ECE 461/661 Spring 2025 Mini-Exam 1

Introduction to Machine Learning for Engineers

Prof. Gauri Joshi and Prof. Carlee Joe-Wong

Mon, February 10, 2025, 9:00am - 9:40am PT/12:00pm - 12:40pm ET/7:00pm - 7:40pm CAT

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 multiple-choice or true/false questions.

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, 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 permitted but not necessary. If you choose to use a calculator, it must be a standalone one. 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**
1-5	True/False	7
6-8	Multiple Choice	6
9	Descriptive	4
10	Descriptive	3
11	Honor Pledge	0
Total		20

1 True or False (7 points)

Problem 1: *[1 points]* The eigenvalues of $\mathbf{A}^\top\mathbf{A}$ or $\mathbf{A}\mathbf{A}^\top$ are squares of the singular values of matrix $\mathbf{A}$.

- ◯ True
- ◯ False

Solution: TRUE. This follows directly from the definition of singular values.

Problem 2: *[1 points]* Suppose you wish to estimate the probabilities of rolling each side of a given six-sided die, based on 100 observed outcomes from rolling it. You may assume the 100 rolls are independent. Based on your knowledge of the manufacturing process for this die, you also have a prior distribution π for the probabilities of rolling each side. Using MAP with the prior distribution π will produce a distribution more similar to the true distribution than using MLE.

- ◯ True
- ◯ False

Solution: FALSE. In this question, we have no information about how well the prior distribution reflects the true parameters, so we do not know if MLE or MAP will produce better parameter estimates.

Problem 3: *[1 points]* Suppose you flipped a coin twice, and that the first time it landed on HEADS and the second time on TAILS. You model the HEADS outcome as having a Bernoulli distribution with parameter π. Then the log-likelihood of your observed data is $\log(\pi) + \log(1-\pi)$.

- ◯ True
- ◯ False

Solution: TRUE. The likelihood of observing first heads and then tails is $\pi(1-\pi)$. Taking the logarithm, we obtain $\log(\pi) + \log(1-\pi)$.

Problem 4: *[1 points]* Given a feature matrix $\mathbf{X}$ and label vector $\mathbf{y}$, linear regression finds a weight vector $\mathbf{w}^*$ such that $\mathbf{X}\mathbf{w}^* = \mathbf{y}$.

- ◯ True
- ◯ False

Solution: FALSE. This is generally not true. Since $\mathbf{y} \notin \text{col}(\mathbf{X})$ is often true, linear regression finds $\mathbf{w}^*$ such that $\mathbf{X}\mathbf{w}^* = \hat{\mathbf{y}}$, where $\hat{\mathbf{y}}$ is the vector in the column space of $\mathbf{X}$ that is closest to $\mathbf{y}$.

Problem 5: Consider a regression problem with a feature vector $\mathbf{x} = [x_1, x_2, x_3] \in \mathbb{R}^3$ and target $y \in \mathbb{R}$, where we find the model $\mathbf{w}$ that minimizes the residual sum of squares function (without adding any regularizer). We consider two different feature transformations of $\mathbf{x}$: $\boldsymbol{\phi}_A = [x_1, x_2, x_3, x_1^2, x_2^2, x_3^2]$, $\boldsymbol{\phi}_B = [\frac{x_1+x_3}{2}, \frac{x_1+x_2}{2}, \frac{x_2+x_3}{2}]$. Indicate whether the following statements are true or false:

a. *[1 points]* The model $\mathbf{w}_A$ learnt using $\boldsymbol{\phi}_A$ has higher bias and lower variance than the model $\mathbf{w}_B$ learnt using ϕ_B.

- ◯ True
- ◯ False

Solution: FALSE. It will have lower bias and higher variance because the model becomes more complex and expressive due to the addition of squares of the features.

b. *[1 points]* The model $\mathbf{w}_B$ learnt using $\boldsymbol{\phi}_B$ has higher bias and lower variance than the model $\mathbf{w}$ learnt using the original feature vector $\mathbf{x}$.

