Experiment No: 2 16|01|2025

Title: K means clustering for data segmentation.

Aim: to implement the k means clustering algorithm for k=2 and k=3 on the data set.

Software used: Matlab R2024b.

Theory:

K-means clustering: K-means clustering is an Unsupervisied machine learning algorithm which groups the unlabeled dataset into different clusters based on feature Similarity It is widely used in various domain such as data mining, image segmentation and pattern recognition. The primary objective of k means is to minimize intra-cluster variance while maximizing inter-cluster separation.

K-means clustering groups data points into k clusters by minimize the variance within each cluster The algorithm works iteratively to assign data points to the closet cluster center and then recalculates the cluster centers based on the mean of the assigned data points.

Algorithm:

The algorithm iterates through the following steps Until convergence.

1. Cluster Assignment : Each data points is assigned to the nearest cluster centroid based on the Encludian distance
2. Centroid Update; The new cluster centroids are computed as the mean of all data points assigned to the respective clusters.

The steps are;

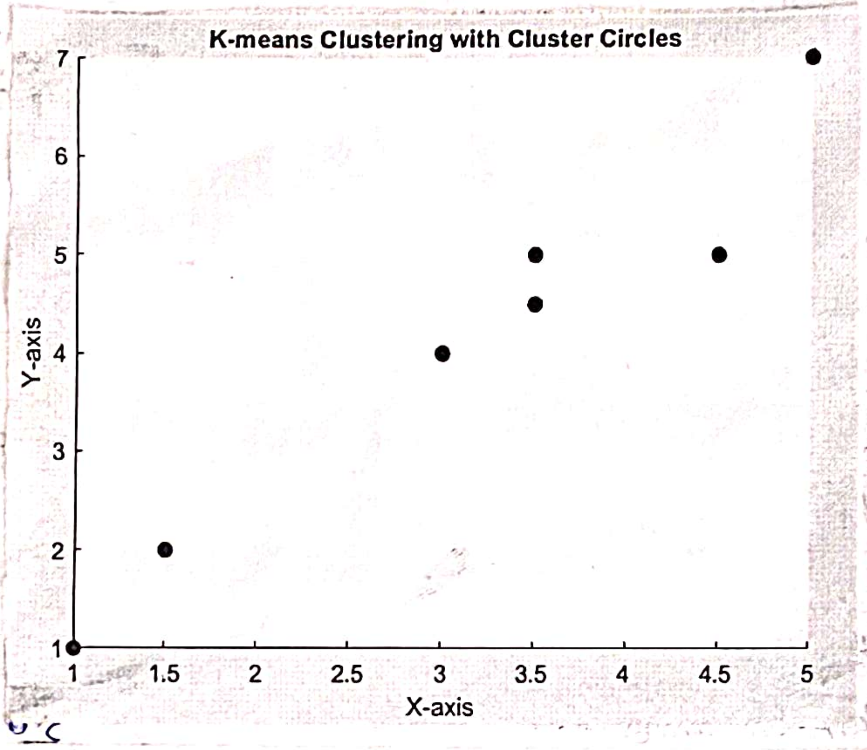
1. Select K initial centroids randomly from the dataset or using heuristic method
2. Cluster Assignment:
   Assign each datapoint to the nearest centroid using the Encludian distance.
3. Centroid update:
   Calculate the new centroids by Computing the mean of all data points within each cluster.
4. Converge Check.
   Repeat steps 2 and 3 Until Centroids no longer change Significantly.

Results

Dataset.

for K=2, K=3

data → (1,1), (1.5,2), (3,4), (5,7), (3.5,5), (4.5,5), (3.5,4.5)



(Visual representation of dataset on 2D plane) for K=2

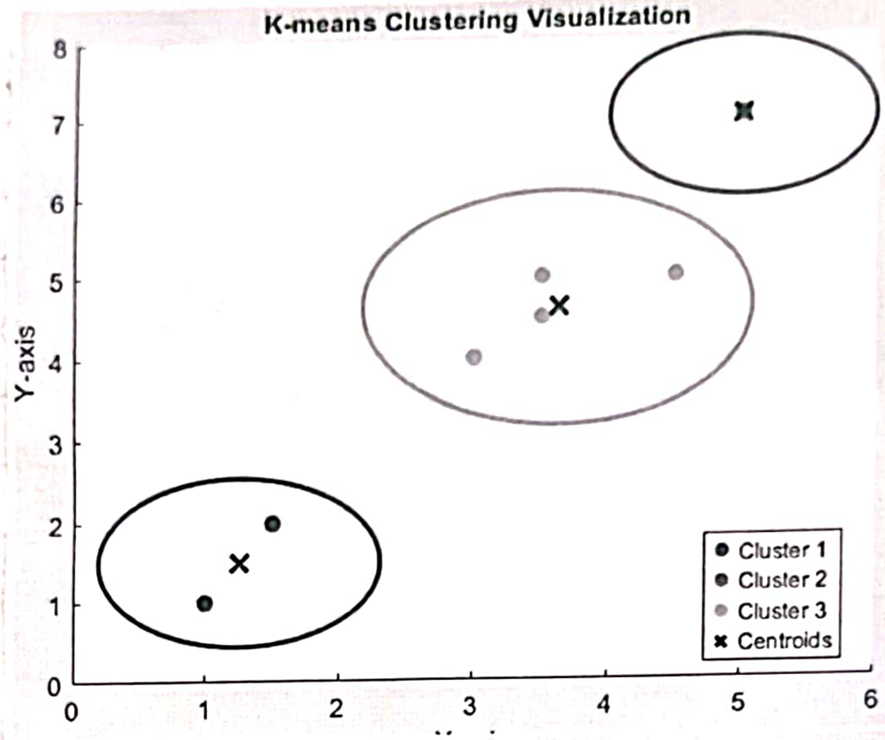
final Centroids → 1→(1.25, 1.5) 2→(3.9, 5.1)

