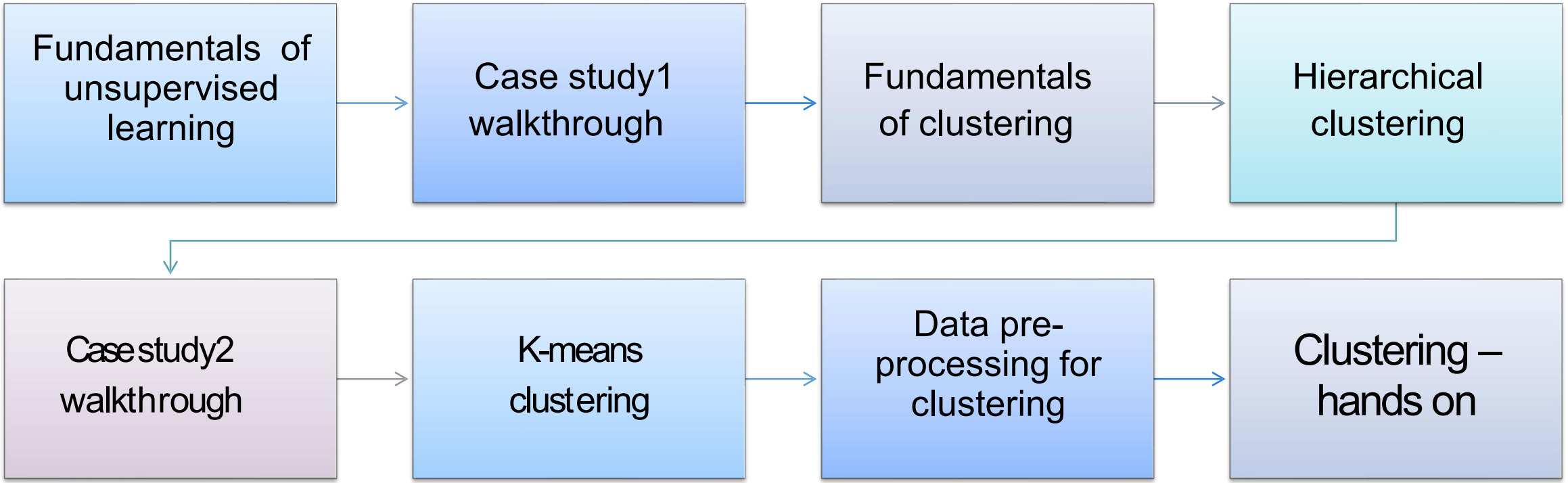
Unsupervised Learning - Clustering

Agenda - Clustering



What is unsupervised learning?

No defined dependent and independent variables.

Patterns in the data are used to identify / group similar observations

Supervised vs unsupervised learning

Supervised learning

- Clearly defined X and Y variables
- Predict a continuous response (Regression)
- categorical response (classification)

Unsupervised learning

- Unlabelled data
- Emerging patterns based on similarity identified
- Clustering
- Association rules (market basket analysis)

What is clustering?

