Pattern Recognition and Machine Learning
End-Semester Exam
Course Instructor : Arun Rajkumar.
Date : April 9, 2023 **Duration** : 12 hours

Name:
Roll Number:
SCORING:The exam contributes 30 points towards your final grade calculation. Each question carries 5 points. Attempt all questions. The best 6 will be considered for grading.

Each question carries 5 points

(a) Consider the scenario where you observe that 8 of your friends $\{f_1, \ldots, f_8\}$ have scored the following marks in a PRML mini-quiz: $\{4, 3, -5, -4, 2, -3, 5, -2\}$ respectively. You would like to cluster your friends based on their marks. You attempt to cluster them into 2 groups using the GMM algorithm. You initialize the algorithm with the following cluster assignment: Cluster 1 contains $\{f_3, f_4, f_7, f_8\}$ and the rest in Cluster 2. You assume that the mean of the clusters are unknown but the standard deviation is known and fixed to be $\sigma = 9$. After *one* step of the GMM algorithm, which of your friends have probabilities greater than 0.5 of being assigned to cluster 1? (You can use the table of density values provided below for your calculations if necessary)

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\mu = -2.5$	0.034	0.036	0.042	0.043	0.039	0.044	0.031	0.044
$\mu = -1.5$	0.036	0.039	0.041	0.042	0.041	0.043	0.034	0.044
$\mu = -0.5$	0.039	0.041	0.039	0.041	0.042	0.042	0.036	0.043
$\mu = 0.5$	0.041	0.042	0.037	0.039	0.044	0.041	0.039	0.043
$\mu = 1.5$	0.043	0.044	0.034	0.037	0.044	0.039	0.041	0.041
$\mu = 2.5$	0.044	0.044	0.031	0.034	0.044	0.037	0.043	0.039

(b) Which of the following matrices *can/cannot* be the covariance matrix of any dataset where the data points are in $\mathbb{R}^2$? For each option, argue why you think Yes or NO.

(a) $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

(b) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$

(d) $\begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}$

(c) You collect height and weight information of all the 100 participants in the PRML class. But instead of computing PCA on the entire dataset, you arbitrarily split it into 2 datasets with 50 points each. One of your friend says that she suffered an error vector of $\begin{bmatrix} -5 \\ 5 \end{bmatrix}$ with respect to the most important direction of the first dataset and an error vector of $\begin{bmatrix} -5 \\ 6 \end{bmatrix}$ with respect to the most important direction of the second dataset. What is the datapoint corresponding to your friend?

(d) Your friend claims that she ran the Llyod's algorithm using the K-means++ initialization technique discussed in class where $K = 3$. Her data looks as follows:

$$\left\{ x_1 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, x_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, x_4 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, x_5 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, x_6 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$$

Answer the following question based on the above data.

i. Your friend says the means selected at initialization were $\{x_6, x_5, x_1\}$ in that order. What is the probability that the procedure resulted in this selection?

ii. Come up with an initialization which is strictly more probable than $\{x_6, x_5, x_1\}$ as per the K-means++ algorithm.

(e) Assume that you have a dataset of 12 points in $\mathbb{R}^3$ and you wish to to perform PCA using a polynomial kernel with degree 2 as discussed in class. What is the exact dimension to which the kernel maps the 3 dimensional datapoints to?

(f) You wish to run spectral clustering on a dataset containing 6 points namely $\{x_1, \ldots, x_6\}$ to partition the data into 2 clusters. You find that the Eigenvectors v_1 and v_2 corresponding the largest and second largest Eigenvalues for the Kernel you chose look as given below:

$$\left\{ v_1 = \begin{bmatrix} 1/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \\ -1/\sqrt{6} \\ 1/\sqrt{6} \\ -1/\sqrt{6} \end{bmatrix}, v2 = \begin{bmatrix} 1/\sqrt{6} \\ -1/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \\ -1/\sqrt{6} \end{bmatrix} \right\}$$

What would be the final clustering of the algorithm if the initialization for the K-means step in Spectral clustering is $\{\mu_1^0 = x_1, \mu_2^0 = x_4\}$? Explain your steps

(g) Assume you are running the EM algorithm for Gaussian Mixture Model and you are in the t-th iteration. The current guess for the parameters is θ^t. You compute λ^{t+1} by maximizing the modified log likelihood function as a function of λ. What can you say about the relation between the log_likelihood(θ^t) and the modified_log_likelihood(θ^t, λ^{t+1})? Justify your answer. Can you use this relation to argue that the EM algorithm must converge?