Practice Questions for the ECE 661 Spring 2025 Midterm Exam

Introduction to Machine Learning for Engineers
Prof. Gauri Joshi and Prof. Carlee Joe-Wong

The number of practice questions included in this document may not be representative of the length of the exam. These are solely intended for you to gain more experience in answering questions similar to what you will see on the exam.

1 True or False

Problem 1: The reason why Naïve Bayes is naïve is that it assumes a prior distribution on the label of a data point.

- ◯ True
- ◯ False

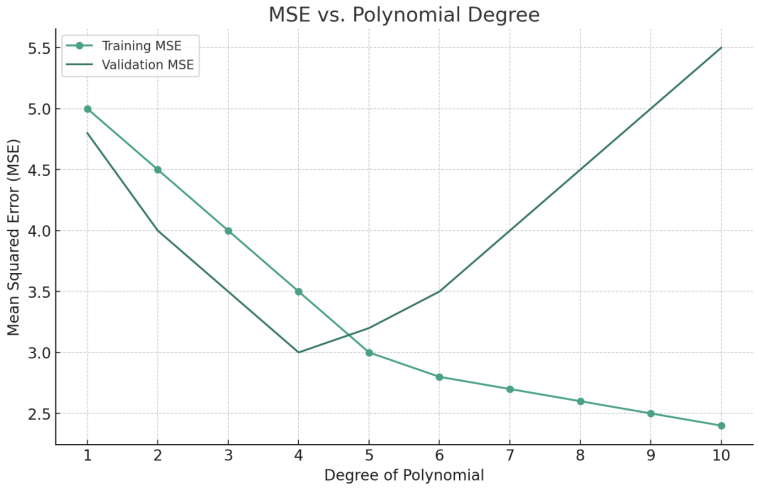
Solution: False

Problem 2: We can use the k-nearest neighbours algorithm to solve regression problems.

- ◯ True
- ◯ False

Solution: True

Problem 3:
Assume you are using polynomial regression to solve a regression problem. You tried training polynomial models of various degrees and compared their Mean Squared Error (MSE) on a validation dataset, the following is a graph of the MSE as the degree of the polynomial increases.



Based on the graph, the degree 4 polynomial model provides a balance between bias and variance, optimizing performance on the validation dataset.

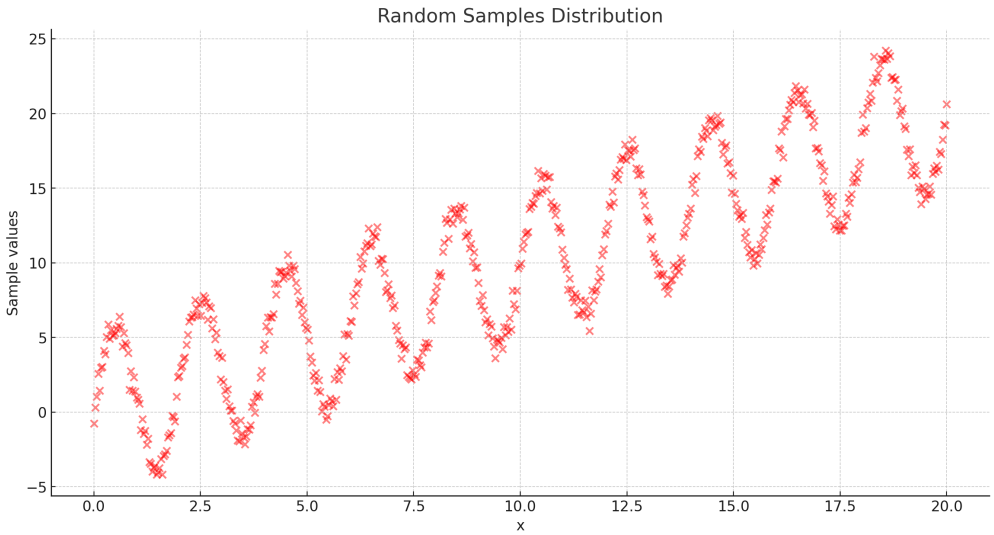
◯ True

◯ False

Solution: True.

