


Module -5

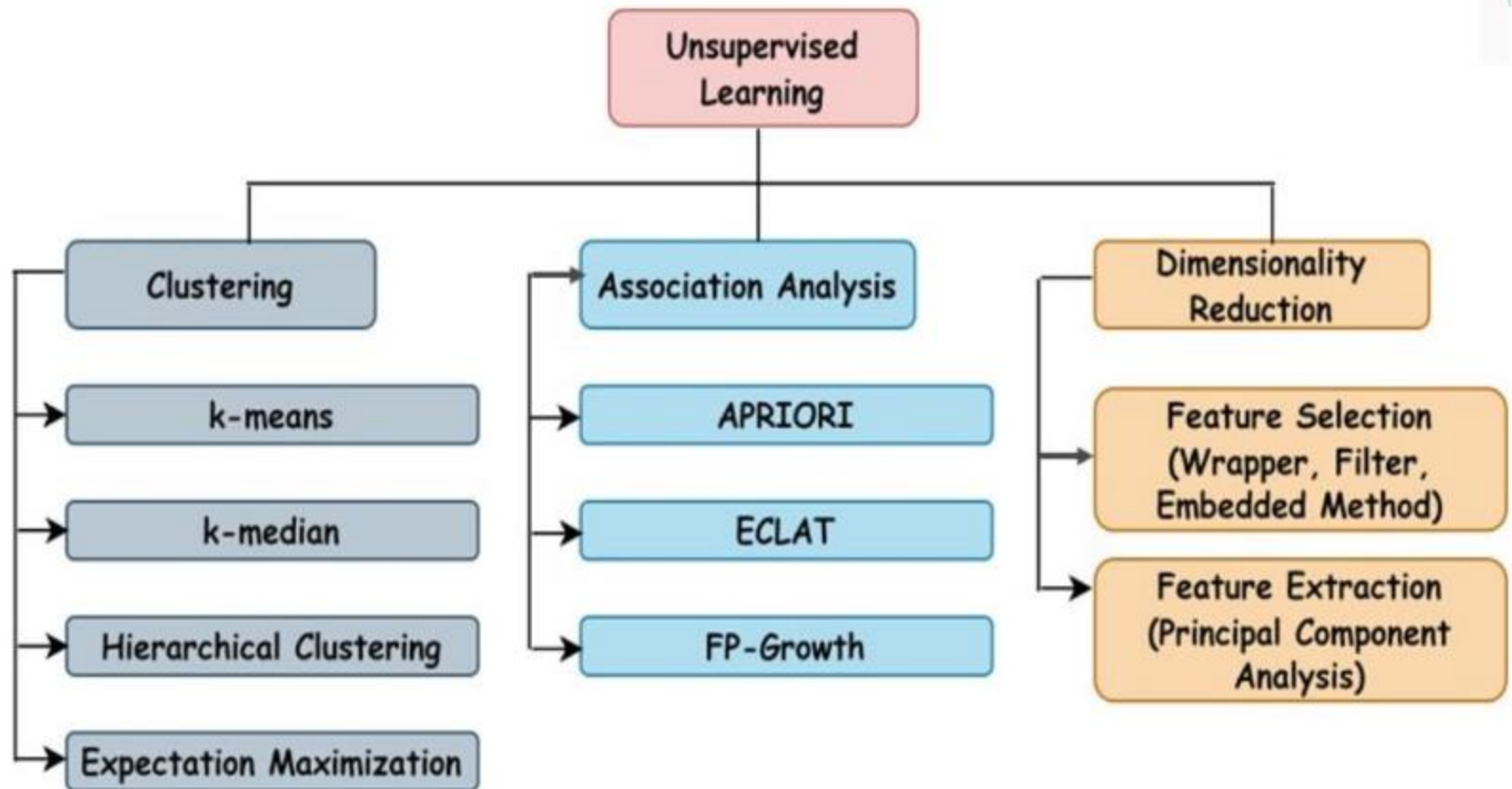
Dr. Parimala M

SCORE, VIT

Module:5	Unsupervised Learning	7 hours
Introduction - K-Means Clustering - Expectation Maximization Algorithm - Supervised Learning after Clustering - Hierarchical Clustering - Density Based Clustering - Evaluation Metrics - Association Rule Learning		