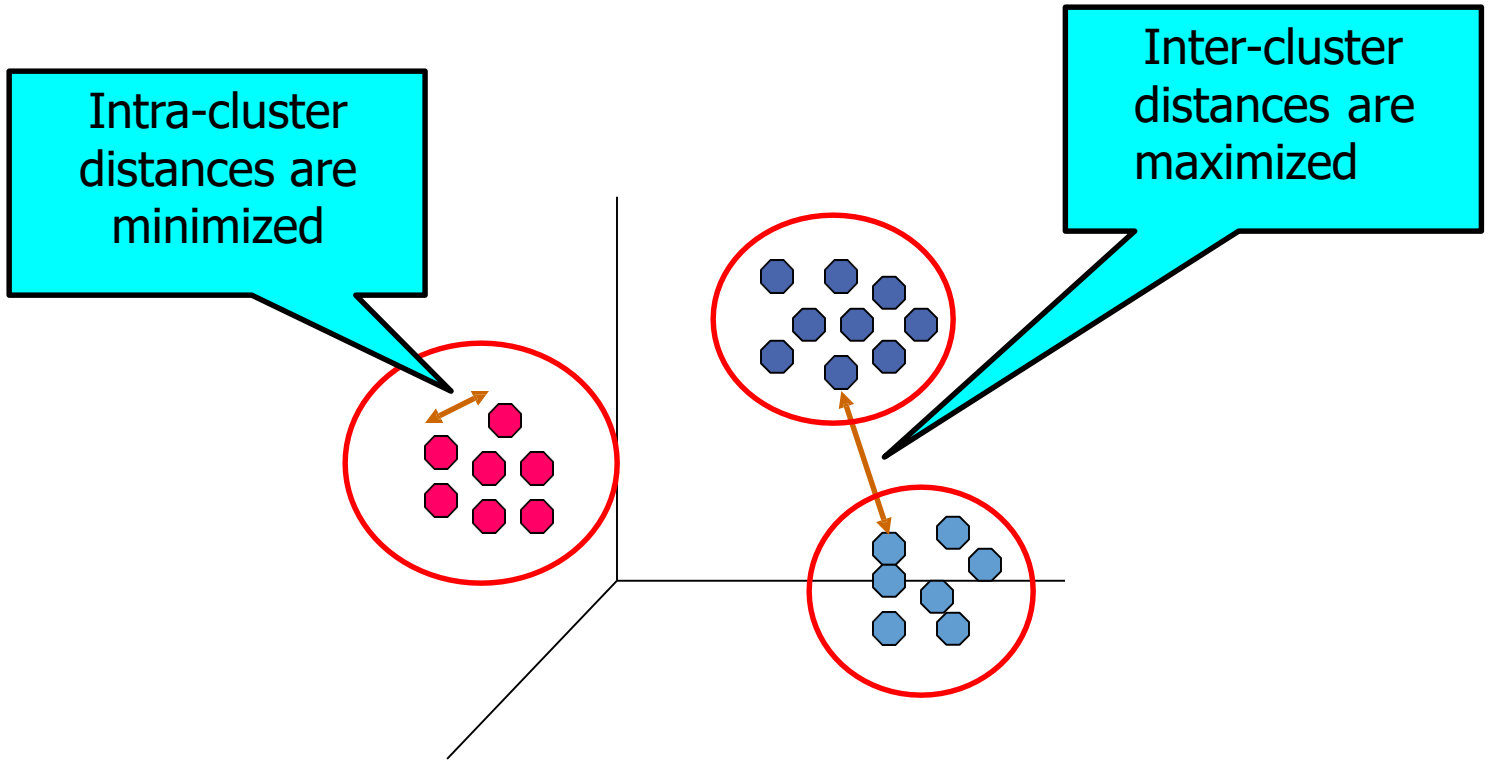
Grouping objects

Heterogeneity between groups

Homogeneity within groups

SSB > SSW

What is clustering?



This is classic case of Unsupervised learning

Why do we cluster?

Group records such that

- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters

Clustering results are used:

- As a stand-alone tool to get insight into data distribution
- Visualization of clusters may unveil important information
- As a preprocessing step for other algorithms

Cluster Analysis – Use cases

Image processing	Web	Market Segmentation	Market structure analysis	Finance
• cluster images based on their visual content	• Cluster groups of users based on their access patterns on webpages • Cluster webpages based on their content	• customers are segmented based on demographic and transaction history information, and a marketing strategy is tailored for each segment	• identifying groups of similar products according to competitive measures of similarity	• cluster analysis can be used for creating *balanced portfolios*

Clustering vs PCA

PCA – grouping variables that relate to each other

Clustering – Segment variables according to the distance between them.

Grouping of similar rows

		COMP1			COMP2				COMP		
	AID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
CLUSTER1	1	2.51	9.19	4.45	5.33	7.27	0.7	5.85	4.01	1.34	6.1
	2	7.51	1.77	2.01	9.31	6.61	7.69	3.29	8.85	0.2	6.35
	3	2.52	2.61	5.65	1.24	0.97	2.85	9.87	3.14	3.7	5.17
	4	6.56	5.9	1.65	6.69	8.04	0.8	1.91	7.42	8.02	1.43
	5	6.91	7.78	5.63	3.84	8.99	1.56	0.13	7.29	6.45	9.58
CLUSTER2	6	2.63	3.16	1.39	0.55	9.85	4.58	0.97	5.89	0.04	3.88
	7	3.78	9.9	5.07	5.41	3.27	4.04	2.11	9.47	4.98	0.32
	8	5.63	6.86	9.24	4.47	5.46	7.05	7.7	9.21	7.99	9.51
	9	6.09	8.36	1.03	1.81	0.58	2.02	9.86	8.2	0.81	0.25
	10	2.26	3.48	7.69	0.9	6.07	0.74	2.31	6.48	0.45	6.78
CLUSTER3	11	3.79	2.52	2.93	1.92	7.12	4.22	2.07	6.73	1.35	6.64
	12	6.37	5.13	4.09	1.39	3.74	3.67	5.46	4.17	1.6	0.92
	13	3.9	8.14	8.91	4.7	8.73	8.5	5.75	6.76	0.17	5.08
	14	2.07	3.23	2.8	0.43	8.51	0.48	2.52	8.83	0.01	0.37
	15	1.39	8.66	3.57	6.68	2.54	4.89	7.27	2.75	7.43	9.89

Role of cluster analysis

Data Reduction	Taxonomy description	Influence of cluster on Y variable
• Classify observations into manageable groups	• Exploratory • Confirmatory	• What is the average sales from each customer segment? • How does churn % vary for each customer segment

The clustering task

Group observations so that the observations belonging in the same group are similar, whereas observations in different groups are dissimilar.

