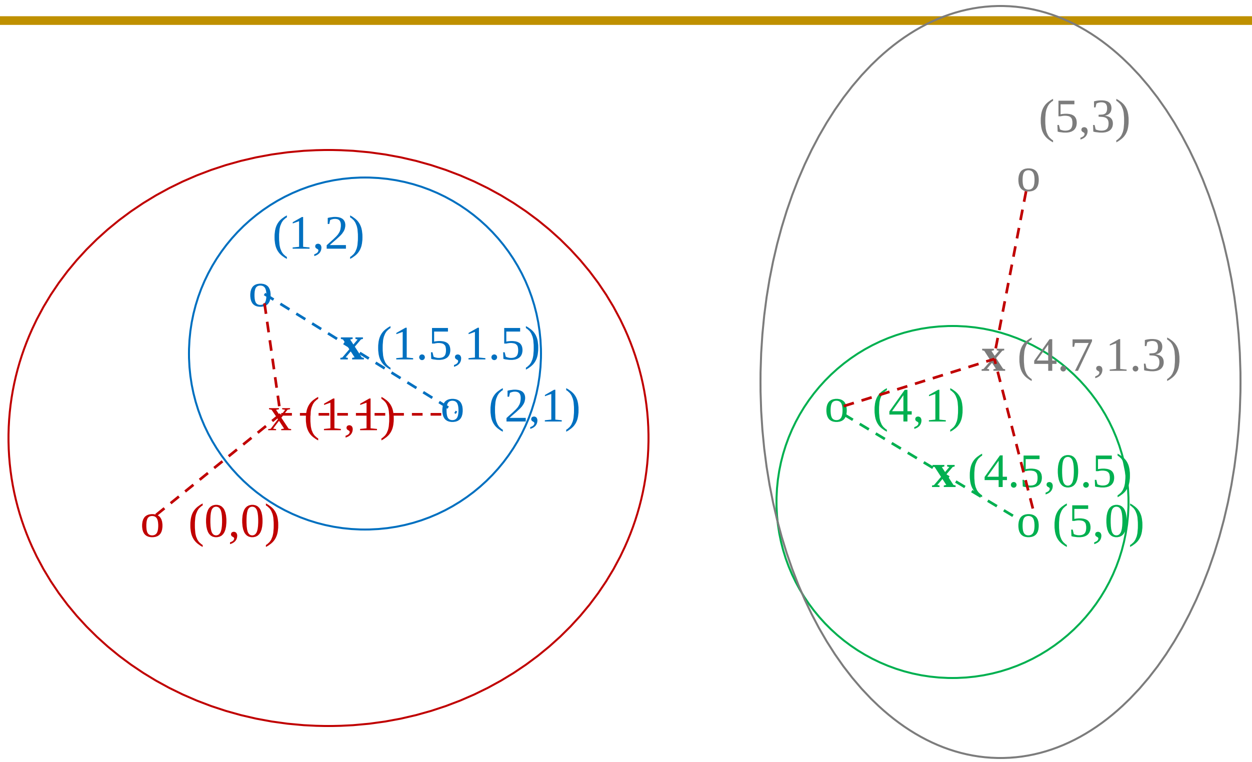
- **Centroid-linkage:** finds centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.

- The selection of linkage method entirely depends on the problem requirement and there is not any method that will always give us good results. Different linkage methods lead to different clusters.



Hierarchical Clustering

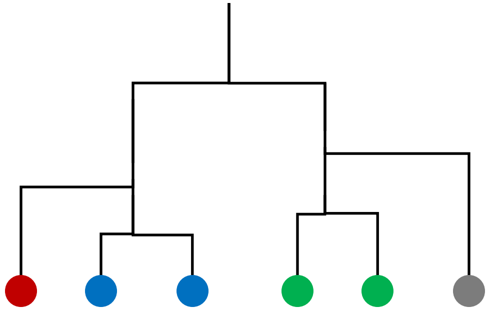
Animation



Data:

o … data point

x … centroid



Dendrogram

Resources/ References

- Discovering Knowledge In Data: An Introduction To Data Exploration, Second Edition, By Daniel Larose And Chantal Larose, John Wiley And Sons, Inc., 2014.
- Introduction to Machine Learning with Python A Guide for Data Scientists, Andreas C. Müller and Sarah Guido, Copyright © 2017, O'Reilly.
- Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning Paperback – 23 Mar. 2018. by. Chris Albon
- Thoughtful Machine Learning by Matthew Kirk Published by O'Reilly Media, Inc., 2014
- https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589
- https://github.com/ApoorvRusia/Naive-Bayes-classification-on-Iris-dataset
- Some images are used from google search repository to enhance the level of learning.
- Sample datasets for Regression: https://www.kaggle.com/tags/regression