Cluster Analysis: Basic Concepts and Methods

- Cluster Analysis: Basic Concepts 
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Evaluation of Clustering
- Summary

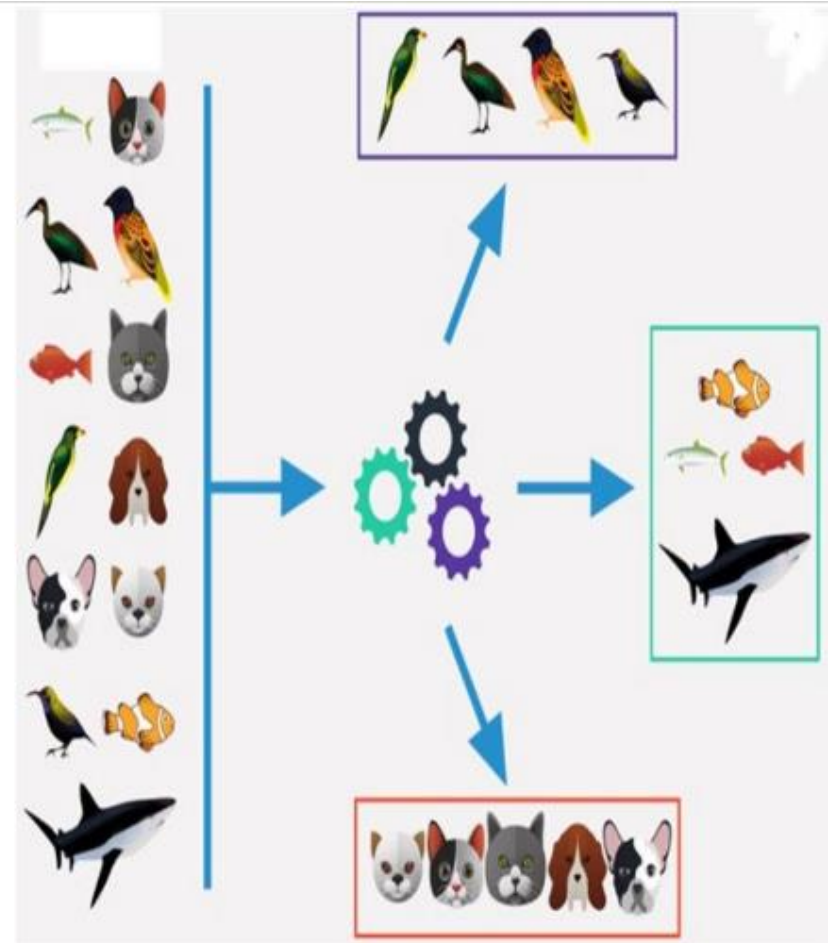


What is Cluster Analysis?

- Cluster: A collection of data objects
 - similar (or related) to one another within the same group
 - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation*, ...)
 - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- **Unsupervised learning**: no predefined classes (i.e., *learning by observations* vs. learning by examples: supervised)
- Typical applications
 - As a **stand-alone tool** to get insight into data distribution
 - As a **preprocessing step** for other algorithms

Clustering

- Aims to group objects into disjoint cluster or groups
- Clusters are formed based on attributes
- All data objects in one cluster are similar in some aspects and vary from other cluster



Applications of Cluster Analysis

- Data reduction
 - Summarization: Preprocessing for regression, PCA, classification, and association analysis
 - Compression: Image processing: vector quantization
- Hypothesis generation and testing
- Prediction based on groups
 - Cluster & find characteristics/patterns for each group
- Finding K-nearest Neighbors
 - Localizing search to one or a small number of clusters
- Outlier detection: Outliers are often viewed as those “far away” from any cluster

