Name: Sai Anish Garapati
UIN: 650208577

1) Question Answering

Q1) K-means is an unsupervised machine learning technique in which we partition the training data into K clusters based on K centroids and the training process involves shifting these centroids to minimize the distance between the centroids and the points in their respective clusters.

Q2) Let initial centroids be
C1 (0, 0); C2 (-1, 0)

For C1 (0, 0)

x1	0	1	0	0	-1
x2	1	0	0	-1	0
Distances (from (0, 0))	1	1	0	1	1

For C2 (-1, 0)

x1	0	1	0	0	-1
x2	1	0	0	-1	0
Distances (from (-1, 0))	$\sqrt{2} = 1.414$	2	1	$\sqrt{2} = 1.414$	0

1) From the above table cluster assignments are as follows:
   a) C1(0, 0) - (0, 1); (1, 0); (0, 0); (0, -1)
   b) C2(-1, 0) - (-1, 0)

2) Updated centroids:
$C1 \leftarrow (\frac{0+1+0+0}{4}, \frac{1+0+0+(-1)}{4}) = (0.25, 0)$
$C2 \leftarrow (-1, 0)$

Q3)
- For every cluster, we initialize means, covariances and mixing coefficients for that cluster
- Then we perform the E step, where for every point in the data, we calculate the responsibility value (γ) of the point with each cluster, which shows the likeliness of the point belonging to that cluster.
- Then we perform the M step, where we update the initialized parameters (means, covariances and mixing coefficients) based on the responsibility values obtained in the E step, to maximize the likelihood function.
- We perform the above steps until the likelihood function or the parameters converge.

Q4) The K-means algorithm is restricted to circular clusters as we consider only the means of clusters as opposed to both means and covariances in the Gaussian mixture model. GMM does well when the distribution of data is not circular or some elliptical shape. GMM does well on data with more complex decision boundaries.

Q5) Let the clusters be C1 {1, 2, 3}; C2 {4, 5}; C3 {7}

- Based on average distance between points in clusters

$$C1\ \&\ C2 = \frac{3+4+2+3+1+2}{6} = \frac{15}{6} = 2.5$$
$$C2\ \&\ C3 = \frac{3+2}{2} = 2.5$$
$$C1\ \&\ C3 = \frac{6+5+4}{3} = 5$$

  Since the average distance between points in clusters C1 & C2 = C2 & C3 < C1 & C3, we can merge the clusters C1 & C2 into a single cluster.
  Resultant clusters
  C1 {1, 2, 3, 4, 5}; C2 {7}

- Based on maximum distance between points in clusters

$$C1\ \&\ C2 - max\{3,4,2,3,1,2\} = 4$$
$$C2\ \&\ C3 - max\{3,2\} = 3$$
$$C1\ \&\ C3 - max\{6,5,4\} = 6$$

  Since the maximum distance between points in clusters C2 & C3 < C1 & C2 < C1 & C3, we can merge the cluster C2 & C3 into a single cluster
  Resultant clusters
  C1 {1, 2, 3}; C2 {4, 5, 7}

- Based on minimum distance between points in clusters

$$C1\ \&\ C2 - min\{3,4,2,3,1,2\} = 1$$
$$C2\ \&\ C3 - min\{3,2\} = 2$$
$$C1\ \&\ C3 - min\{6,5,4\} = 4$$

  Since the minimum distance between points in clusters C1 & C2 < C2 & C3 < C1 & C3, we can merge the clusters C1 & C2 into a single cluster
  Resultant clusters
  C1 {1, 2, 3, 4, 5}; C2 {7}

- Based on distance between means in clusters

$$C1\ \&\ C2 - abs(2-4.5) = 2.5$$
$$C2\ \&\ C3 - abs(4.5-7) = 2.5$$
$$C1\ \&\ C3 - abs(2-7) = 5$$