Observations to cluster - dimensions

Real-value attributes	• salary, height
Binary attributes	• gender (M/F), has_cancer(T/F)
Nominal (categorical) attributes	• religion (Christian, Muslim, Buddhist, Hindu, etc.)
Ordinal/Ranked attributes	• military rank (soldier, sergeant, lutenant, captain, etc.)

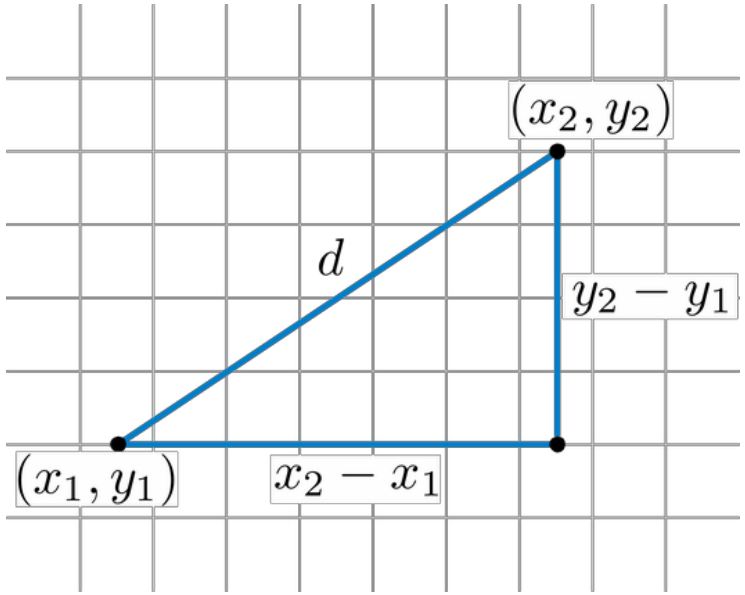
Measuring similarity - Distances

- Eucledian distance
- Manhattan distance
- Chebyshev distance

Euclidean distance

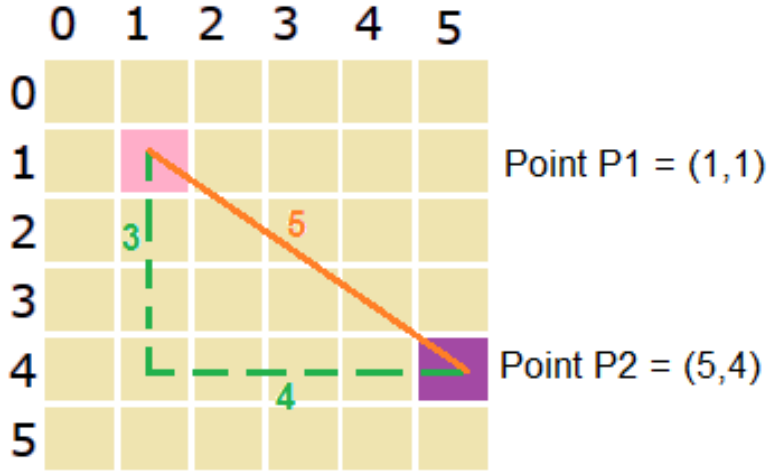
$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

Where x_1 to x_p are the independent variables of i and j



Manhattan distance (city –block distance)

- Distance between the projection of points on the axis.



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

Chebyshev Distance (chessboard distance)