Clustering: Application Examples

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering
- 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
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean
- Economic Science: market research

Basic Steps to Develop a Clustering Task

- Feature selection
 - Select info concerning the task of interest
 - Minimal information redundancy
- Proximity measure
 - Similarity of two feature vectors
- Clustering criterion
 - Expressed via a cost function or some rules
- Clustering algorithms
 - Choice of algorithms
- Validation of the results
 - Validation test (also, *clustering tendency* test)
- Interpretation of the results
 - Integration with applications

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters
 - high intra-class similarity: **cohesive** within clusters
 - low inter-class similarity: **distinctive** between clusters
- The quality of a clustering method depends on
 - the similarity measure used by the method
 - its implementation, and
 - Its ability to discover some or all of the hidden patterns

Measure the Quality of Clustering

- Dissimilarity/Similarity metric
 - Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
 - The definitions of distance functions are usually rather different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables
 - Weights should be associated with different variables based on applications and data semantics
- Quality of clustering:
 - There is usually a separate “quality” function that measures the “goodness” of a cluster.
 - It is hard to define “similar enough” or “good enough”
 - The answer is typically highly subjective

Considerations for Cluster Analysis

- Partitioning criteria
 - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)
- Separation of clusters
 - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- Similarity measure
 - Distance-based (e.g., Euclidian, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- Clustering space
 - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

Requirements and Challenges

- Scalability
 - Clustering all the data instead of only on samples
- Ability to deal with different types of attributes
 - Numerical, binary, categorical, ordinal, linked, and mixture of these
- Constraint-based clustering
 - User may give inputs on constraints
 - Use domain knowledge to determine input parameters
- Interpretability and usability
- Others
 - Discovery of clusters with arbitrary shape
 - Ability to deal with noisy data
 - Incremental clustering and insensitivity to input order
 - High dimensionality


Major Clustering Approaches (I)

- Partitioning approach:
 - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
 - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical approach:
 - Create a hierarchical decomposition of the set of data (or objects) using some criterion
 - Typical methods: Diana, Agnes, BIRCH, CAMELEON
- Density-based approach:
 - Based on connectivity and density functions
 - Typical methods: DBSACN, OPTICS, DenClue
- Grid-based approach:
 - based on a multiple-level granularity structure
 - Typical methods: STING, WaveCluster, CLIQUE

Major Clustering Approaches (II)

- Model-based:
 - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
 - Typical methods: EM, SOM, COBWEB
- Frequent pattern-based:
 - Based on the analysis of frequent patterns
 - Typical methods: p-Cluster
- User-guided or constraint-based:
 - Clustering by considering user-specified or application-specific constraints
 - Typical methods: COD (obstacles), constrained clustering
- Link-based clustering:
 - Objects are often linked together in various ways
 - Massive links can be used to cluster objects: SimRank, LinkClus

Chapter 10. Cluster Analysis: Basic Concepts and Methods

- Cluster Analysis: Basic Concepts
- Partitioning Methods 
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Evaluation of Clustering
- Summary

