ECE 461/661 Spring 2025 Mini-Exam 3

Introduction to Machine Learning for Engineers

Prof. Gauri Joshi and Prof. Carlee Joe-Wong

Wed, April 16, 2025, 9:00am - 9:55am PT/12:00pm - 12:55pm ET/6:00pm - 6:55pm CAT

Name ______________________________ **Andrew ID** ________________________

Instructions

a. If a problem asks you which of its choices is TRUE, you should treat choices that may be either true or false as FALSE.

b. **Unless otherwise stated, only one option is correct in each multiple-choice question.** No partial credit will be given for true/false questions or multiple choice questions where only one option is correct.

c. For descriptive questions, make sure to explain your answers and reasoning. We will give partial credit for wrong answers if portions of your reasoning are correct. Conversely, correct answers accompanied by incomplete or incorrect explanations may not receive full credit.

d. You are allowed one **physical, handwritten, single-sided** US-letter or A4 sized cheat sheet. No other notes or material (aside from blank pieces of scratch paper) may be used.

e. You may only use a pen/pencil, eraser, and scratch paper. The backside of each sheet in the exam can also be used as scratch paper. If you do not wish for us to grade your scratch work, please clearly indicate which parts of your work we should ignore.

f. Calculators are not permitted and not necessary. No other electronic devices such as phones, tablets or laptops can be used during the exam.

g. If you would like to ask a clarification question during the exam, raise your hand and an instructor or TA will come over. Note that we will not help you answer the questions but can give clarifications.

Problem	**Type**	**Points**
1-4	True/False	4
5-9	Multiple Choice	10
10	Descriptive	6
11	Honor Pledge	0
Total		20

1 True or False (4 points)

Problem 1: *[1 points]* In asynchronous distributed SGD (stochastic gradient descent), the parameter server updates the global model as soon as any worker finishes and sends its gradient. That worker then receives the updated global model from the parameter server, while the other workers asynchronously continue their gradient computations. The update rule at the parameter server at the tth global update is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta g(\mathbf{w}_{\tau(t)}) \tag{1}$$

where $\mathbf{w}_s$ denotes the model parameters after the sth global update, $g(\cdot)$ denotes the gradient of the model on the most recent worker's data, and $\tau(t) \leq t$ is the index of the model at which the most recent worker evaluated the gradient. A small value of $\tau(t)$ means that the gradient is more stale.

○ True

○ False

Solution: True. There is a larger gap $t - \tau(t)$ between the parameter server's model and the worker's model.

Problem 2: *[1 points]* The features selected by Principal Component Analysis (PCA), i.e., the principal components, are linear combinations of the original features.

○ True

○ False

Solution: True

Problem 3: *[1 points]* Generating n tokens using an autoregressive decoder model requires a single forward pass of the model.

○ True

○ False

Solution: False, to generate n tokens using an autoregressive decoder model would require n forward passes of the model

Problem 4: *[1 points]* In a Gaussian Mixture Model (GMM), the Expectation-Maximization (EM) algorithm guarantees convergence to the global maximum of the log-likelihood function regardless of initialization.

○ True

○ False

Solution: False. The EM algorithm for Gaussian Mixture Models is guaranteed to monotonically increase the log-likelihood at each iteration and converge to a *local maximum* (or saddle point), but not necessarily the *global maximum*. Its final result depends heavily on the initialization of the parameters.

2 Multiple Choice (10 points)

Problem 5: *[2 points]* Consider a system of m worker nodes and a parameter server. In synchronous SGD, the m worker nodes fetch the current version of the model from the parameter server, compute a mini-batch SGD gradient and send the gradients back to the parameter server. The parameter server averages these gradients and updates the model to complete 1 iteration of the synchronous SGD algorithm.
To save the communication cost, in local-update SGD, the m worker nodes perform τ local SGD iterations and their updated models are averaged by the parameter server in every round. One round consists of τ iterations.
Suppose each local gradient computation takes constant time Y and the two-way communication with the parameter server takes constant time D. What is the ratio of the per-iteration runtimes of synchronous and local-update SGD?

○ $\frac{T_{\text{sync}}}{T_{local}} = \frac{D+Y}{D+\tau Y}$

○ $\frac{T_{\text{sync}}}{T_{local}} = \frac{D+Y}{\tau D+Y}$

○ $\frac{T_{\text{sync}}}{T_{local}} = \frac{D+Y}{D/\tau+Y}$

○ $\frac{T_{\text{sync}}}{T_{local}} = \frac{D+Y}{D+Y/\tau}$

Solution: C: $\frac{T_{\text{sync}}}{T_{local}} = \frac{D+Y}{D/\tau+Y}$
The time taken to complete one iterations using synchronous SGD is $T_{\text{sync}} = Y + D$.
The time taken to complete τ iterations using local-update SGD is $\tau Y + D$, and this the per-iteration runtime is $T_{\text{local}} = Y + \frac{D}{\tau}$.

