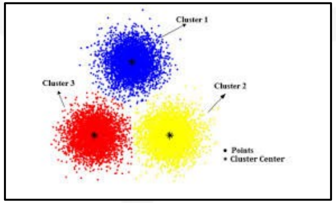
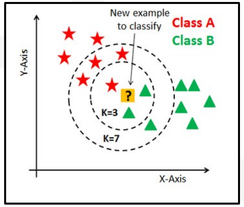
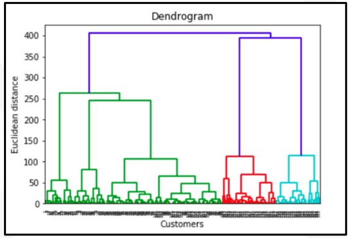
Hierarchical Clustering

Clustering Overview

- K-Means
  - Good if you have a lot of data
  - Estimates the underlying group structure of the population
- DBSCAN
  - Good for non-spherical clusters with similar density
  - Good for cases with noise/outliers
- Hierarchical
  - Clusters don’t have to be the same size or density





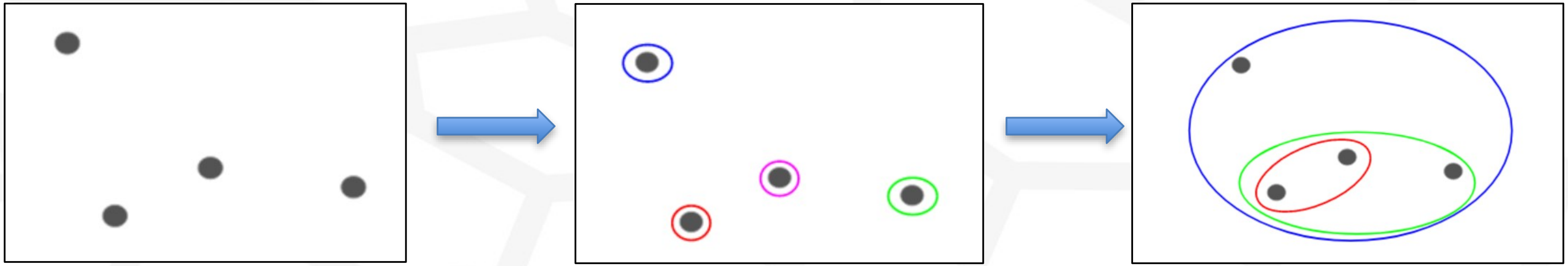


Hierarchical Clustering - Overview

Start with data in some feature space

Assign a single cluster to each point

Iteratively merge closest clusters until we have one

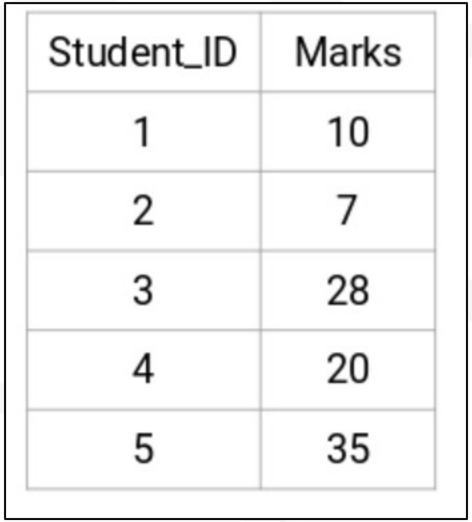


Hierarchical Clustering - Example

- A teacher wants to assign students to groups based on their grades on an assignment

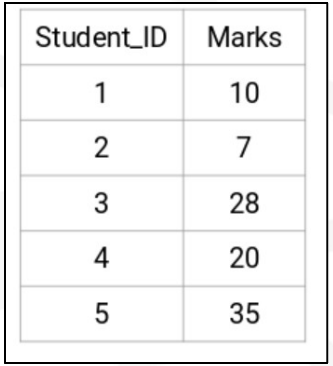
- There’s no fixed target on how many groups there should be

