Assignment -6

Name - Gaurang Patil.
PRN - 1032221535.
RollNo - 20.
T.Y B.tech CSF.

Aim → W.A.P to implement k-Means clustering.

objective → To study k-means clustering algorithm.

Theory →

i) k-means clustering

An i) → k-means clustering is a method that sorts data into a set number of clusters, denoted as k. It start by Randomly placing cluster centres (centroids), then a assign each data point to the closet centroid. The also updates the centroids based on the points assigned to them and respect this until the clusters stabilize.

ii) Steps of k-means clustering Algo.

Ans i) Step 1 → Select the no. of k to decide the number of clusters

Step 2 → Select random k points or centroid (It can be other from the input dataset)

Step 3 → Assign each data point to their closest centroid, which will form the predefined k clusters.

Step 4 → Calculate the variance and place a new centroid of each cluster.

step 5 → Repeat the third step, which means Reassign each datapoint to the new closet centroid of each cluster

step 6 → If any Reassignment occur, then go to step 4 else go to finish

step 7 → The model is ready

3) Objective function of the k-means algorithm.

→ The objective function of the k-means clustering algo is to minimize the within-cluster sum of squares (WCSS)

- Minimize within cluster Scatter - minimize the distance btw data points assigned to be cluster.
- Maximize between-cluster Scatter - Maximize the distance between cluster points assigned to different clusters.

FAQs →

1) How, Determine the k using the Elbow method?

Ans. i) Run k-means for a Range k.

ii) Calculate Intorial for each k

iii) Plot Inertia v/s k on a graph

iv) Find the elbow point. The optimal k is where the Curve bend.

2) Describe the initialization step in the k-means algo.

→ Random Initialization - Select k data points Randomly as the initial centroid.

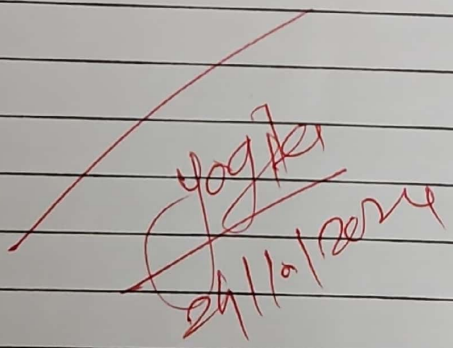
3) k Means the Initialization - Select the first centroid Randomly and Subsequent centroid are chosen based on optimizing the maximum distance from centroid.

Importance of Initial Centroids:

Impact on convergence: Poor initialization can lead to slower convergence.

Risk of local minima: Random centroid might cause the algo to get such in suboptimal cluster.

Consistency of Result: Different initial centroid can yield diff clusters, good initialization helps provide consistent results.



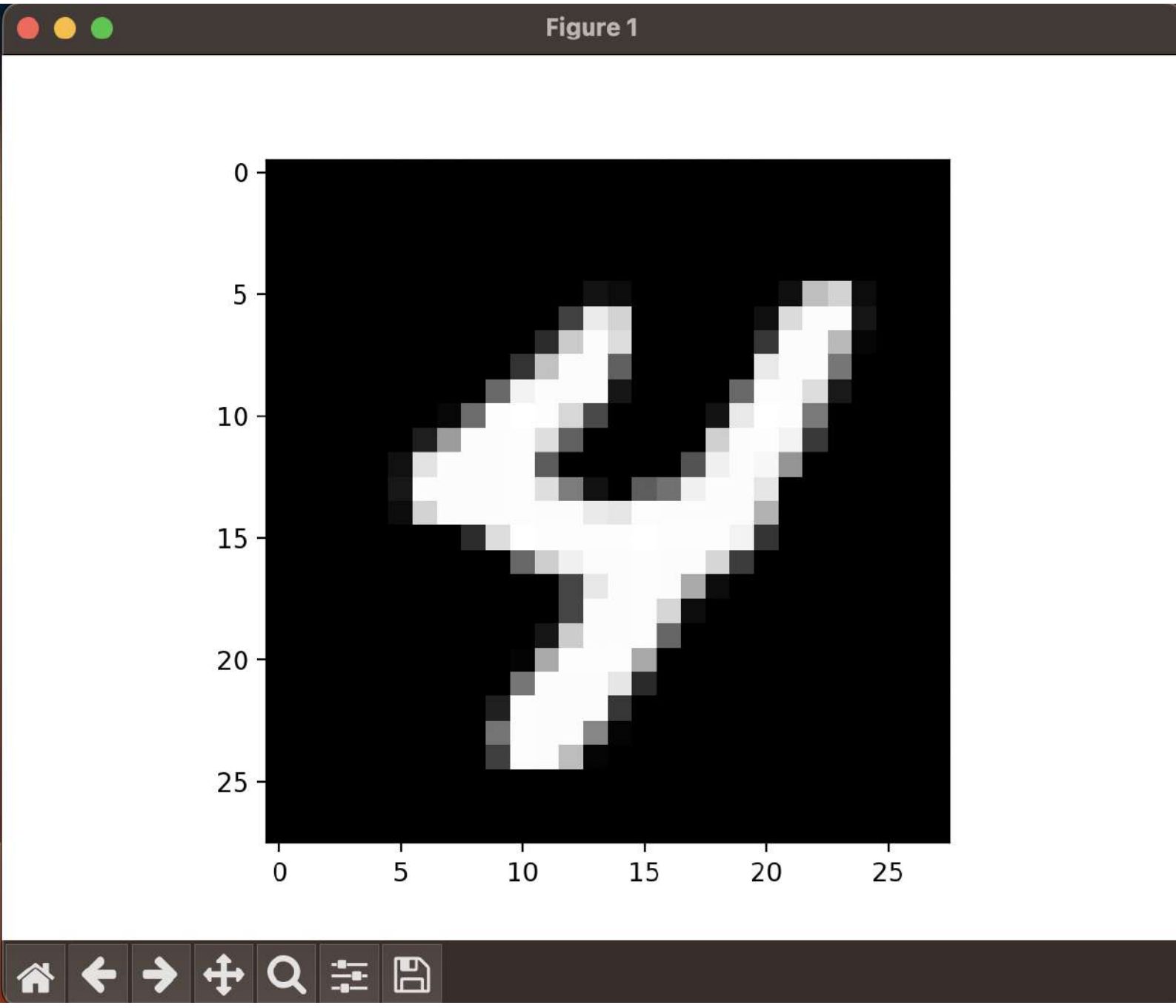


Figure 1
0
5
10
15
20
25
0
5
10
15
20
25

```
Accuracy: 0.8262195121951219
Iteration: 350
Accuracy: 0.828317073170731
Iteration: 360
Accuracy: 0.830780487804878
Iteration: 370
Accuracy: 0.8323658536585365
Iteration: 380
Accuracy: 0.8342682926829268
Iteration: 390
Accuracy: 0.8360731707317073
Iteration: 400
Accuracy: 0.8376829268292683
Iteration: 410
Accuracy: 0.8390731707317073
Iteration: 420
Accuracy: 0.8404390243902439
Iteration: 430
Accuracy: 0.8422926829268292
Iteration: 440
Accuracy: 0.8437317073170731
Iteration: 450
Accuracy: 0.845170731707317
Iteration: 460
Accuracy: 0.8462439024390244
Iteration: 470
Accuracy: 0.8473414634146341
Iteration: 480
Accuracy: 0.8486097560975610
Iteration: 490
Accuracy: 0.8496829268292683
Prediction:  [4]
Label:  4
```