Partitioning Algorithms: Basic Concept

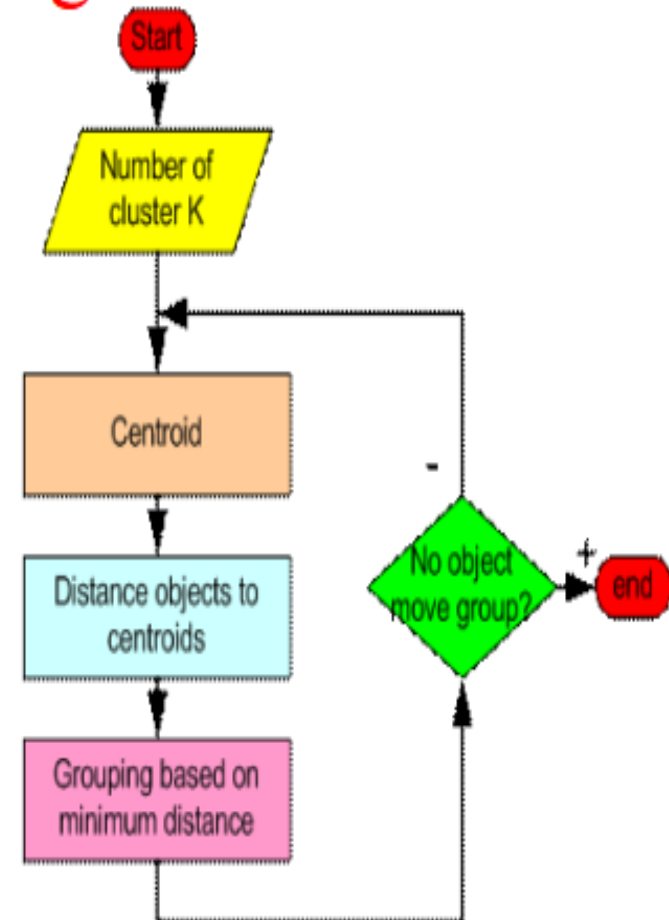
- Partitioning method: Partitioning a database ***D*** of ***n*** objects into a set of ***k*** clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)

$$E = \sum_{i=1}^k \sum_{p \in C_i} (d(p, c_i))^2$$

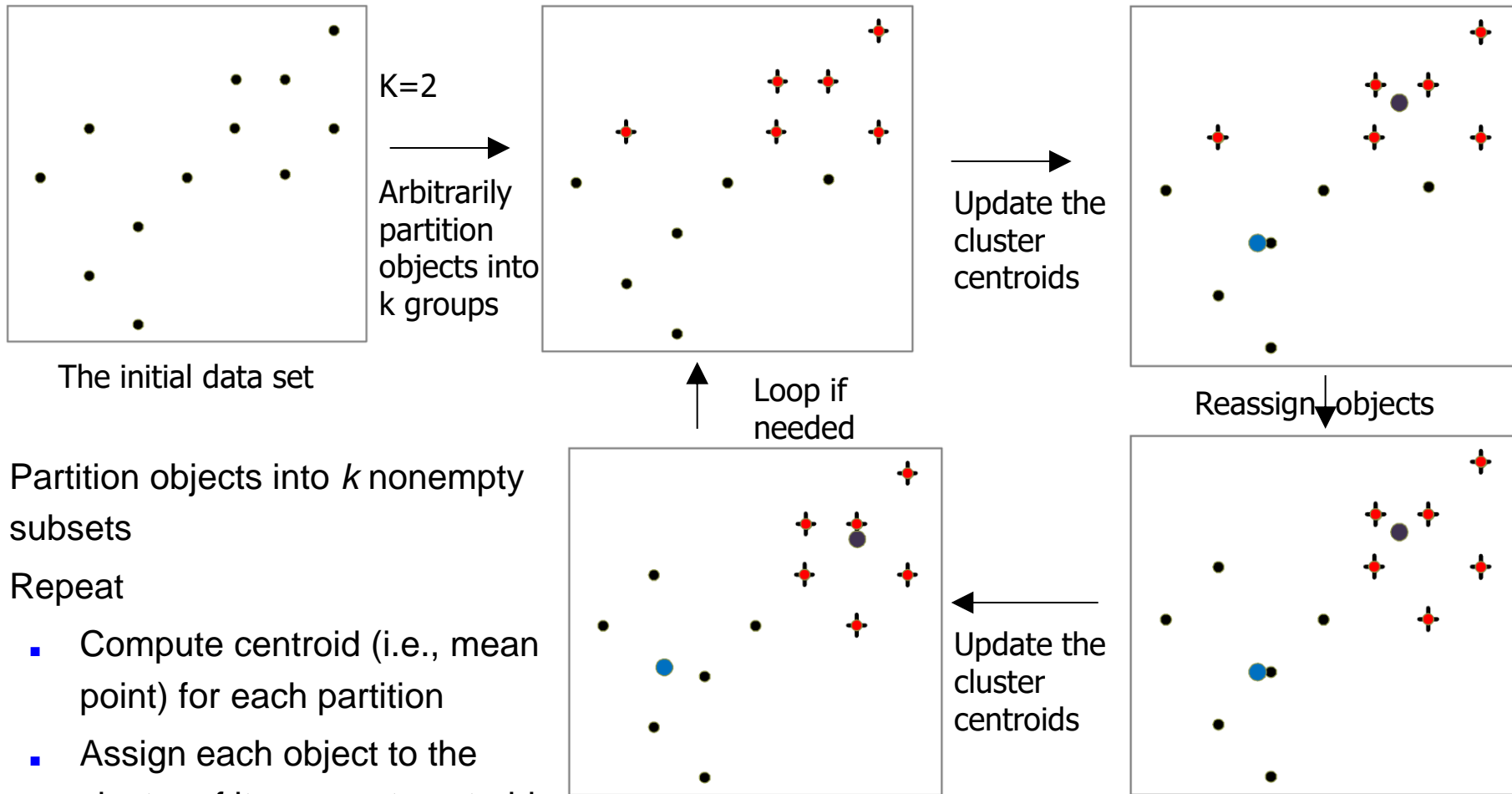
- Given k , find a partition of k clusters that optimizes the chosen partitioning criterion
 - Global optimal: exhaustively enumerate all partitions
 - Heuristic methods: *k-means* and *k-medoids* algorithms
 - *k-means* (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster
 - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