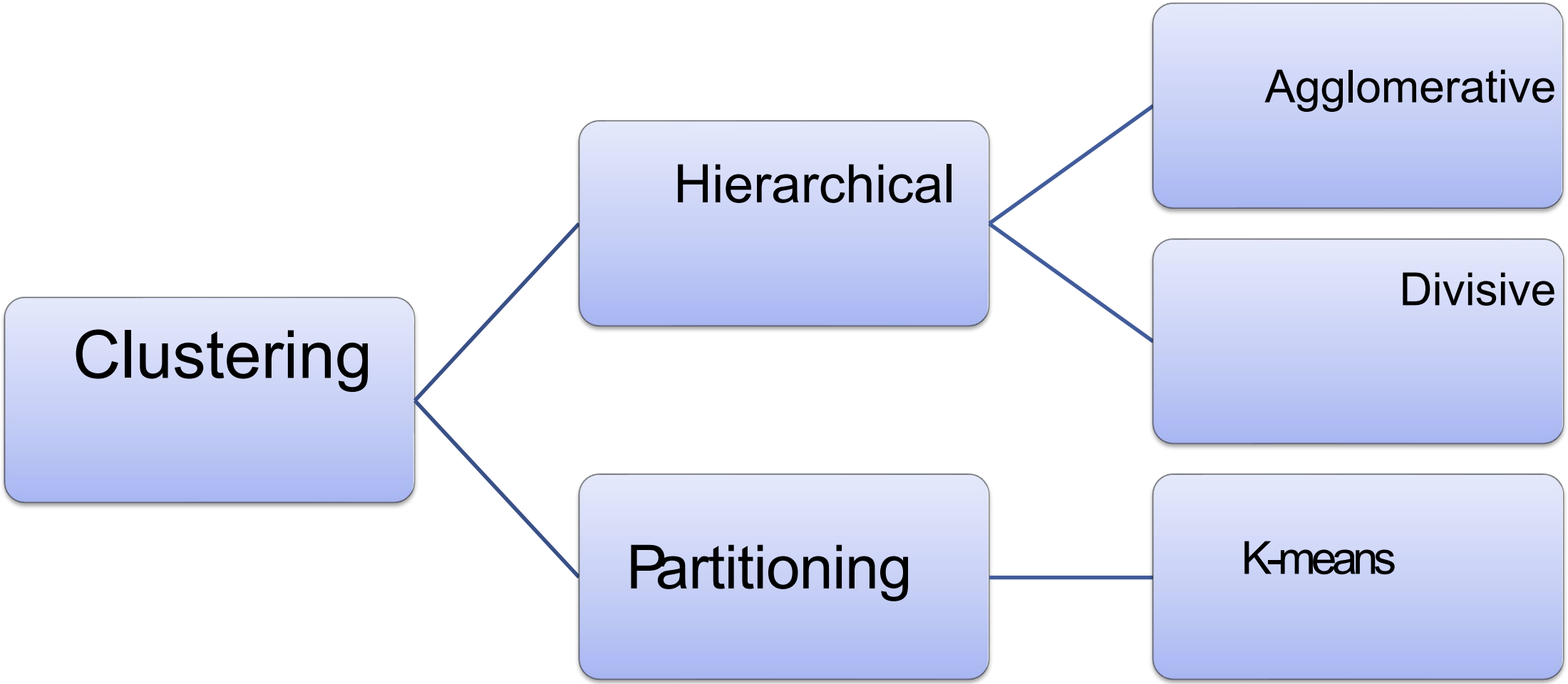
Max (|x1-x2|, |y1-y2|, |z1-z2|, ….)

Minkowski Distance

Mathematical formula: $\left(\sum_{i=1}^{m}|x_i - y_i|^p\right)^{1/p}$

- If p=2, then the above equation resembles the equation of Euclidean Distance.
- If p=1, then the above equation resembles the equation of Manhattan Distance.

Types of clustering



Clustering types

Agglomerative clustering	Divisive approach	Partitioning
• Bottom up approach • start with each object forming a separate group • It keeps on merging the objects or groups that are close to one another	• Top down approach • start with all of the objects in the same cluster • a cluster is split up into smaller clusters	• constructs ‘k’ partition of data • Each partition will represent a cluster and k ≤ n

Hierarchical clustering

- Records are sequentially grouped to create clusters, based on distances between records and distances between clusters.
- Hierarchical clustering also produces a useful graphical display of the clustering process and results, called a dendrogram.

Strengths of Hierarchical Clustering

- No assumptions on the number of clusters
- Any desired number of clusters can be obtained by ‘cutting’

