Introduction to ML (CS771), 2024-2025-Sem-I	Total Marks	25
Quiz 4. November 11, 2024	Duration	45 minutes
Name	Roll No.	

Instructions:

1.	Clearly write your name (in block letters) and roll number in the provided boxes above.
2.	Write your final answers concisely in the provided space. You may use blue/black pen.
3.	We won't be able to provide clarifications during the quiz. If any aspect of some question appears ambiguous/unclear to you, please state your assumption(s) and answer accordingly.

Question 1: Write **T** or **F** for True/False in the box next to each question given below, with a brief (1-2 sentences at most) explanation in the provided space in the box below the question. Marks will be awarded only when the answer (T/F) and explanation both are correct. (**3 x 2 = 6 marks**)

1.1	EM or ALT-OPT will be required for doing MLE for the parameters of a supervised generative classification model with Gaussian class-conditionals.	**F**

In the supervised setting, all the labels are known for the training data, so there are no latent variables. Thus EM or ALT-OPT is not required.

1.2	Projecting D dimensional inputs to a different co-ordinate system with D dimensions using linear PCA will incur zero loss of information.	**T**

In PCA, the total variance captured after the projection is the sum of variances across all the D projected dimensions. We will not incur if we use all the D dimensions.

1.3	Kernel PCA can also be used for doing linear dimensionality reduction.	**T**

Yes, if we use a linear kernel in kernel PCA.

Question 2: Answer the following questions concisely in the space provided below the question.

2.1	In 1-2 sentences, briefly state what distortion error is in the context of dimensionality reduction. Given N inputs $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ an encoder function f and a decoder function g, write down the expression of the total distortion error assuming squared Euclidean distance as the distortion error. **(3 marks)**

Distortion error measures the difference in the original input $\boldsymbol{x}_n$ and its reconstructed (decoded) version from the encoded version $\boldsymbol{z}_n = f(\boldsymbol{x}_n)$. Its expression of $\sum_{n=1}^{N} \|\boldsymbol{x}_n - g(f(\boldsymbol{x}_n)\|^2$

2.2	Given the top K eigenvectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_K$ computed using a set of inputs $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$, write down the expression of the K-dim projection $\boldsymbol{z}_n$ of any input $\boldsymbol{x}_n \in \mathbb{R}^D$. **(2 marks)**

$\boldsymbol{z}_n = [z_{n1,} z_{n2}, \dots, z_{nK}]$ where $z_{nk} = \boldsymbol{w}_k^\top \boldsymbol{x}_n$ for $k = 1,2, \dots, K$, or $\boldsymbol{z}_n = \boldsymbol{W}^\top \boldsymbol{x}_n$ where $\boldsymbol{W}$ is the $D \times K$ matrix with its columns consisting of the top K eigenvectors

2.3	Using appropriate notation, write down the general expression of the loss function (assuming squared Euclidean distance based loss) for a matrix factorization problem for an $N \times M$ matrix $\boldsymbol{X}$ that may have some missing entries and write the expression required for the task of matrix completion, i.e., for predicting the value of some missing entry X_{ij}. **(4 marks)**

Denoting the (row,column) indices of all the observed entries by the set Ω, the loss function can be expressed as

$$\mathcal{L} = \sum_{(i,j)\in\Omega} \left(X_{ij} - \boldsymbol{u}_i^\top \boldsymbol{v}_j\right)^2$$

where $\boldsymbol{u}_i$ $(i = 1,2,\dots,N)$ and $\boldsymbol{v}_j$ $(j = 1,2,\dots,M)$ are the unknowns to be estimated by minimizing the above loss functions. Once we have estimated these unknowns, the predicted value of some missing entry can be computed as $X_{ij} \approx \boldsymbol{u}_i^\top \boldsymbol{v}_j$,

2.4	For a K-component Gaussian mixture model with parameters $\Theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, starting with the expression of the joint distribution $p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta)$, show the steps that to obtain $p(\boldsymbol{x}_n|\Theta)$ and write down the final expression of $p(\boldsymbol{x}_n|\Theta)$. **(3 marks)**

$p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta) = p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)p(\boldsymbol{z}_n|\Theta)$. Since z_n is discrete with K possible values, we can obtain $p(\boldsymbol{x}_n|\Theta)$ using the sum rule as $p(\boldsymbol{x}_n|\Theta) = \sum_{k=1}^N p(\boldsymbol{x}_n|\boldsymbol{z}_n = k, \Theta)p(\boldsymbol{z}_n = k|\Theta)$

$p(\boldsymbol{x}_n|\boldsymbol{z}_n = k, \Theta) = \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)$ and $p(\boldsymbol{z}_n = k|\Theta) = \pi_k$

Therefore $p(\boldsymbol{x}_n|\Theta) = \sum_{k=1}^N \pi_k \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)$

2.5	For a latent variable model with data $\boldsymbol{X}$, latent variables $\boldsymbol{Z}$, and parameters Θ, EM computes the MLE of Θ by solving $\Theta_{MLE} = \text{argmax}_\Theta f(\boldsymbol{X}|\Theta)$. Write down the general expression of $f(\boldsymbol{X}|\Theta)$, clearly specifying and defining the various terms in the expression. **(3 marks)**

$\Theta_{MLE} = \mathbb{E}_{\boldsymbol{p}(\boldsymbol{Z}|\boldsymbol{X},\Theta)}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)]$ where $\boldsymbol{p}(\boldsymbol{Z}|\boldsymbol{X}, \Theta)$, w.r.t. which the expectation is taken, is the conditional posterior of the latent variables $\boldsymbol{Z}$ given data $\boldsymbol{X}$ and the parameters Θ.

2.6	For an MLP with L hidden layers and nonlinearity g in each hidden layer, clearly and briefly write down the expressions that will be used to compute the real-valued output $\hat{y}_n$ for an input $\boldsymbol{x}_n \in \mathbb{R}^D$. You may use any terms/notation necessary for these expressions. **(4 marks)**

This MLP will compute a sequences of nonlinear transformations:

$h_n^{(\ell)} = g\left(W^{(\ell)\top} \boldsymbol{h}_n^{(\ell-1)}\right)$ for $\ell = 1,2,3,\dots,L$, and $\boldsymbol{h}_n^{(0)} = \boldsymbol{x}_n$

where $h_n^{(\ell)}$ is of size $K_\ell \times 1$, $W^{(\ell)}$ is of size $K_{\ell-1} \times K_\ell$ and $K_0 = D$

Finally, the MLP computes the real-valued output $\hat{y}_n$ using a linear model as $\hat{y}_n = \boldsymbol{v}^\top \boldsymbol{h}_n^{(L)}$