- The teacher doesn’t know what type of student should be in which group (unsupervised learning problem)



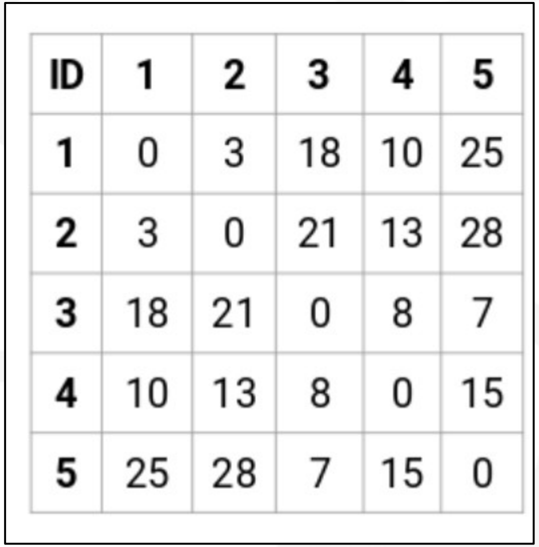
Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

Example – Create a Proximity matrix



Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

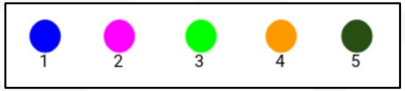
$$Distance(Proximity) = |m_a - m_b|$$

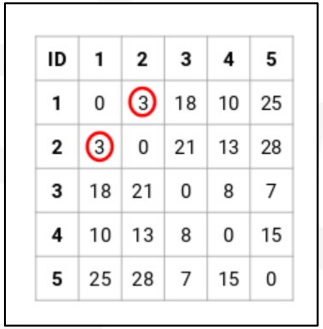


ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

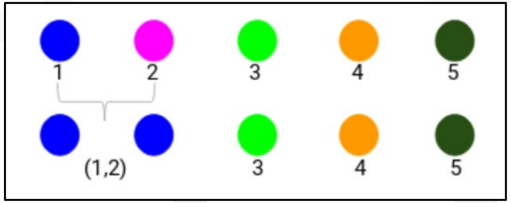
Example – Clustering Process

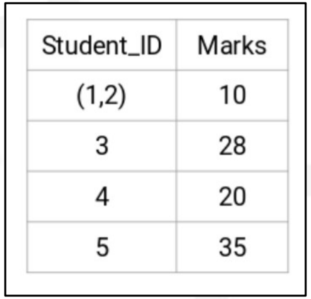
Find closest two clusters
and merge them





ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

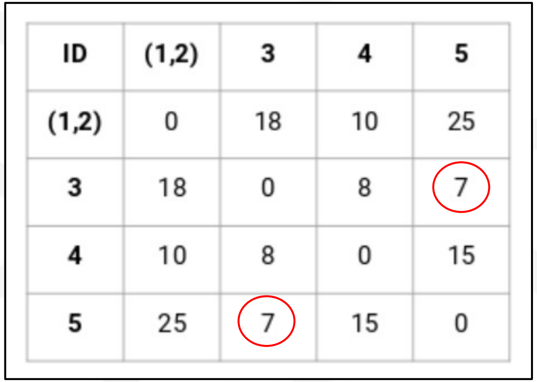




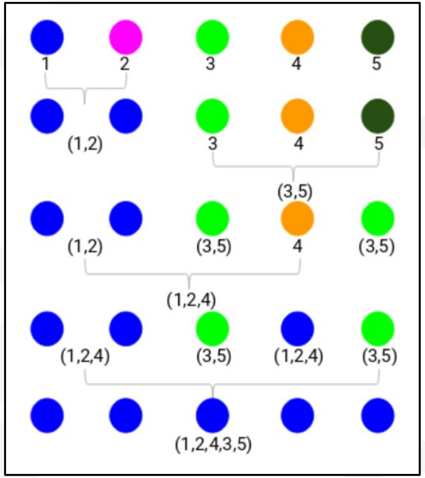
Student_ID	Marks
(1,2)	10
3	28
4	20
5	35

Example – What Number of Clusters?

