CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2024)
Quiz - 4

Date of Examination: 12/12/2024 Duration: 40 mins Total Marks: 8 marks

---

Instructions –

- Attempt all questions.
- MCQs may have multiple correct options.
- State any assumptions you have made clearly.
- Standard institute plagiarism policy holds.
- No evaluation without suitable justification.
- 0 marks if the option or justification of MCQs is incorrect.

---

1. **[1 marks]** The Gaussian Mixture Model (GMM) and the k-means algorithm are closely related—the latter is a special case of GMM. The likelihood of a GMM with Z denoting the latent components can be expressed typically as

$$P(X) = \sum_{Z} P(X|Z)P(Z),$$

where $P(X|Z)$ is the (multivariate) Gaussian likelihood conditioned on the mixture component, and $P(Z)$ is the prior on the components.

Such a likelihood formulation can also be used to describe a k-means clustering model. Which of the following statements is/are true? Choose all correct options if there are multiple ones.

(A) $P(Z)$ is uniform in k-means but this is not necessarily true in GMM.

(B) The values in the covariance matrix in $P(X|Z)$ tend towards zero in k-means, but this is not so in GMM.

(C) The values in the covariance matrix in $P(X|Z)$ tend towards infinity in k-means, but this is not so in GMM.

(D) The covariance matrix in $P(X|Z)$ in k-means is diagonal, but this is not necessarily the case in GMM.

Correct Answer: A and B. 1 mark for correct answer and correct reason.
Explanation: The correct options are (A) and (B).
(A): In k-means, P(Z) is uniform because all clusters are equally likely, while in GMM, P(Z) is determined by the mixing coefficients, which are not necessarily uniform.
(B): In k-means, the covariance matrix values conceptually tend to zero because it assumes clusters are concentrated at their centers. In GMM, the covariance matrix explicitly models the spread and orientation of clusters and does not shrink to zero.
Options (C) and (D) are incorrect because the covariance does not tend to infinity in k-means, and k-means does not explicitly use a diagonal covariance matrix.

2. **[1 marks]** Which of the following can act as possible termination conditions in K-Means?

   1. For a fixed number of iterations.
   2. The assignment of observations to clusters does not change between iterations, except for cases with a bad local minimum.
   3. Centroids do not change between successive iterations.
   4. Terminate when RSS (Residual Sum of Squares) falls below a threshold.

(A) 1, 3 and 4

(B) 1, 2 and 3

(C) 1, 2 and 4

(D) All of the above

Correct Answer: D. 1 mark for correct answer and correct reason.
Reason:
All four conditions can be used as possible termination conditions in K-Means clustering:
A. This condition limits the runtime of the clustering algorithm, but in some cases, the clustering quality will be poor because of an insufficient number of iterations.
B. This produces good clustering except for cases with a bad local minimum, but runtimes may be unacceptably long.
C. This also ensures that the algorithm has converged at the minimum.
D. Terminate when RSS falls below a threshold. This criterion ensures that the clustering is of the desired quality after termination. Practically, combining it with a bound on the number of iterations to guarantee termination is a good practice.

3. **[1 marks]** Which of the following is **NOT** a desirable property of a distance measure in clustering?

   (A) Symmetry

   (B) Positivity

   (C) Triangle inequality

   (D) Dependency on labels

Correct Answer: D: Dependency on labels, 1 mark for correct answer and correct reason.
Reason:
Clustering methods, including distance measures, operate in an unsupervised setting with no labels. Dependency on labels contradicts this principle.

4. **[1 marks]** In which of the following cases will K-Means clustering fail to give good results?

   1. Data points with outliers
   2. Data points with different densities
   3. Data points with round shapes
   4. Data points with non-convex shapes

   Options:

   A. 1 and 2

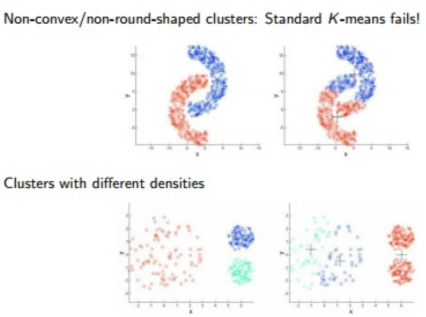
   B. 2 and 3

   C. 2 and 4

   D. 1, 2 and 4

   E. 1, 2, 3 and 4

Correct Answer: D: 1, 2 and 4, 1 mark for correct answer and correct reason.
Reason:
The K-Means clustering algorithm fails to give good results when the data contains outliers, the density spread of data points across the data space differs, and the data points follow non-convex shapes.



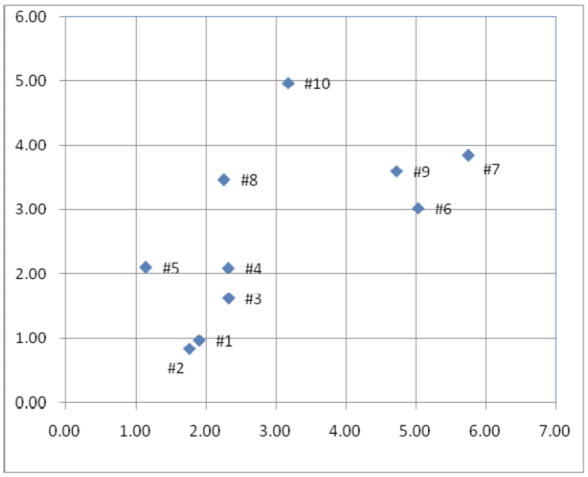


Figure 1: Question 5

Data #	x	y
1	1.90	0.97
2	1.76	0.84
3	2.32	1.63
4	2.31	2.09
5	1.14	2.11
6	5.02	3.02
7	5.74	3.84
8	2.25	3.47
9	4.71	3.60
10	3.17	4.96

Table 1: $\langle x, y \rangle$ Pairs

5. **[4 marks]** Suppose you are given the following $\langle x, y \rangle$ pairs. You will simulate the k-means algorithm and Gaussian Mixture Models (GMM) learning algorithm to identify **two clusters** in the data.