K-means clustering

- K-means algorithm is an unsupervised learning algorithm
- Given a data set of items, with certain features, and values for these features, the algorithm will categorize the items into k groups or clusters of similarity.
- To calculate the similarity, we can use the Euclidean distance, Manhattan distance, Hamming distance, Cosine distance as measurement.
- K-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group.
- K is positive integer number.
- The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.



An Example of *K-Means* Clustering



- Partition objects into k nonempty subsets
- Repeat
 - Compute centroid (i.e., mean point) for each partition
 - Assign each object to the cluster of its nearest centroid

■ Until no change

A Simple example showing the implementation of k-means algorithm (using $K=2$)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 1:

Initialization:

Randomly we choose following two centroids ($k=2$) for two clusters.

In this case the 2 centroid are: $m_1=(1.0,1.0)$ and $m_2=(5.0,7.0)$.

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

$$d(m_1, 2) = \sqrt{|1.0 - 1.5|^2 + |1.0 - 2.0|^2} = 1.12$$

$$d(m_2, 2) = \sqrt{|5.0 - 1.5|^2 + |7.0 - 2.0|^2} = 6.10$$

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Step 2:

- Thus, we obtain two
- clusters containing:
 $\{1,2,3\}$ and $\{4,5,6,7\}$.
- Their new centroids are:

$$m_1 = \left(\frac{1}{3}(1.0 + 1.5 + 3.0), \frac{1}{3}(1.0 + 2.0 + 4.0) \right) = (1.83, 2.33)$$

$$\begin{aligned} m_2 &= \left(\frac{1}{4}(5.0 + 3.5 + 4.5 + 3.5), \frac{1}{4}(7.0 + 5.0 + 5.0 + 4.5) \right) \\ &= (4.12, 5.38) \end{aligned}$$

Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

Step 3:

Now using these centroids we compute the Euclidean distance of each object, as shown in table.