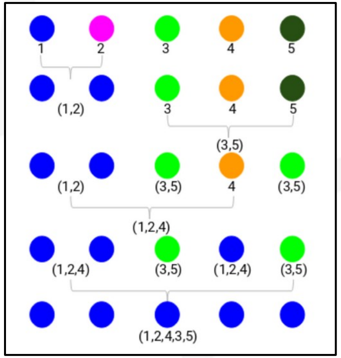
Keep iterating until you
have one cluster



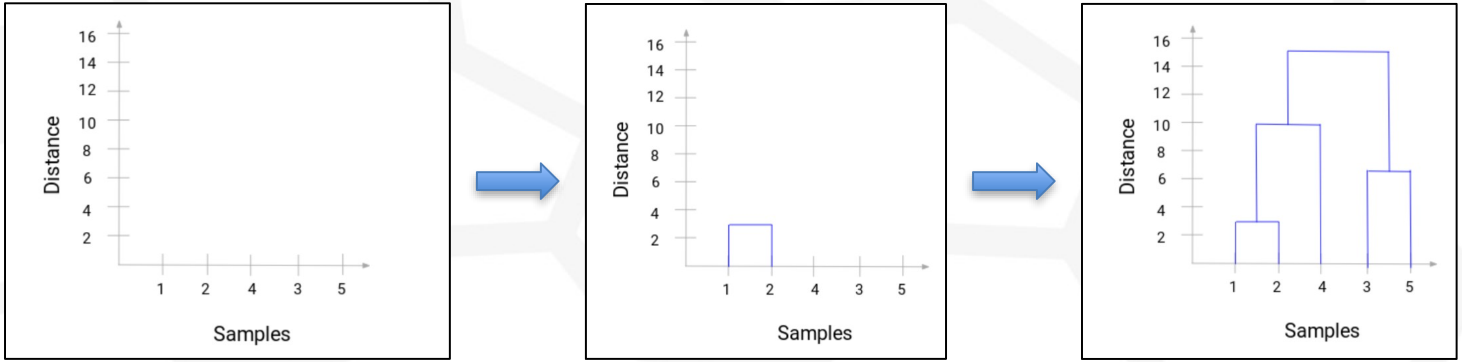
ID	(1,2)	3	4	5
(1,2)	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0



Example – Iteration

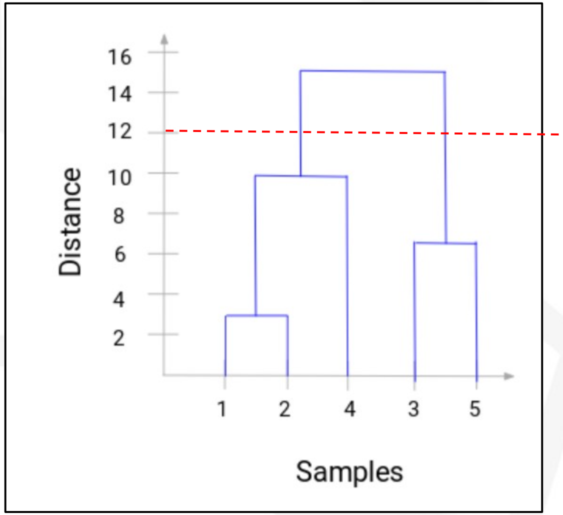


- <u>Dendogram</u>
  - – Horizontal lines indicate the distance at where clusters were merged
  - – The more distance of the vertical lines, the more distance between those clusters



Example – Setting the Threshold

- Dendogram
  - The number of clusters will be the number of vertical lines intersected by the line drawn using the threshold
  - Example on right shows a threshold of 12 which results in 2 clusters
  - Rule of thumb – threshold is placed where the longest vertical gap between clusters resides


Distance
Samples