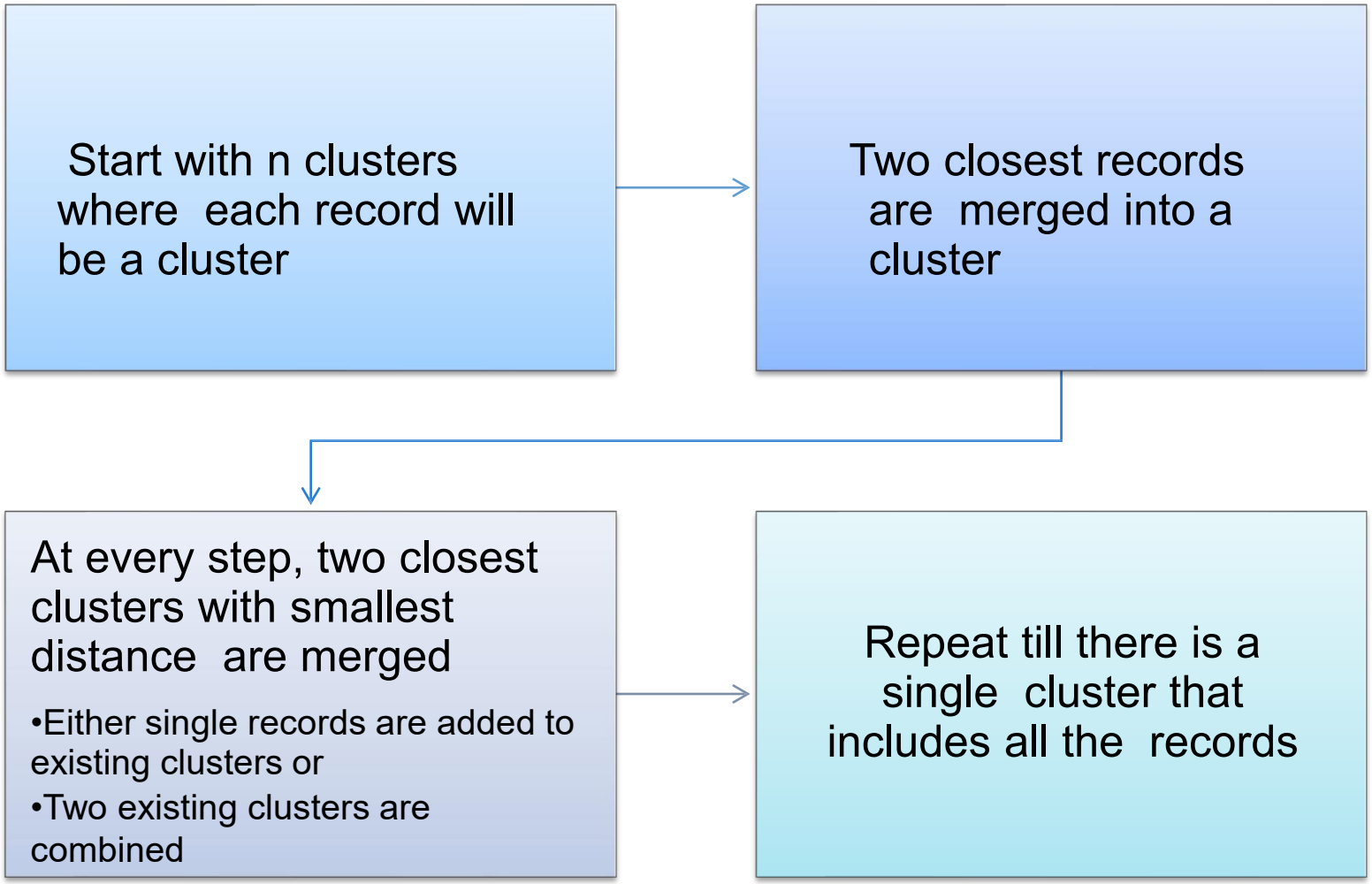
the dendrogram at the proper level

- Hierarchical clustering may correspond to meaningful taxonomies

Disadvantages of Hierarchical clustering

- Time complexity: not suitable for larger data sets.
- Very sensitive to outliers