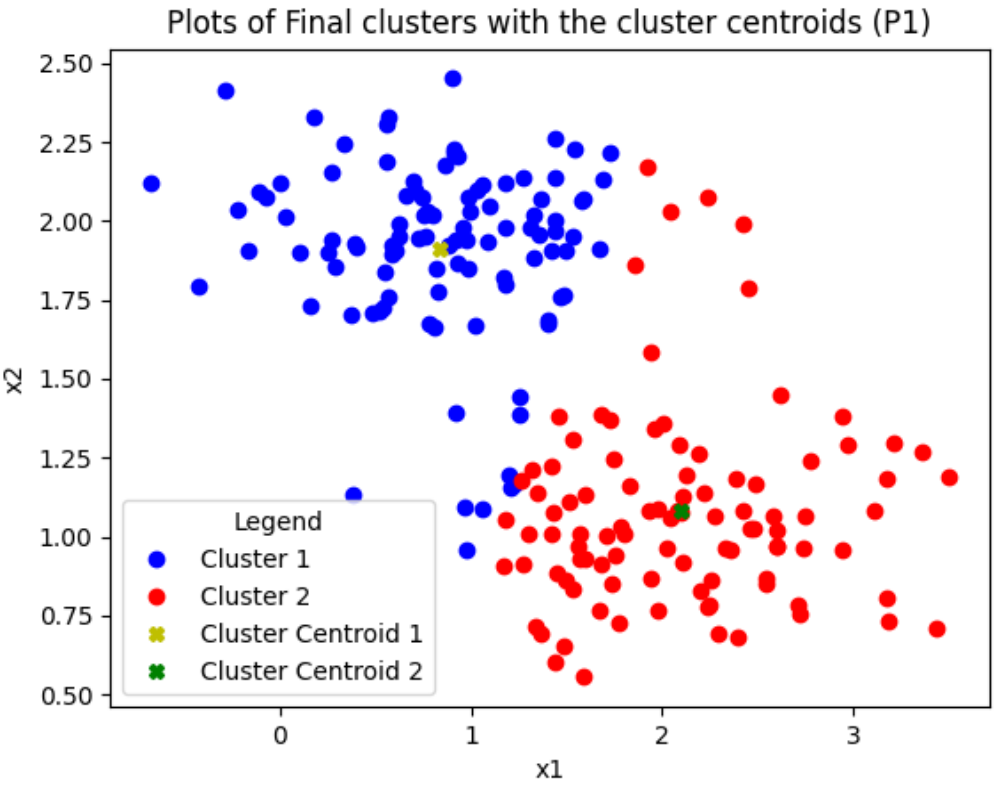
  Since the distance between means in cluster C1 & C2 = C2 & C3 < C1 & C3, we can merge the clusters C1 & C2 into a single cluster
  Resultant clusters
  C1 {1, 2, 3, 4, 5}; C2 {7}

Different cluster similarities do not give the same result from the above cases and in general it is not guaranteed that we will get the same results for different cluster similarity metrics as it varies depending on the distribution of data.

2) Programming Questions

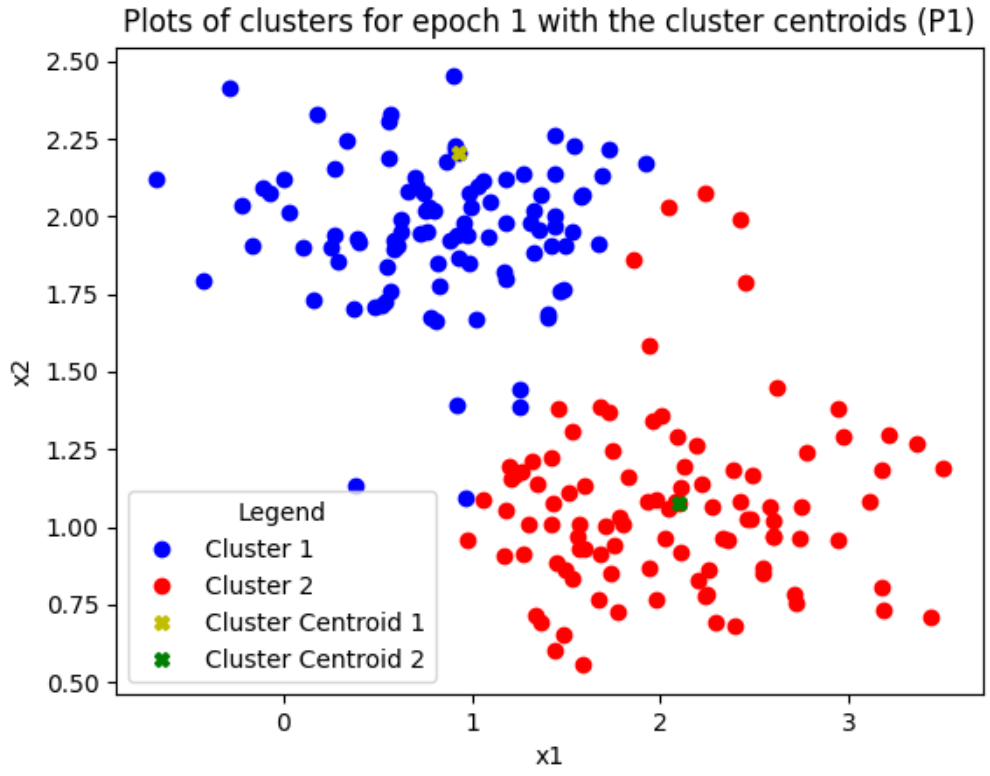
P1)