Problem 6: *[2 points]* Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ be a set of n data points in d-dimensional space. The k-means algorithm aims to partition X into k disjoint clusters $C_1, C_2, \ldots, C_k$ such that the following objective function is minimized:

$$\mathcal{J} = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|^2$$

where μ_i is the centroid of cluster C_i. Which of the following statements about k-means is/are true? **More than one option can be correct.**

□ The k-means algorithm always converges to the global minimum of $\mathcal{J}$ in a finite number of steps, regardless of initialization.

□ The centroid μ_i for each cluster C_i is updated to the median of all points in C_i at every iteration.

□ The k-means algorithm guarantees a non-increasing objective value $\mathcal{J}$ at each iteration and converges to a local minimum.

□ None of the above.

Solution:

- **False.** The k-means algorithm is sensitive to the initial positions of the centroids. It uses a greedy approach and may converge to a local minimum of $\mathcal{J}$, not necessarily the global minimum.
- **False.** k-means updates the centroid μ_i as the mean of all points in cluster C_i. Using the median would correspond to the k-medians algorithm, which minimizes the sum of absolute deviations.

- **True.** The k-means algorithm alternates between assigning points to the nearest centroid and updating centroids as the mean of the assigned points. Each step guarantees a non-increasing objective function $\mathcal{J}$. Since there are finitely many possible clusterings and $\mathcal{J}$ decreases or stays constant at each step, the algorithm converges in a finite number of iterations to a local minimum.

- **False.**

Problem 7: *[2 points]* Suppose that you wish to partition a dataset $\mathcal{D} \subset \mathbb{R}^2$ into $k = 2$ clusters. You decide to run both the k-means (with Euclidean distances) and Gaussian mixture model (GMM) algorithms on your dataset and find that they yield identical cluster assignments. That is, all data points have the same cluster labels under the k-means and GMM clustering methods.
Which of the following datasets would yield this result? **More than one option may be correct.**

□ $\mathcal{D} = \left\{ \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix} \right\}$

□ $\mathcal{D} = \left\{ \begin{bmatrix}0\\0.1\end{bmatrix}, \begin{bmatrix}0.1\\0\end{bmatrix}, \begin{bmatrix}0\\0.2\end{bmatrix}, \begin{bmatrix}-0.1\\0\end{bmatrix}, \begin{bmatrix}0\\0.3\end{bmatrix}, \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0\\0.4\end{bmatrix}, \begin{bmatrix}0\\0.5\end{bmatrix}, \begin{bmatrix}-0.2\\0\end{bmatrix}, \begin{bmatrix}0.2\\0\end{bmatrix} \right\}$

□ $\mathcal{D} = \left\{ \begin{bmatrix}0.1\\0\end{bmatrix}, \begin{bmatrix}0\\10.1\end{bmatrix}, \begin{bmatrix}-0.1\\0\end{bmatrix}, \begin{bmatrix}0\\9.9\end{bmatrix}, \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0\\10\end{bmatrix} \right\}$

□ This is an impossible result; no dataset will give the same cluster labels for GMMs and k-means.

Solution:

A is correct, since the dataset consists of two points. k-means will assign cluster centers at $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$, and GMMs will assign Gaussian means at the same centers. This will lead to all points $\begin{bmatrix}1\\0\end{bmatrix}$ being assigned to one cluster and all points $\begin{bmatrix}0\\1\end{bmatrix}$ assigned to another.

B is incorrect, since GMMs should learn two clusters along the vertical and horizontal axes. k-means cannot do so since we use Euclidean distances in k-means.

C is correct, since the data points are divided into well-separated clusters with no overlap.

D is incorrect; see the examples above.

Problem 8: *[2 points]* Consider a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where each data point $\mathbf{x}_n$ is associated with a hidden label $\mathbf{z}_n$. Which of the following statements is/are true about the expectation-maximization (EM) algorithm? **More than one option may be correct.**

□ The EM algorithm's objective is to maximize the complete log-likelihood $\ell(\boldsymbol{\theta}) = \sum_n \sum_{\mathbf{z}_n} \log p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\theta})$ of dataset $\mathcal{D} = \{\mathbf{x}_n\}$ with respect to the set of parameters $\boldsymbol{\theta}$.

□ The EM algorithm's objective is to maximize the incomplete log-likelihood $\ell(\boldsymbol{\theta}) = \sum_n \log p(\mathbf{x}_n | \boldsymbol{\theta})$ of dataset $\mathcal{D} = \{\mathbf{x}_n\}$ with respect to the set of parameters $\boldsymbol{\theta}$.

□ In the E-step at round t, we set the distribution of the hidden labels $\mathbf{z}_n$, $q_t(\mathbf{z}_n)$, to $p(\mathbf{z}_n | \mathbf{x}_n; \boldsymbol{\theta}_t)$, so that $\ell(\boldsymbol{\theta}_t) = \sum_n \sum_{\mathbf{z}_n} q_t(\mathbf{z}_n) \log p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\theta}_t)$.

□ In the M-step, we choose the parameters θ that maximize an upper bound on the objective function $\ell(\boldsymbol{\theta})$.