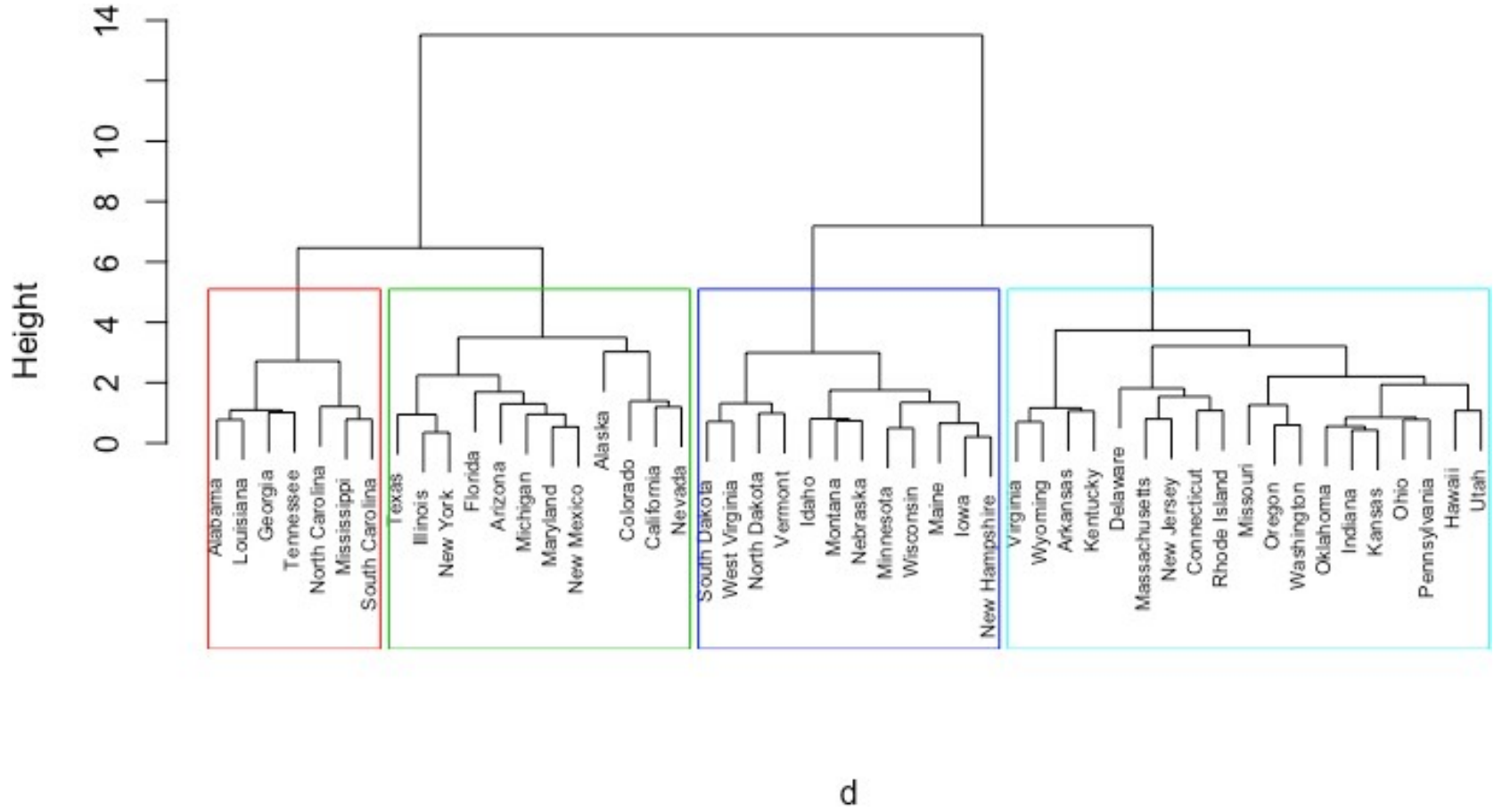
Hierarchical clustering - Steps



Dendrograms

- A *dendrogram* is a treelike diagram that summarizes the process of clustering
- On the *x*-axis are the records
- Similar records are joined by lines whose vertical length reflects the distance between the records
- the greater the difference in height, the more dissimilarity
- By choosing a cutoff distance on the *y*-axis, a set of clusters is created

Dendrograms



Distance between two clusters

- Each cluster is a set of points
- How do we define distance between two sets of points

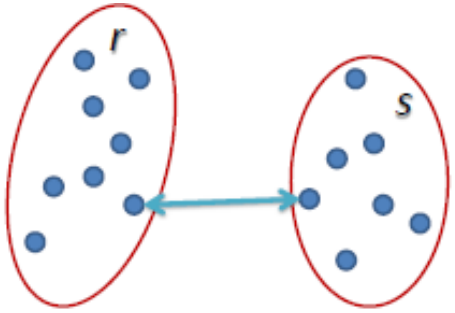
Hierarchical clustering – Distance between clusters

Linkage types

- Single linkage
- Complete linkage
- Average linkage
- Centroid linkage
- Ward's Method

Single Linkage

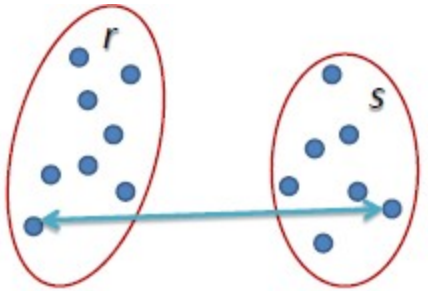
- Distance between two clusters is defined as the shortest distance between two points in each cluster.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete linkage