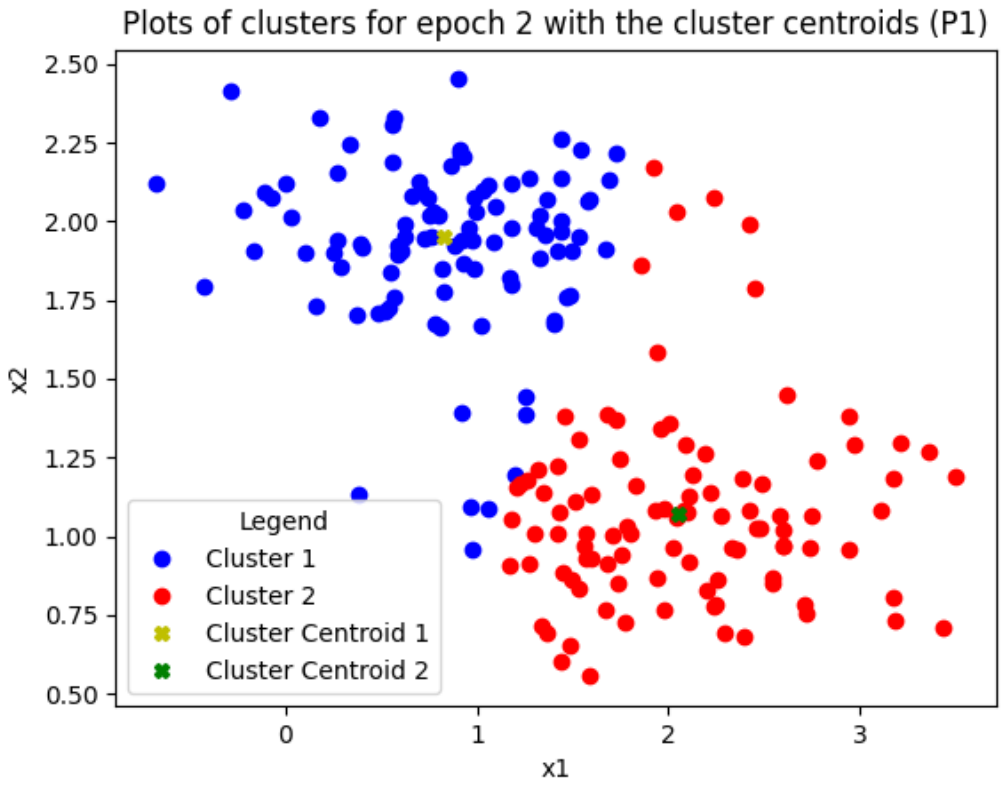
2)