Problem 4: Assume you are given a sample training dataset which consists of real numbers $(x_i, y_i) \in \mathbb{R}^2$, for $i = 1, 2, ..., N$, where the x_i are one-dimensional input features and y_i the labels. To understand the data, you plotted those points as x values corresponding to the horizontal axis and the y values corresponding to the vertical axis. Your friend used the following nonlinear basis for the same problem but did not get a good test performance. That is, the test samples were far away from the fitted curve:

$$\boldsymbol{\phi}(x) = \begin{bmatrix} 1 & x & x^2 & \dots & x^{42} & \sin(\pi x) & \cos(\pi x) & \dots & \cos(61\pi x) & e^x \end{bmatrix}^T$$



The mostly likely explanation for this performance is that the model is underfitting the data.

◯ True

◯ False

Solution: False. Your friend's model is likely overfitting to the data.

2 Multiple Choice

Problem 5: Consider a multi-class classification problem with 3 classes (c_1, c_2, c_3). To solve this problem, you learn three sets of parameters, $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3$, each associated with one class. A data point $(\mathbf{x}_i, y_i)$ is of class 3 if:

- ◯ $\mathbf{w}_3^\top \mathbf{x}_i \geq \mathbf{w}_2^\top \mathbf{x}_i$
- ◯ $\mathbf{w}_3^\top \mathbf{x}_i < \mathbf{w}_1^\top \mathbf{x}_i$ and $\mathbf{w}_2^\top \mathbf{x}_i < \mathbf{w}_3^\top \mathbf{x}_i$
- ◯ $(\mathbf{w}_3 - \mathbf{w}_1)^\top \mathbf{x}_i > 0$ and $\mathbf{w}_2^\top \mathbf{x}_i < \mathbf{w}_3^\top \mathbf{x}_i$
- ◯ $\mathbf{w}_1^\top \mathbf{x}_i \leq \mathbf{w}_2^\top \mathbf{x}_i$

Solution: C. $(\mathbf{w}_3 - \mathbf{w}_1)^\top \mathbf{x}_i > 0$ and $\mathbf{w}_2^\top \mathbf{x}_i < \mathbf{w}_3^\top \mathbf{x}_i$

Problem 6: In the context of machine learning model performance, the error of a model can be decomposed into three distinct components. Consider the expression for the expected prediction error R for a model $h_D(x)$ trained on dataset D as follows:

$$\mathbb{E}_D\left[R[h_D(x)]\right] = \text{variance} + \text{bias}^2 + \text{noise}$$

. What is the relationship of the variance, bias, and noise terms?
[Hint:] The expression $a \propto b$ means that the quantities a and b are proportional to each other, i.e., one is a constant multiple of the other.

- ◯ $\text{bias}^2 \propto \text{variance}$; $\text{bias}^2 + \text{variance} \propto \text{noise}$
- ◯ $\text{bias}^2 \propto \frac{1}{\text{variance}}$; $\text{bias}^2 \propto \text{noise}$
- ◯ $\text{bias}^2 \propto -\text{variance}$; $\text{variance} \propto \text{noise}$
- ◯ None of the above.

Solution: D. The trade-off does not imply proportionality. There is no direct mathematical relationship between the three.

Problem 7: Suppose you are using k-nearest neighbors to solve a binary classification problem with training dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_i), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_{999}, y_{999})\}$ and validation dataset $\mathcal{D}_{val}$. Let $c_{val}^{(k)}$ denote the classification accuracy on a validation dataset consisting of 50 data points, given the value of k in k-nearest neighbors. 25 of the data points in the validation dataset belong to one class, and 25 belong to the other. You tried three values of k (α, β, and γ), and find that $c_{val}^{(\alpha)} < \frac{1}{2} < c_{val}^{(\gamma)} < c_{val}^{(\beta)}$. You further find that α-nearest neighbors has higher classification accuracy than β- or γ-nearest neighbors on the training dataset. Which of the following are possible values of α, β, and γ?