Therefore, the new clusters are: $\{1,2\}$ and $\{3,4,5,6,7\}$

Next centroids are:
 $m1=(1.25,1.5)$ and
 $m2 = (3.9,5.1)$

Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.84	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

Step 4 :

The clusters obtained are:

$\{1,2\}$ and $\{3,4,5,6,7\}$

Therefore, there is no change in the cluster.

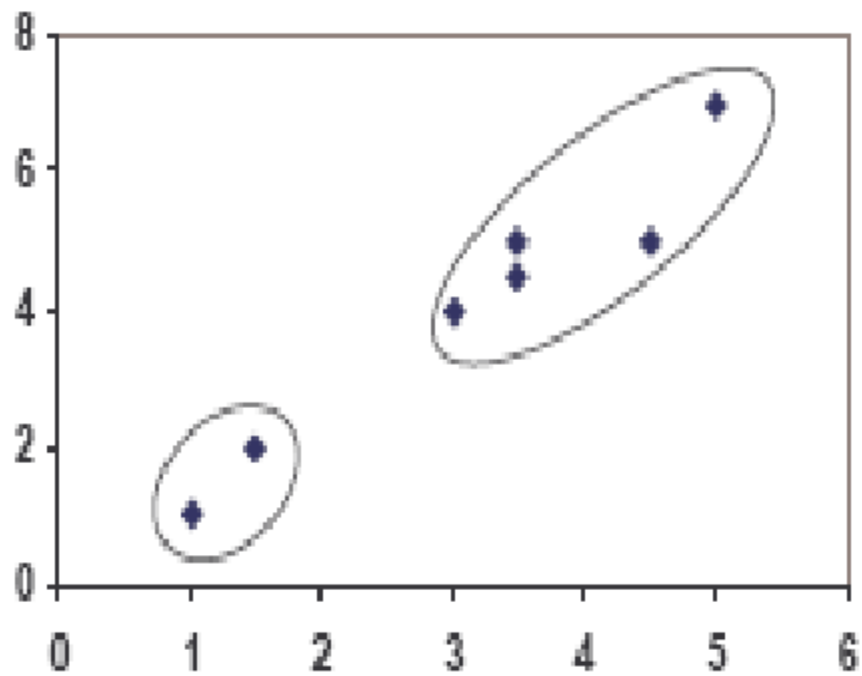
Thus, the algorithm comes to a halt here and final result consist of

2 clusters $\{1,2\}$ and

$\{3,4,5,6,7\}$.

Individual	Centroid 1	Centroid 2
1	0.58	5.02
2	0.58	3.92
3	3.05	1.42
4	6.88	2.20
5	4.18	0.41
6	4.78	0.81
7	3.75	0.72

PLOT



(with K=3) Step 1

Individual	$m_1 = 1$	$m_2 = 2$	$m_3 = 3$	cluster
1	0	1.11	3.81	1
2	1.12	0	2.5	2
3	3.81	2.5	0	3
4	7.21	6.10	3.81	3
5	4.72	3.81	1.12	3
6	5.31	4.24	1.80	3
7	4.30	3.20	0.71	3

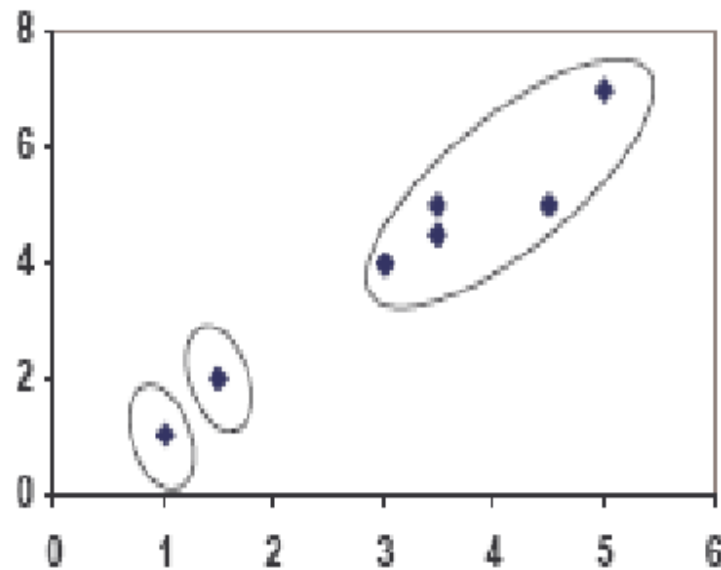
} C_3

clustering with initial centroids (1, 2, 3)