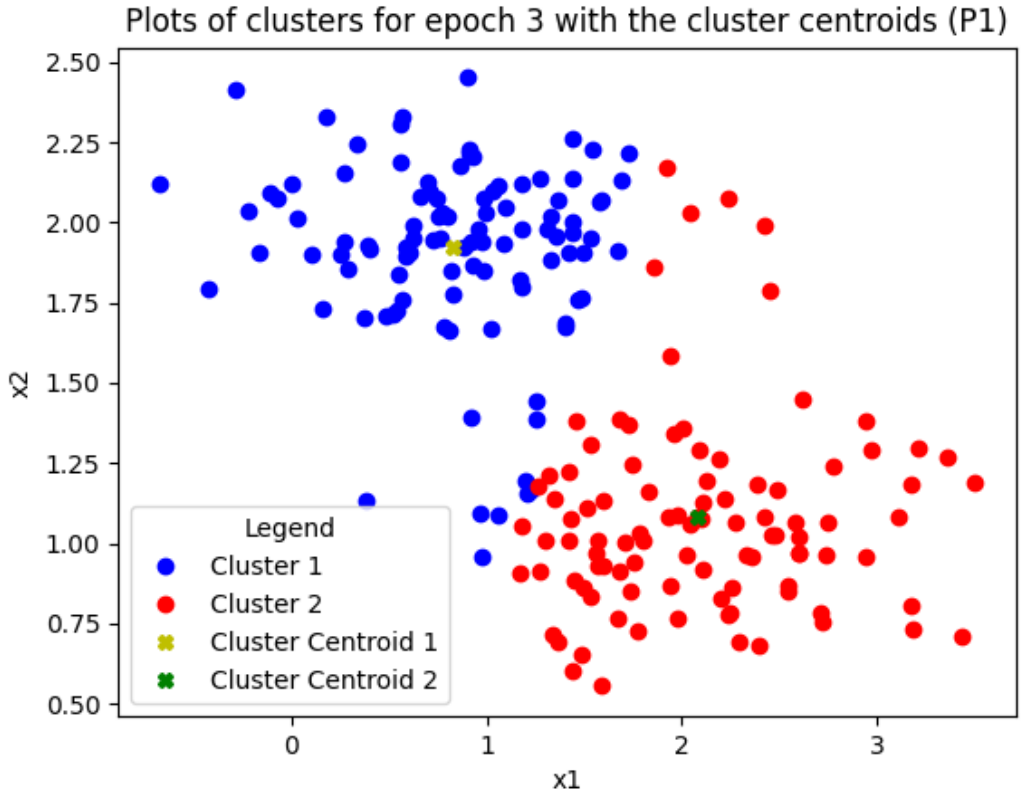


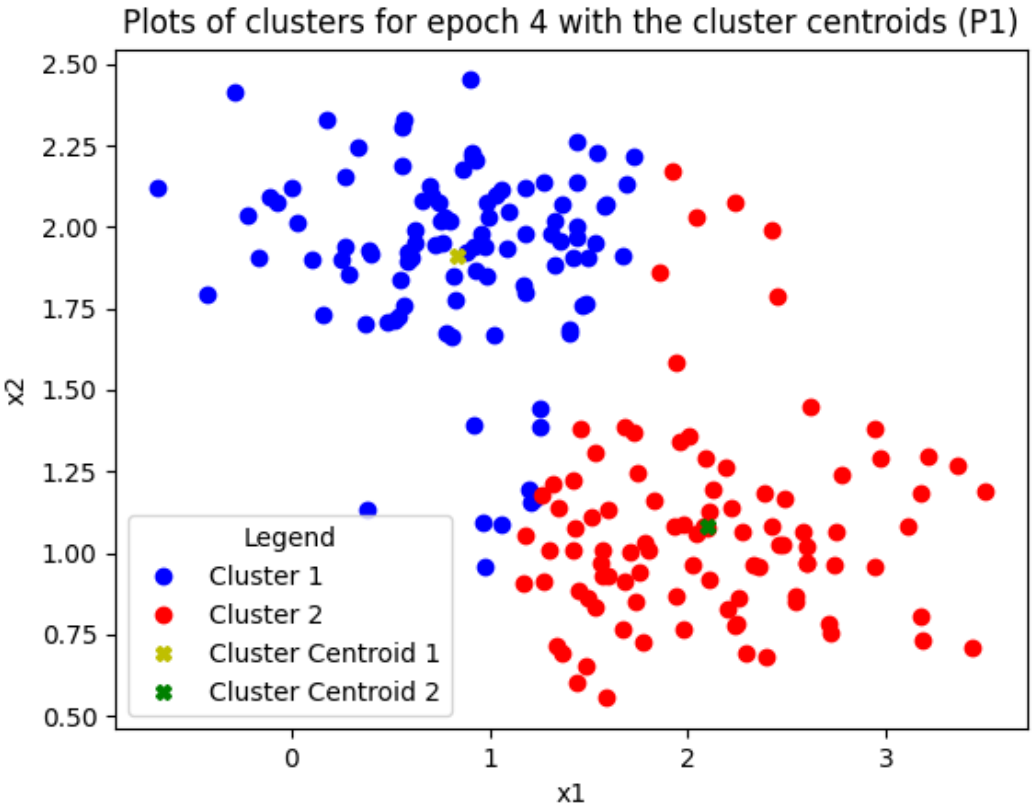
3) K_means objective function for final cluster assignment: 76.7325735871905