- ◯ $\alpha = 11$, $\beta = 1$, $\gamma = 3$
- ◯ $\alpha = 5$, $\beta = 999$, $\gamma = 7$
- ◯ $\alpha = 1$, $\beta = 15$, $\gamma = 3$
- ◯ $\alpha = 1$, $\beta = 5$, $\gamma = 999$

Solution: We can eliminate option A since 1-nearest neighbors always has the highest training accuracy of 100% (each training data point is its own nearest neighbor). Since β-nearest neighbors and γ-nearest neighbors both have validation accuracies above $\frac{1}{2}$, we know that $\beta \neq 999$ and $\gamma \neq 999$: taking $k = 999$ means that all validation data points would be classified into the same class, and we know that the validation

dataset is equally divided between the two classes (i.e., the accuracy with $k = 999$ would be exactly $\frac{1}{2}$). Thus, we also eliminate options B and D, so option C is the only one remaining.

Problem 8: You are tasked with developing a predictive model for the daily number of visitors to a national park. Given a dataset of daily visitor counts over several years, you decide to model the visitor count Y using a Poisson distribution, where Y represents the number of visitors per day and λ is the rate parameter of the distribution. To estimate λ, you consider using both Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) estimation methods. You also have prior knowledge that the visitor count has seasonal variations, which you decide to model with a Gamma prior on λ, given its flexibility in representing rate parameters. The Gamma prior is parameterized by shape α and rate β.
Which of the following statements correctly distinguishes the application of MLE and MAP in estimating λ for this scenario, and what could be a potential outcome of using each method?

- ◯ MLE will likely outperform MAP in terms of predictive accuracy for future visitor counts because MLE does not rely on prior assumptions, making it more adaptable to changes in visitor trends over time.
- ◯ MAP estimation, by incorporating the Gamma prior reflecting seasonal variations, can potentially provide a more accurate and stable estimate of λ during different seasons compared to MLE, which might overlook these variations due to its reliance solely on observed data.
- ◯ Using a Gamma prior in MAP estimation will constrain the model to only predict visitor counts that directly match the historical seasonal patterns, leading to poor performance in predicting any new trends or changes in visitor behavior.
- ◯ MLE and MAP will produce identical estimates for λ regardless of the choice of prior, because the Poisson distribution's properties ensure that prior information does not influence the estimation of rate parameters.

Solution: B.
A is incorrect since even though MLE is adaptable and directly reflects observed data, it does not necessarily outperform MAP in scenarios where prior knowledge is relevant and can improve model accuracy.

B is correct as MAP's incorporation of a Gamma prior, designed to reflect known seasonal variations in visitor counts, allows for a more accurate estimation of λ that accounts for these expected fluctuations. This can lead to more accurate predictions during different seasons than MLE.

C is incorrect because incorporating a prior in MAP estimation does not really constrain the model to historical patterns. Instead, it balances prior knowledge with observed data to adjust predictions, allowing for flexibility in capturing new trends as more data becomes available.

D is incorrect because the properties of the Poisson distribution do not negate the influence of the prior in MAP estimation. The choice of prior can significantly affect the MAP estimate.

Problem 9: Consider training a ridge regression model in which the bias term is accidentally included in the regularization term $\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$ of the ridge regression objective. What likely impact would this have on the model compared to not including the bias term in the regularization? You should assume that the true model has nonzero bias term.

- ◯ model training loss decreases, model test loss decreases
- ◯ model training loss decreases, model test loss increases
- ◯ model training loss increases, model test loss decreases

◯ model training loss increases, model test loss increases

Solution: D. Regularizing the bias term would encourage the intercept to move toward zero. If the true model has nonzero bias, the bias term would be important for calibration. Therefore, this leads to an increase in the model's training loss and a decrease in model performance.