- ◯ True
- ◯ False

Solution: FALSE. Since we get just doing a linear transformation of the features, the bias and variance will not change.

c. *[1 points]* The model $\mathbf{w}_B$ learnt using $\boldsymbol{\phi}_B$ is identical to the model $\mathbf{w}$ learnt using the original feature vector $\mathbf{x}$, that is, $\mathbf{w}_B = \mathbf{w}$.

- ◯ True
- ◯ False

Solution: FALSE. The predictions $\mathbf{w}^\top \mathbf{x}$ and $\mathbf{w_B}^\top \boldsymbol{\phi}_B$ are identical, not the models $\mathbf{w}$ and $\mathbf{w}_B$ are not.

2 Multiple Choice (6 points)

Problem 6: *[2 points]* If $\lambda_1, \ldots \lambda_n$ are the eigenvalues of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, what are the eigenvalues of the matrix $\mathbf{A}^2$?

- ◯ $\lambda_1, \ldots, \lambda_n$
- ◯ $2\lambda_1, \ldots, 2\lambda_n$
- ◯ $\lambda_1^2, \ldots, \lambda_n^2$
- ◯ None of the above

Solution: C. $\lambda_1^2, \ldots, \lambda_n^2$

$$\mathbf{A}^2\mathbf{x} = \mathbf{A}(\mathbf{A}\mathbf{x}) = \mathbf{A}\lambda_i\mathbf{x} = \lambda_i(\mathbf{A}\mathbf{x}) = \lambda_i^2\mathbf{x}$$

Problem 7: *[2 points]* Let $D = \{x_1, x_2, \ldots, x_n\}$ be an i.i.d. (independent and identically distributed) dataset sampled from a probability distribution $p(x \mid \theta)$, where θ is an unknown parameter. We use Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) to estimate θ. Which of the following statements is/are true? **More than one option can be true**.

- ☐ If the prior $p(\theta)$ is uniform, then MAP and MLE always give different estimates.
- ☐ Both MLE and MAP will give biased estimates of θ, i.e., their expected estimates will not equal the value of θ.
- ☐ If two priors $p_1(\theta)$ and $p_2(\theta)$ give the same MAP estimate, they must be identical distributions.
- ☐ None of the above is true.

Solution:

- ☐ **False**. If the prior $p(\theta)$ is uniform, then the MAP estimate reduces to the MLE estimate. This means they are the same, not different.
- ☐ **False**. The MLE estimate can be unbiased depending on how p depends on θ. For example, MLE produces an unbiased estimate of the Bernoulli distribution parameter. The MAP estimate can be unbiased as well, e.g., a uniform prior when estimating the Bernoulli distribution parameter.
- ☐ **False**. Two different priors $p_1(\theta)$ and $p_2(\theta)$ can lead to the same MAP estimate while being different distributions. This can happen when the likelihood function dominates the posterior distribution in such a way that the different priors do not affect the mode. For example, $p_1(\theta)$ and $p_2(\theta)$ may both have a peak at the maximizer of the likelihood function.
- ☐ **True**. Since all the previous statements are false, this option is correct.

One point was given if only one incorrect option was selected. Half a point was given if two incorrect options were selected.

Problem 8: *[2 points]* Recall that ridge regression learns a parameter vector $\mathbf{w}$ given a feature matrix $\mathbf{X}$ and label vector $\mathbf{y}$. Which of the following statements is/are true about ridge regression? **More than one option can be true.**

- ☐ Suppose that our data has d features. For sufficiently large d, ridge regression ensures that some entries in $\mathbf{w}$ will be equal to 0.
- ☐ Ridge regression adds an ℓ_2 regularization term to the standard linear regression formulation.

☐ Ridge regression worsens the model's generalization to unseen data, compared to the model obtained with linear regression.

☐ Ridge regression helps limit the magnitude of the entries of $\mathbf{w}$.

Solution: B, D. Ridge regression uses an ℓ_2 regularization term (B is correct), which helps limit the magnitude of the entries of $\mathbf{w}$ (D is correct). However, this does not ensure that any entries of $\mathbf{w}$ will actually be equal to 0 (A is incorrect). Ridge regression typically improves generalization to unseen data, as it reduces overfitting (C is incorrect).
One point was given for answers that selected both B and D, as well as A or C. Half a point was given for answers that selected one of B or D and one of A or C.