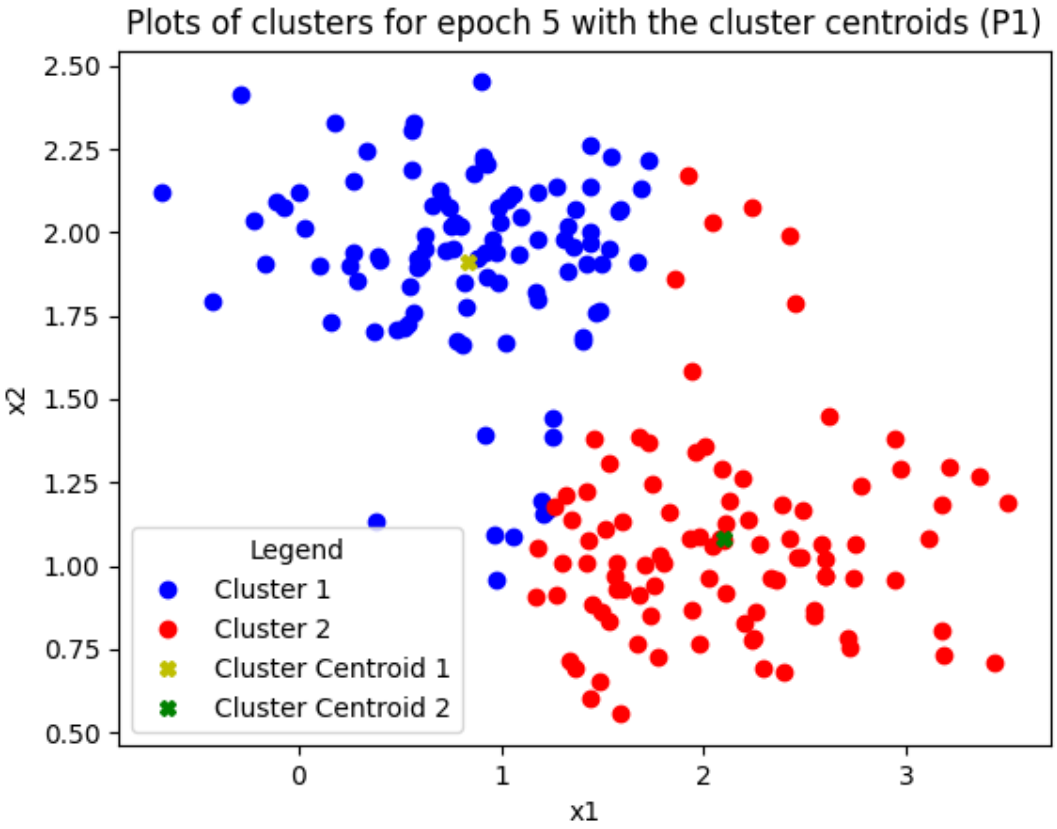
4)




Plots of clusters for epoch 2 with the cluster centroids (P1)
x2
x1
Legend
Cluster 1
Cluster 2
Cluster Centroid 1
Cluster Centroid 2