Problem 10: You are training a linear regression model on the dataset you just collected for your ambitious machine learning project. During the data analysis, you have found that the input features are not directly proportional to the target label. Among the following, what are the best modelling approaches for your dataset to ensure low test loss?

◯ Using ridge regression with clever choice of regularization parameter will achieve the lowest error on your validation set. Ridge regression is suited for any kind of dataset.

◯ Using a polynomial model of carefully chosen order might achieve the lowest error on the validation set.

◯ Using a complex model for this dataset is probably not the right approach. A very simple linear model without regularization will achieve the lowest error on validation set.

◯ Removing some of the training data to ensure that linear regression will yield a low training loss on the reduced dataset.

Solution: B. Polynomial model is a good fit for non-linear data given the dataset description(non-linearity in input and target feature).
A is incorrect, since ridge regression is still linear model. C is incorrect, since a very simple linear model is unlikely to work in this dataset. D is incorrect, since we still learn a linear model.

3 Descriptive

Problem 11: Suppose a student is taking an exam that consists of M multiple-choice questions. Due to a lack of preparation, the student decides to guess on all questions. However, the student has partial knowledge that influences their guessing accuracy, making the probability of correctly guessing a question unknown.
Let X be the random variable representing the number of multiple-choice questions the student answers correctly. Assume that the student's guesses are independent from one question to another.

a. What is the log-likelihood of observing x correct answers? The binomial distribution can be used to model the probability a student answers exactly k questions correctly (i.e., $X = k$) out of the total number of questions:

$$P(X = k) = \binom{M}{k} p^k (1-p)^{M-k}$$

b. Suppose the student correctly answers x out of M questions. Using your answer to part (a), find the MLE of p.

   For the next part of the problem, assume you have a prior belief about the student's guessing ability, which can be modeled with a Beta distribution with parameters α, β:

$$\text{Beta}(p|\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$$

c. Write the log-likelihood function of observing x correct answers using Maximum A Posteriori (MAP), incorporating the Beta prior.

d. Calculate the MAP estimate of p given the prior parameters α and β, and the observed data of x correct answers out of M questions.

Solution:
1)

$$L(p) = P(X = x) = \binom{M}{x} p^x (1-p)^{M-x}$$
$$\log L(p) = \log\left(\binom{M}{x}\right) + x\log(p) + (M-x)\log(1-p)$$

2)

$$\frac{d}{dp}\log L(p) = \frac{x}{p} - \frac{M-x}{1-p} = 0$$
$$\frac{x}{p} = \frac{M-x}{1-p}$$
$$x(1-p) = p(M-x)$$
$$x - xp = pM - px$$
$$x = pM$$
$$p = \frac{x}{M}$$

3)

$$\begin{aligned} L(p) &\propto P(X = x) \times \text{Beta}(p|\alpha, \beta) \\ &\propto p^x(1-p)^{M-x} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{x+\alpha-1}(1-p)^{M-x+\beta-1} \\ \log(L(p)) &= (x+\alpha-1)\log(p) + (M-x+\beta-1)\log(1-p) \end{aligned}$$

4)

$$\begin{aligned} \frac{d}{dp}\log(L(p)) &= \frac{x+\alpha-1}{p} - \frac{M-x+\beta-1}{1-p} = 0 \\ (x+\alpha-1)(1-p) &= (M-x+\beta-1)p \\ x+\alpha-1 &= Mp - xp + \beta p - p + p\,(x+\alpha-1) \\ x+\alpha-1 &= Mp + \beta p + \alpha p - p \\ p &= \frac{x+\alpha-1}{M+\beta+\alpha-2} \end{aligned}$$

Problem 12: Suppose that you wish to learn a linear regression model on the training dataset $\{\boldsymbol{\phi}(x_i), y_i; i = 1, 2, \ldots, n\}$, where $\boldsymbol{\phi}$ is a given nonlinear transformation of the one-dimensional input feature x and we use y_i to denote the label of each training data point $i = 1, 2, \ldots, n$.

a. Write out the residual sum of squares (RSS) of the above nonlinear regression problem in terms of $\boldsymbol{w}$ and $\{\boldsymbol{\phi}(x_i), y_i; i = 1, 2, \ldots, N\}$.