3 Descriptive (7 pts)

Problem 9: *[4 points]* Consider the dataset given by $\left\{\left(\begin{bmatrix}1\\0\end{bmatrix}, y_1\right), \left(\begin{bmatrix}0\\1\end{bmatrix}, y_2\right)\right\}$. You decide to apply linear regression to this dataset, i.e., to solve the following optimization problem:

$$\min_{w_1,w_2} \left\| \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} \begin{bmatrix}w_1\\w_2\end{bmatrix} - \begin{bmatrix}y_1\\y_2\end{bmatrix} \right\|_2^2. \tag{1}$$

a. Find the values of w_1 and w_2 that solve the linear regression problem in Equation (1), and find the resulting squared ℓ_2 loss of this model. Your answer may be in terms of y_1 and y_2.

Solution: We can see by inspection that if we take $w_1 = y_1$, $w_2 = y_2$, then the ℓ_2 loss in Equation (1) will be equal to 0, i.e., we will perfectly fit our two data points.

Alternatively, we can calculate this solution from the closed-form solution to the linear regression problem:

$$\begin{bmatrix}w_1\\w_2\end{bmatrix} = \left(\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}^T \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)^{-1} \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} \begin{bmatrix}y_1\\y_2\end{bmatrix} = \begin{bmatrix}y_1\\y_2\end{bmatrix}.$$

b. You now decide to compare your model from part (a) to the solution of a ridge regression model with regularization parameter λ. In other words, your new parameters w_1, w_2 will solve the optimization problem

$$\min_{w_1,w_2} \left\| \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} \begin{bmatrix}w_1\\w_2\end{bmatrix} - \begin{bmatrix}y_1\\y_2\end{bmatrix} \right\|_2^2 + \lambda\left(w_1^2 + w_2^2\right). \tag{2}$$

Find a closed-form solution for w_1, w_2 that solves Equation (2). Your answer may be in terms of y_1, y_2, λ.

Does the squared ℓ_2 loss

$$\left(\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}w_1\\w_2\end{bmatrix} - y_1\right)^2 + \left(\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}w_1\\w_2\end{bmatrix} - y_1\right)^2$$

increase or decrease compared to the loss value you found in part (a)?

Solution: We can calculate the solution from the closed-form solution to the linear regression problem:

$$\begin{bmatrix}w_1\\w_2\end{bmatrix} = \left(\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}^T \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} + \begin{bmatrix}\lambda & 0\\0 & \lambda\end{bmatrix}\right)^{-1} \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} \begin{bmatrix}y_1\\y_2\end{bmatrix} = \begin{bmatrix}1+\lambda & 0\\0 & 1+\lambda\end{bmatrix}^{-1} \begin{bmatrix}y_1\\y_2\end{bmatrix} = \begin{bmatrix}\frac{1}{1+\lambda} & 0\\0 & \frac{1}{1+\lambda}\end{bmatrix} \begin{bmatrix}y_1\\y_2\end{bmatrix} = \begin{bmatrix}\frac{y_1}{1+\lambda}\\\frac{y_2}{1+\lambda}\end{bmatrix}.$$

The loss value will increase compared to part (a), since in part (a) we were able to obtain 0 loss.

c. Now suppose that a new data point, $\left(\begin{bmatrix}0\\0\end{bmatrix}, 1\right)$ has arrived. You decide to re-solve for the linear regression model with all three of your data points $\left\{\left(\begin{bmatrix}1\\0\end{bmatrix}, y_1\right), \left(\begin{bmatrix}0\\1\end{bmatrix}, y_2\right), \left(\begin{bmatrix}0\\0\end{bmatrix}, 1\right)\right\}$, i.e., you wish to solve

$$\min_{w_1,w_2} \left\| \begin{bmatrix}1 & 0\\0 & 1\\0 & 0\end{bmatrix} \begin{bmatrix}w_1\\w_2\end{bmatrix} - \begin{bmatrix}y_1\\y_2\\1\end{bmatrix} \right\|_2^2. \tag{3}$$

Find a closed-form solution for w_1, w_2 that solves Equation (3). Your answer may be in terms of y_1, y_2.