   Suppose you are given the initial assignment of cluster centres as:

   $$\text{Cluster 1: } \#1, \quad \text{Cluster 2: } \#10.$$

   The first data point is used as the centre for the first cluster, and the 10th data point is used as the centre for the second cluster.

   1. Please simulate the k-means algorithm ($k = 2$) for one iteration. What are the cluster assignments after one iteration? Assume k-means uses Euclidean distance.
   2. What are the cluster assignments until convergence?

Correct Answer
Initialization::
Number of clusters (K) = 2, centroid for cluster0 (C1)= (1.90, 0.97) and centroid for cluster1 (C2) = (3.17, 4.96). We use Euclidean distance to find the closest point to centroids.

Record Number	Close to C0 (1.90, 0.97)	Close to C1 (3.17, 4.96)	Assign to Cluster
1 (1.90, 0.97)	dist(1, C1) = 0.0	dist(1, C2) = 4.19	Cluster0
2 (1.76, 0.84)	dist(2, C1) = 0.19	dist(2, C2) = 4.35	Cluster0
3 (2.31,1.63)	dist(3, C1) = 0.78	dist(3, C2) = 3.44	Cluster0
4 (2.31, 2.09)	dist(4, C1) = 1.19	dist(4, C2) = 3.00	Cluster0
5 (1.14, 2.11)	dist(5, C1) = 1.37	dist(5, C2) = 3.50	Cluster0
6 (5.02, 3.02)	dist(6, C1) = 3.73	dist(6, C2) = 2.68	Cluster1
7 (5.74, 3.84)	dist(7, C1) = 4.79	dist(7, C2) = 2.80	Cluster1
8 (2.25, 3.47)	dist(8, C1) = 2.52	dist(8, C2) = 1.75	Cluster1
9 (4.71, 3.60)	dist(9, C1) = 3.85	dist(9, C2) = 2.05	Cluster1
10 (3.17, 4.96)	dist(10, C1) = 4.19	dist(10, C2) = 0.00	Cluster1

Table 2: After Iteration 1 (1 mark)

Thus, we obtain two clusters containing(0.5 each):
Cluster0 {1, 2, 3, 4, 5} and Cluster1 {6, 7, 8, 9, 10}.
For the updated cluster, we calculate centroids:

$$C_0 = \left(\frac{1.90 + 1.76 + 2.32 + 2.31 + 1.14}{5}, \frac{0.97 + 0.84 + 1.63 + 2.09 + 2.11}{5}\right)$$
$$= (1.886, 1.528)$$
$$C_1 = \left(\frac{5.02 + 5.74 + 2.25 + 4.71 + 3.17}{5}, \frac{3.02 + 3.84 + 3.47 + 4.96}{5}\right)$$
$$= (4.178, 3.778)$$

Part 2:(2 Marks) Run until convergence:
Total Iterations it took: 2 or 3
Final Class Assignments

Record Number	Close to C0 (1.886, 1.528)	Close to C1 (4.178, 3.778)	Assign to Cluster
1 (1.90, 0.97)	dist(1, C1) = 0.56	dist(1, C2) = 3.62	Cluster0
2 (1.76, 0.84)	dist(2, C1) = 0.70	dist(2, C2) = 3.81	Cluster0
3 (2.31,1.63)	dist(3, C1) = 0.45	dist(3, C2) = 2.84	Cluster0
4 (2.31, 2.09)	dist(4, C1) = 0.70	dist(4, C2) = 2.52	Cluster0
5 (1.14, 2.11)	dist(5, C1) = 0.95	dist(5, C2) = 3.47	Cluster0
6 (5.02, 3.02)	dist(6, C1) = 3.47	dist(6, C2) = 1.13	Cluster1
7 (5.74, 3.84)	dist(7, C1) = 4.49	dist(7, C2) = 1.56	Cluster1
8 (2.25, 3.47)	dist(8, C1) = 1.98	dist(8, C2) = 1.95	Cluster1
9 (4.71, 3.60)	dist(9, C1) = 3.50	dist(9, C2) = 0.56	Cluster1
10 (3.17, 4.96)	dist(10, C1) = 3.66	dist(10, C2) = 1.55	Cluster1

Table 3: After Convergence (1 mark)

Final Centroids (0.5 each)

$$C_0 = \left(\frac{1.90 + 1.76 + 2.32 + 2.31 + 1.14}{5}, \frac{0.97 + 0.84 + 1.63 + 2.09 + 2.11}{5}\right)$$
$$= (1.886, 1.528)$$
$$C_1 = \left(\frac{5.02 + 5.74 + 2.25 + 4.71 + 3.17}{5}, \frac{3.02 + 3.84 + 3.47 + 4.96}{5}\right)$$
$$= (4.178, 3.778)$$

2 marks for simulate the k-means algorithm (k = 2) for one iteration and 2 marks for "What are the cluster assignments until convergence" Solution.