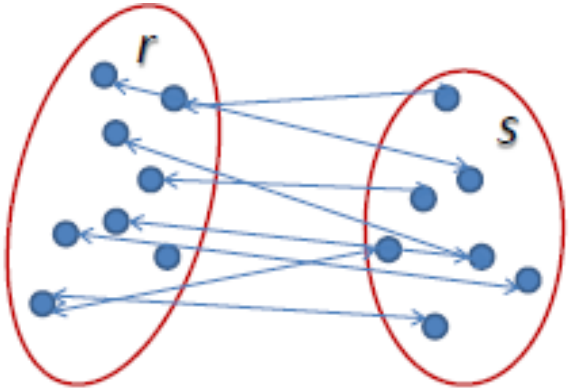
- Distance between two clusters is defined as the longest distance between two points in each cluster.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average linkage

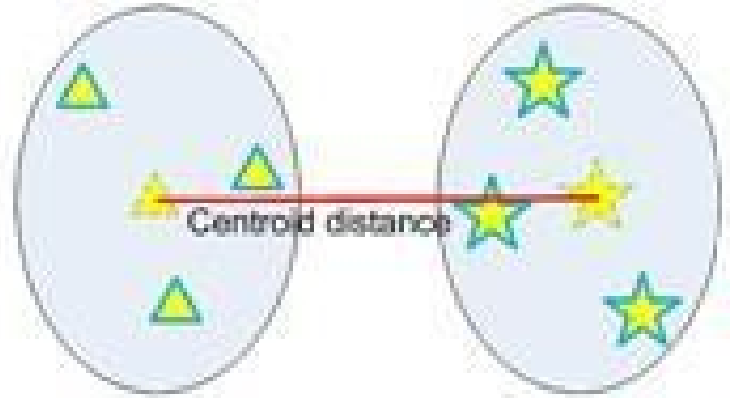
- Distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Centroid linkage

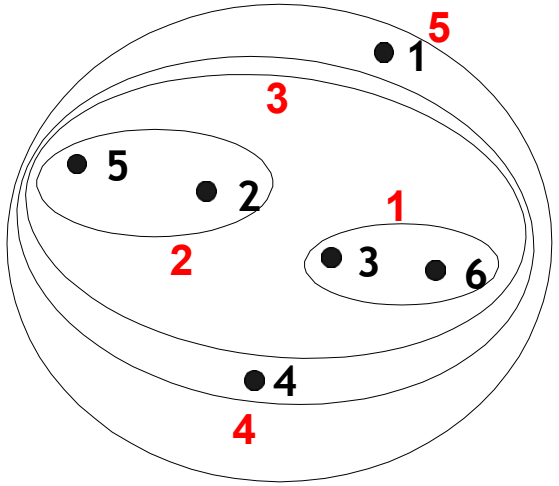
- Based on centroid distance. clusters are represented by their mean values for each variable, which forms a vector of means.
- Distance between 2 clusters is distance between the 2 vectors



Ward's linkage

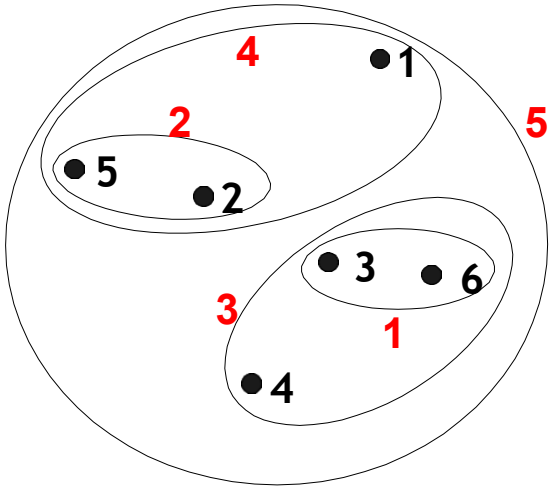
- Similar to group average and centroid distance
- joins records and clusters together progressively to produce larger and larger clusters, but operates slightly differently from the general approach.