Solution: $RSS = \sum_{i=1}^{N} (y_i - \boldsymbol{w} \cdot \boldsymbol{\phi}(x_i))^2$

b. You are given the following hint about the minimizer of the RSS that you found in part (a): $\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i \boldsymbol{\phi}(x_i)$, where $\alpha_i \in \mathbb{R}$.That is, the optimal $\boldsymbol{w}$ is a linear combination of $\{\boldsymbol{\phi}(x_i); i = 1, 2, \ldots, N\}$. Using this hint, write an expression for the RSS in terms of $\{\alpha_i, \boldsymbol{\phi}(x_i), y_i; i = 1, 2, \ldots, N\}$.

Solution:

RSS $= \sum_{i=1}^{n} \left(y_i - \boldsymbol{w}^{*T}\boldsymbol{\phi}(x_i)\right)^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \alpha_j \boldsymbol{\phi}(x_j)^T \boldsymbol{\phi}(x_i)\right)^2$

c. Find the $\boldsymbol{\alpha}$ that minimizes the RSS, using the expression you found in part (b). It may be helpful to use the $n \times n$ matrix $\boldsymbol{M}$, where each (i, j) entry of M is defined as $\boldsymbol{M}_{i,j} = \boldsymbol{\phi}(x_i)^T \boldsymbol{\phi}(x_j)$. You may assume that $\boldsymbol{M}^T\boldsymbol{M}$ is invertible.

Solution: We can write the RSS as $||\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}||^2$. Thus, we solve for $\boldsymbol{\alpha} = (\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T\boldsymbol{y} = \boldsymbol{M}^\dagger\boldsymbol{y}$.

Problem 13: Consider using nearest neighbors to classify a binary dataset consisting of the following two sets of training samples:

- Label $y = 0$: n_0 points lying on a circle of radius r_0 centered at $(0, 0)$
- Label $y = 1$: n_1 points lying on a circle of radius r_1 centered at $(0, 0)$, where $n_1 > n_0$ and $r_1 > r_0$

What is the predicted label ($\hat{y} = 0$ or $\hat{y} = 1$) of a query point $(0, 0)$ using k-nearest neighbors for the following values of k? Explain the reasoning behind each answer in one sentence.

a. $k = 1$

b. $k = n_0$

c. $k = 2n_0 + 1$

d. $k = n_0 + n_1$

To give more weightage to points that are closer, we decide to use weighted nearest neighbors where each of the k nearest points is assigned a weight that is inversely proportional to its distance from the query point.

e. Evaluate the range of $k \in [1, n_0 + n_1]$ for which the predicted label of the query point $(0, 0)$ is 0. Assume that ties are broken in favor of label 0.

Solution:

a. $\hat{y} = 0$

b. $\hat{y} = 0$

c. $\hat{y} = 1$

d. $\hat{y} = 1$

e. We declare the label as 0 if:

$$\frac{\min(n_0, k)}{r_0} \geq \frac{\min(n_1, k - n_0)}{r_1}$$
$$\frac{\min(n_0, k)}{r_0} \geq \frac{k - n_0}{r_1}$$

Thus, if $k \leq n_0$, the label is 0. If $n_0 + n_1 \geq k > n_0$, the decision rule becomes,

$$\frac{n_0}{r_0} \geq \frac{k - n_0}{r_1}$$
$$n_0 \left(\frac{r_1}{r_0} + 1 \right) \geq k$$

Thus, putting these two ranges together, the label is predicted as 0 for all $1 \leq k \leq n_0 \left(\frac{r_1}{r_0} + 1 \right)$.