Solution: B, C

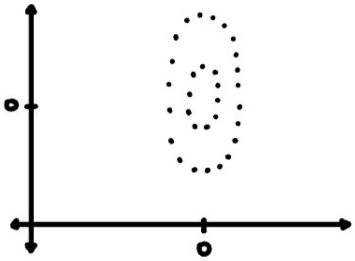
A: False, since we do not observe $\mathbf{z}_n$ it is impossible to compute and maximize the complete log-likelihood

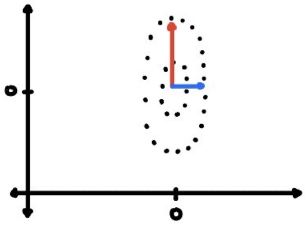
B: True, since we only observer $\mathbf{x}_n$, we can only maximize this incomplete log-likelihood $\ell(\boldsymbol{\theta}) = \sum_n \log p(\mathbf{x}_n | \boldsymbol{\theta})$ with respect to θ

C: True. In the E-step we set $q(\mathbf{z}_n)$ to the posterior distribution $p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\theta}_t)$

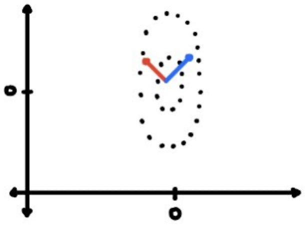
D: False, In the M-step, we choose the parameters θ that maximize a **lower bound** on the objective function

Problem 9: *[2 points]* Which of the following depictions correctly draws the first and second principal components on the below dataset in $\mathbb{R}^2$? Each black dot represents a data point. **Note: the first principal component is drawn in red and the second principal component is drawn in blue**

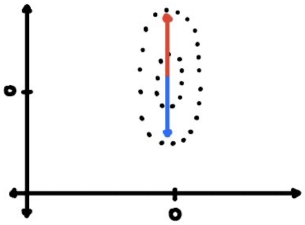




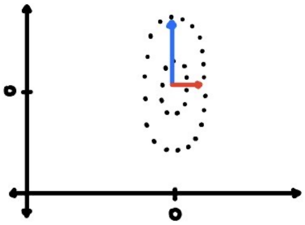
○



○



○



○

Solution: A. the vertical component explains a majority of the variance in the dataset and the horizontal component explains the rest of the variance

3 Descriptive (6 pts)

Problem 10: *[6 points]* Consider a dataset $\mathcal{D}$ on which we wish to run PCA. Each data point $\mathbf{x} \in \mathcal{D}$ is 4-dimensional, and we wish to transform this data into $k = 2$-dimensional representations. Suppose that we find that the covariance matrix of $\mathcal{D}$ has the eigenvalue decomposition

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

a. *[2 points]* Find the representation of the vector $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ in terms of the first two principal components.

Solution: We can read off the principal components (in descending order of significance) from the eigenvalue decomposition of the covariance matrix as

$$\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Thus, the first two principal components are $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix}$. We therefore find the representation

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \approx \left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} + \left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} + \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix}.$$

Full credit should be given for the answer $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$.

b. *[2 points]* Recall that we can use the two-dimensional PCA representation of a vector $\mathbf{x} \in \mathbb{R}^4$ to estimate or reconstruct $\mathbf{x}$. Find all vectors $\mathbf{v} \in \mathbb{R}^4$ whose reconstructed estimates from their two-dimensional PCA representations have an error of 0, i.e., which equal their reconstructed estimates.

Remember to show your work.

Solution: We can rewrite $\mathbf{v}$ in terms of all four principal components as

$$\left(\mathbf{v}^T \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} + \left(\mathbf{v}^T \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix} + \left(\mathbf{v}^T \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right) \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \left(\mathbf{v}^T \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right) \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

In order for this expression to equal $\mathbf{v}$, we must have $\mathbf{v}^T \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \mathbf{v}^T \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = 0$, i.e., $\mathbf{v} = \begin{bmatrix} a \\ b \\ 0 \\ 0 \end{bmatrix}$ for some real values a, b.

c. *[2 points]* Now suppose that all data points in $\mathcal{D}$ belong to the set you identified in part (b), i.e., they have no error when represented in terms of the first two principal components. You decide to apply k-means clustering, with Euclidean distances, to both the original data points and their two-dimensional PCA representations. Will we obtain the same cluster assignments and centers? Briefly explain your answer in 1-2 sentences.

Solution: Using the representation in terms of the first two PCA components corresponds to an orthonormal change-of-basis on our data points (in fact, the principal components in this problem would correspond to a 45-degree rotation in two-dimensional space), since the principal components are orthonormal. This will preserve Euclidean distances, and thus the assignment of data points to clusters. The cluster centers will be the same up to this change of basis.

WRITE YOUR ANSWER TO PROBLEM 10 BELOW

CONTINUE YOUR ANSWER TO PROBLEM 10, IF NEEDED

CONTINUE YOUR ANSWER TO PROBLEM 10, IF NEEDED

4 Honor Pledge

Problem 11: *[0 points]* To affirm that you did not cheat on the exam, please write out the below statement. Sign your name beneath it. **Failure to do so will be taken as a sign that you have cheated on the exam.**

I pledge my honor that I neither gave nor received unauthorized assistance on this examination.