Step 2

Individual	m_1 (1.0, 1.0)	m_2 (1.5, 2.0)	m_3 (3.9, 5.1)	cluster
1	0	1.11	5.02	1
2	1.12	0	3.92	2
3	3.81	2.5	1.42	3
4	7.21	6.10	2.20	3
5	4.72	3.81	0.41	3
6	5.31	4.24	0.81	3
7	4.30	3.20	0.72	3

PLOT



Limitations

- K-means is extremely sensitive to cluster center initializations
- Bad initialization can lead to Poor convergence speed
- Bad initialization can lead to bad overall clustering

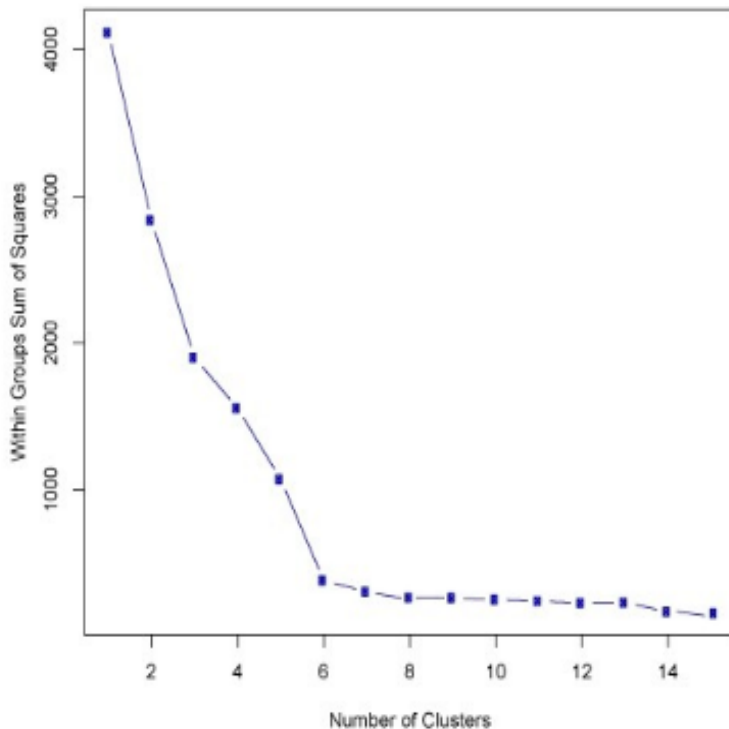
Advantages:

- **Simplicity:** K-means is easy to understand and implement. The core idea of minimizing distances between points and their cluster centers is intuitive.
- **Scalability:** It works well with large datasets due to its efficient iterative approach.
- **Interpretability:** The clusters formed are easy to interpret because they represent groups of similar data points. This makes it easier to understand the structure of the data.
- **Guaranteed Convergence:** K-means is guaranteed to converge on a solution, even if it might not be the optimal one

Disadvantages:

- **Predefined Clusters (k):** You need to specify the number of clusters (k) beforehand. Choosing the right k can be tricky and significantly impacts the results. There are techniques to help decide on k, but it's not always straightforward.
- **Sensitivity to Initialization:** K-means can get stuck in local minima depending on the initial placement of cluster centers. Running it multiple times with different starting positions helps mitigate this but adds complexity.

'k' value



- Elbow method
- Within Group Sum of Square(WGSS)
- Convergence value will be chosen