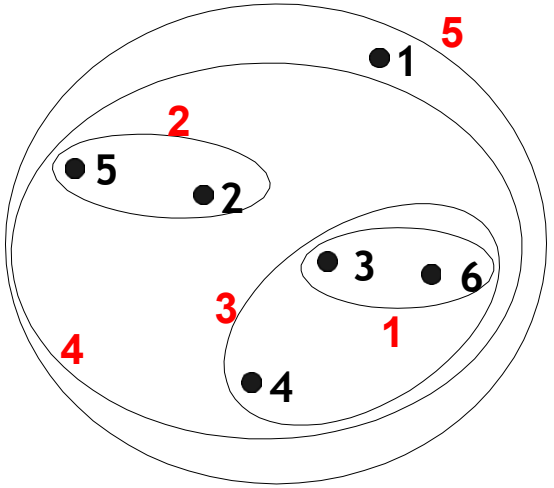
Hierarchical Clustering: Comparison



MIN

MAX





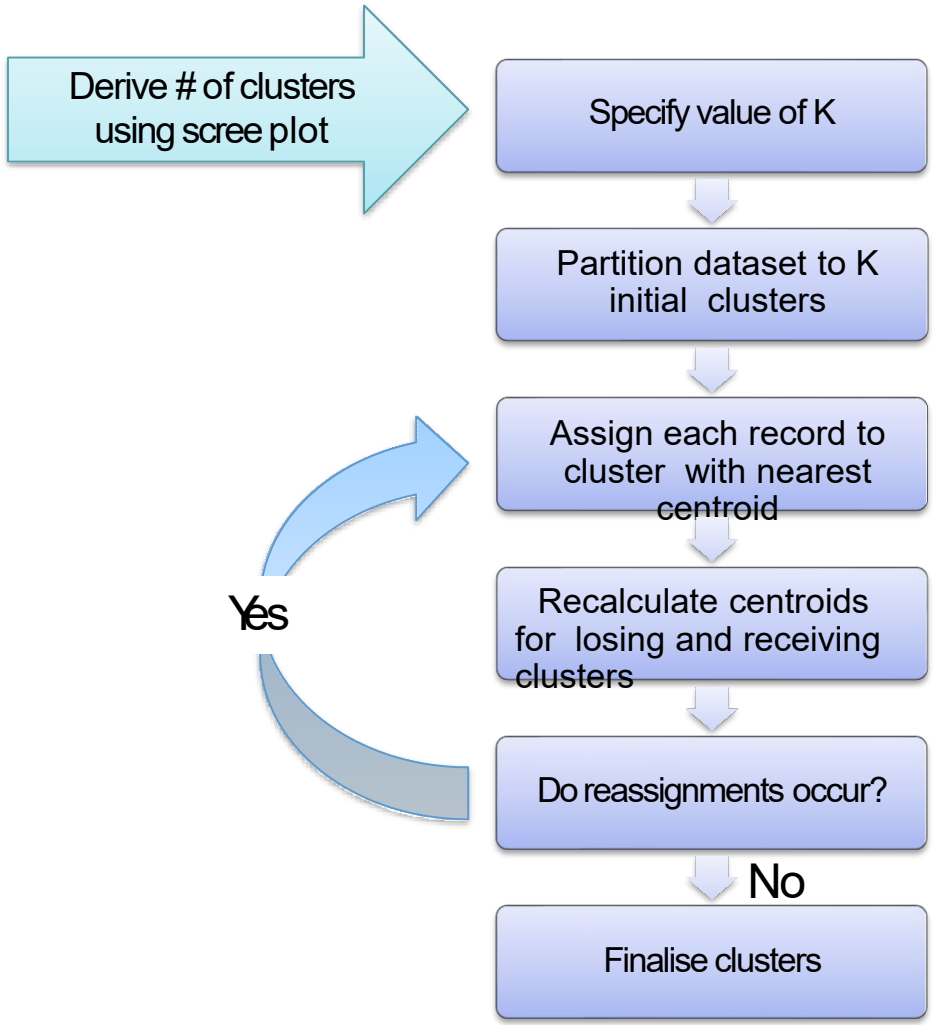
Group Average

K-MEANS CLUSTERING

K-means Clustering

- A non-hierarchical approach to forming good clusters is to pre-specify a desired number of clusters, *k*
- *A*ssign each record to one of the *k* clusters, according to their distance from each cluster
- So as to minimize a measure of dispersion within the clusters
- *The 'means'* in the K-means refers to averaging of the data; that is, finding the centroid
- *K*-means clustering is widely used in large dataset applications

How does k-means clustering work?



Scaling – Z scaling & Min-max scaling

Z Scaling

- Features will be rescaled
- Have the properties of a standard normal distribution
- μ=0 and σ=1

$$z = \frac{x - \mu}{\sigma}$$

Min Max scaling

- The data is scaled to a fixed range - 0 to 1.
- The cost of having this bounded range - smaller standard deviations, which can suppress the effect of outliers

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where is scaling used?

k-nearest neighbors

k-means

Perceptron , Neural networks

principal componen t analysis

Validating Clusters

- The resulting clusters should be valid to generate insights
- Cluster interpretability
- Cluster stability
- Cluster separation
- Number of clusters

Questions?