Solution: We first calculate that

$$\left(\begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\end{bmatrix}^T \begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\end{bmatrix}\right)^{-1} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}^{-1} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

. Thus, our closed-form solution becomes

$$\begin{bmatrix}w_1\\ w_2\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\end{bmatrix}^T \begin{bmatrix}y_1\\ y_2\\ 1\end{bmatrix} = \begin{bmatrix}y_1\\ y_2\end{bmatrix}.$$

Alternatively, we can see by inspection that our linear model will predict the value of 0 for the third data point, no matter which w_1, w_2 we choose. Thus, since this new data point will always have ℓ_2 loss of 1, we obtain the same model as in part (a).

d. Suggest a way to modify the linear regression problem from part (c) that will allow you to reduce the ℓ_2 loss on your three data points $\left\{\left(\begin{bmatrix}1\\ 0\end{bmatrix}, y_1\right), \left(\begin{bmatrix}0\\ 1\end{bmatrix}, y_2\right), \left(\begin{bmatrix}0\\ 0\end{bmatrix}, 1\right)\right\}$. Your answer should be 1-2 sentences long.

Solution: We can introduce a bias term, which gives us an additional degree of freedom in our optimization problem and allows us to make a non-zero prediction for the third data point $\begin{bmatrix}0\\ 0\end{bmatrix}$.

Note that introducing a regularization term will not reduce the ℓ_2 loss on the three training data points. *Any* combination of model weights will result in a prediction of 0 on the data point $\begin{bmatrix}0\\ 0\end{bmatrix}$, yielding an ℓ_2 loss of 1 on that data point, and from part (c), we cannot reduce the ℓ_2 loss on our other two data points any further (it is already 0). To reduce the ℓ_2 loss on these three data points, we must add a bias term or use another feature transformation to obtain nonzero feature values for this third data point. We also accepted answers that said to add more data to the training dataset.

WRITE YOUR ANSWER TO PROBLEM 9 BELOW

CONTINUE YOUR ANSWER TO PROBLEM 9, IF NEEDED

Problem 10: *[3 points]*
Suppose that the time X taken by a generative AI model to respond to a prompt follows an exponential distribution with rate μ, where the probability density function of the exponential distribution is given by

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{for all } x \geq 0$$

For 5 prompts, we observe that the response times are 1, 0.5, 2, 1.25 and 0.25 respectively.

a. What is the maximum likelihood estimate of λ?

b. Suppose we put a prior on λ whose probability density lies in the range $[0, 1]$ and is concentrated around 0.5, and we find the MAP estimate λ_{MAP}. Is λ_{MAP} greater or less than the MLE estimate λ_{MLE} that you found in part (a)? Explain your reasoning in 1-2 sentences.

WRITE YOUR ANSWER TO PROBLEM 10 BELOW

Solution:

a. The log likelihood of the data is:

$$\begin{aligned}
\log P(\mathcal{D}|\lambda) &= \log \left(\prod_{i=1}^{5} \lambda e^{-\lambda x} \right) \\
&= 5 \log \lambda - \lambda \sum_{i=1}^{5} x_i \\
&= 5 \log \lambda - \lambda(1 + 0.5 + 2 + 1.25 + 0.25) \\
&= 5 \log \lambda - 5\lambda
\end{aligned}$$

Taking the derivative of the log likelihood and setting to zero, we can solve for λ_{MLE}:

$$\begin{aligned}
\frac{5}{\lambda} - 5 &= 0 \\
\lambda_{\text{MLE}} &= 1
\end{aligned}$$

Since the log likelihood is a concave function, this is the unique global maximum.

b. The MAP estimate λ_{MAP} will be smaller than $\lambda_{\text{MLE}} = 1$ because it will take the prior, which is concentrated around 0.5 into consideration.

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.