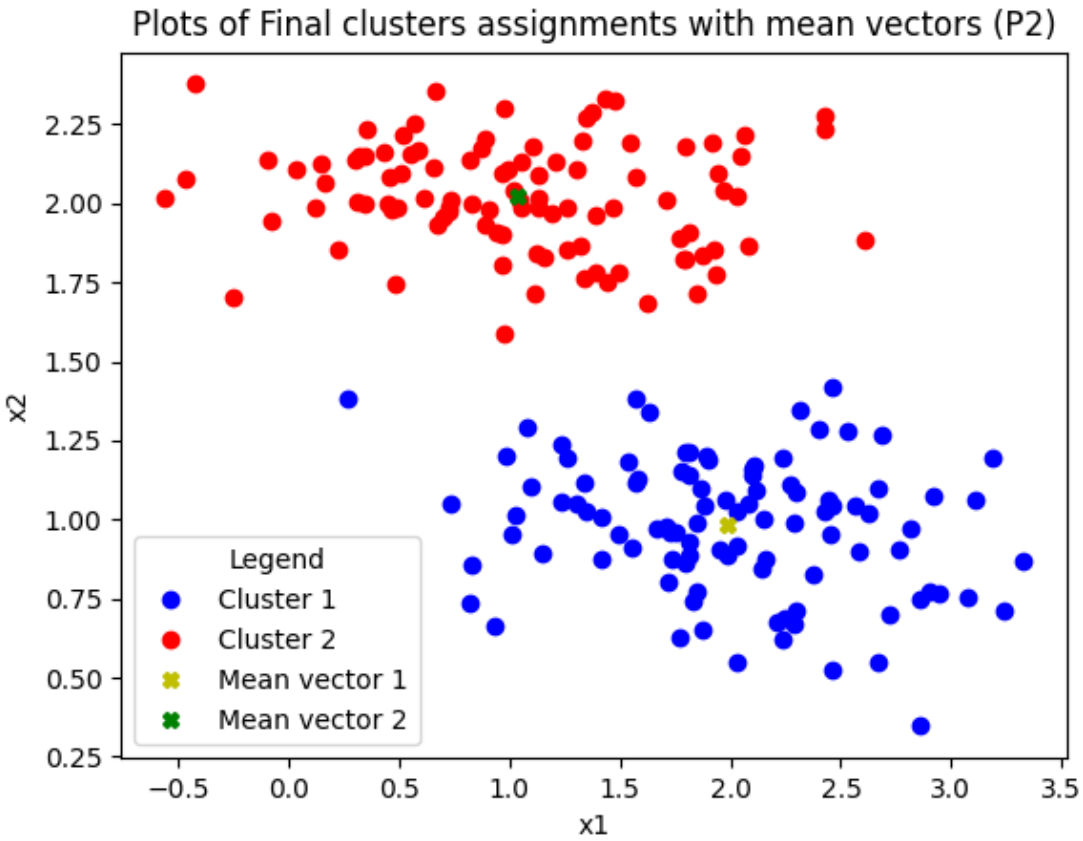

Plots of clusters for epoch 3 with the cluster centroids (P1)
x2
x1
Legend
Cluster 1
Cluster 2
Cluster Centroid 1
Cluster Centroid 2


Plots of clusters for epoch 4 with the cluster centroids (P1)
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
x2
0
1
2
3
x1
Legend
Cluster 1
Cluster 2
Cluster Centroid 1
Cluster Centroid 2


Plots of clusters for epoch 5 with the cluster centroids (P1)
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
x2
0
1
2
3
x1
Legend
Cluster 1
Cluster 2
Cluster Centroid 1
Cluster Centroid 2

5)
K_means objective function for epoch 1 = 84.74314723541929
K_means objective function for epoch 2 = 77.00707834483596
K_means objective function for epoch 3 = 76.79205248879292
K_means objective function for epoch 4 = 76.7325735871905
K_means objective function for epoch 5 = 76.7325735871905

P2)
2)



3) Parameters learned through gmm classifier
Mean: [[1.98909542 0.9817223]
[1.03318008 2.02460907]]

Weights: [0.50177179 0.49822821]

Covariance Matrix: [[[0.37983328 -0.03175925]
[-0.03175925 0.04656601]]

[[0.44774435 -0.01485542]
[-0.01485542 0.0279981]]]

4) responsibility values corresponding to two components: [[0.86998467]] [[0.13001533]]

5) Predicted class value using responsibility values: 0 (which is cluster 1 from the above graph).
Predicted class using gmm classifier: 0 (which is cluster 1 from the above graph).
Yes, the predicted class values for (1.5, 1.5) are same in both the cases.