datapoint	dist(centroid1)	dist (cent-2)	finl assigned Centroi
(1,1)	0.55	5.022	1
(1.5,2)	0.55	3.92	1
(3,4)	3.05	1.42	2
(5,7)	6.65	2.19	2
(3.5,5)	4.14	0.41	2
(4.5,5)	4.77	0.60	2
(3.5,4.5)	3.75	0.72	2

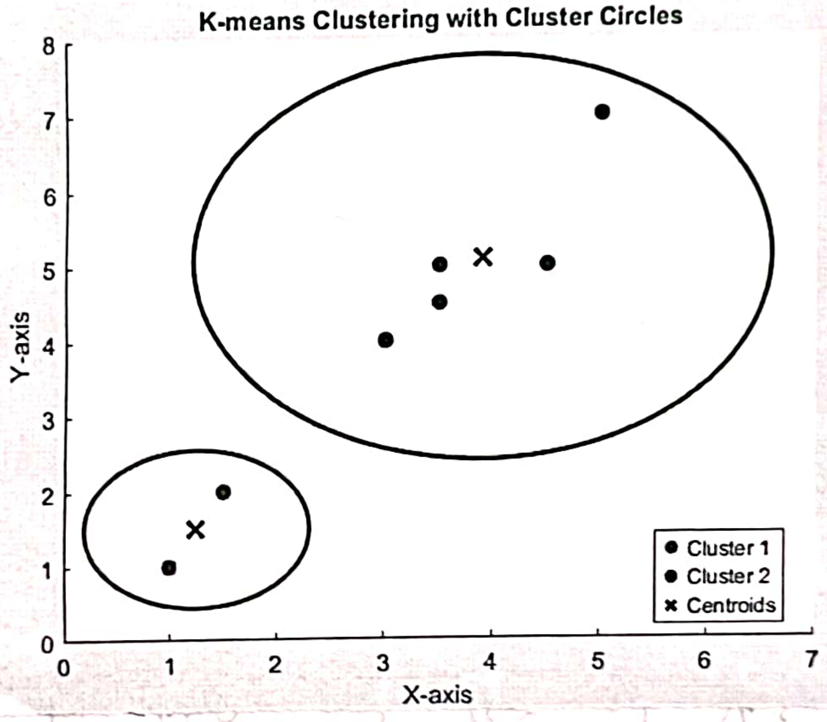
for K=3 centroids [(1.25,1.5), (5,7), (3.625, 4.625)]

data	dist_1	dist_2	dist_3	assign_ centroid
(1,1)	0.55	7.21	4.47	1
(1.5,2)	0.55	6.10	3.37	1
(3,4)	3.05	3.60	0.88	3
(5,7)	6.65	0	2.74	2
(3.5,5)	4.16	2.5	0.39	3
(4.5,5)	4.77	2.06	0.95	3
(3.5,4.5)	3.75	2.91	0.17	3

Results


K-means Clustering Visualization

for K=3


K-means Clustering with Cluster Circles

for K=2

The datapoints are visualized in 2D space with each cluster representing by different Color
Cluster centroids are shown as black cross.('X') on the plot, and data as ('o')
Circles representing cluster boundaries are drawn around centroids with a radius Slightly larger than the farthest data point in each cluster.
Axes are labeled, and a legend is provided for clarity.

Discussion:

→ The algorithm starts with predefined intial centroids based on dataset indices (1,4) & (14,6).

→ Each cluster Contains datapoints that are relatively close to their respective centroids, indicating well-defined Clusters.

→ The datapoints are distinctly grouped with minimal Overlap. Cluster boundaries show moderate Compactness.

→ for K=3, Allow for a more detailed partitioning of the data. Identifies smaller sub groups within the dataset.

→ Each cluster centroid different data characteristics, making it useful for Application needing detailed Segmentations.

→ for K=2, the centroids position themselves in broader region, leading to a larger spread of points in each clusters.

→ points are assigned to Only 2 groups, which might result in some dissimilar data points being grouped together.

→ the two clusters encompass more data points, leading to larger cluster radius.

→ Some points in the middle might be ambiguously assigned.

Conclusions

The dataset has naturaly separation into 3 groups is preferable as it offers better separation and compactness. If the dataset is more uniform for K=2 might suffice for simpler analysis and reduced computational effort. To determine the optimal value of K, techniques such as the elbow method or ~~sithout~~ analysis can be applied.

→ for the less value of K. might group distinct datapoint into same cluster and loss of finer details in the data.

→ for the high value of K clusters might force artificial boundaries.

23/01/24