CSE 446 Spring 2024 Midterm Exam

May 6, 2024

Name ________________________________ **UW NetID** _____________________

Please **wait** to open the exam until you are instructed to begin, and please take out your Husky Card and have it accessible when you turn in your exam.

Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

- For each multiple choice and True/False question, clearly indicate your answer by filling in the letter(s) associated with your choice.
- Multiple choice questions marked with $\boxed{\text{One Answer}}$ should only be marked with one answer. All other multiple choice questions are select all that apply, in which case any number of answers may be selected (**including none, one, or more**).
- For each short answer question, please write your answer in the provided space.
- If you need to change an answer or run out of space, please very clearly indicate what your final answer is and what you would like graded. Responses where we cannot determine the selected option will be marked as incorrect.
- Please remain in your seats for the last 10 minutes of the exam. If you complete the exam before the last 10 minutes, you may turn in your exam and note sheet by handing them to a TA.

1. In a machine learning classification problem, you have a dataset with two classes: Positive (P) and Negative (N). The probability of a randomly selected sample being Negative is 0.6. The probability of a correct classification given that the sample is Positive is 0.8, and the probability of a correct classification given that the sample is Negative is 0.6. What is the probability that a randomly selected sample is Positive given that it has been classified as Positive?

   (a) $\frac{4}{7}$

   (b) $\frac{8}{17}$

   (c) $\frac{4}{5}$

   (d) $\frac{4}{15}$

Correct answers: (a)

2. Consider the closed form of the optimal weight for Ridge Regression, as derived in a previous homework (HW1):

$$\widehat{W} = (X^\top X + \lambda I)^{-1} X^\top Y,$$

where $X = [x_1 \cdots x_n]^\top \in \mathbb{R}^{n\times d}$ and $Y = [y_1 \cdots y_n]^\top \in \mathbb{R}^{n\times k}$.

What is **NOT** true in the following statements?

(a) When $\lambda > 0$, the matrix $X^\top X + \lambda I$ is invertible.

(b) The identity I is a $d \times d$ matrix.

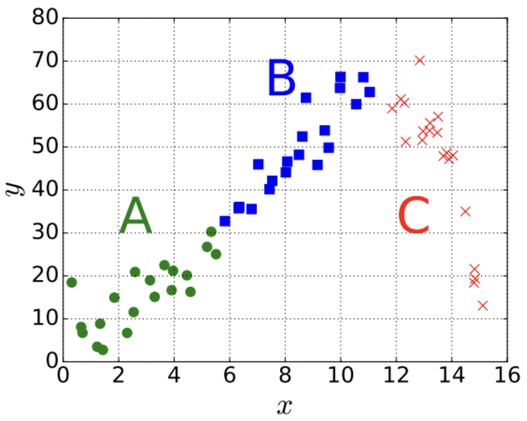
(c) When $\lambda = 0$, the matrix is not full-rank, so there is no solution for Ridge Regression.

(d) If we apply a unitary transform $U \in \mathbb{R}^{d\times d}$ $(U^\top U = I)$ on the input X and output Y to get another dataset (UX, UY), the new estimated weight would still be $\widehat{W}$.

The next two questions are based on this preamble and the Figure below.

As a fresh graduate of CSE 446, your biologist friend seeks your help in modeling the relationship between concentration y of amino acid Arginine in blood plasma after interacting with a reagent for x hours. The experiment was designed to measure the concentration within 3 distinct time blocks (**A,B,C**):

- **A**: time $x = 0$ to around 6 hours (circles)
- **B**: time x around 6 hours to x around 12 hours (squares)
- **C**: time x to around 12 hours to x around 16 hours (the symbol x)



Correct answers: (c)

3. Suppose your hypothesis is that Arginine concentration is linearly related to time, i.e. $y = \theta x + \theta_0$. You employ mean square error (MSE) for the objective function, and use dataset **A** for **training**, and dataset **B** for **testing**. Let us say that MSE below 30 is LOW, and MSE above 100 is HIGH. Judging from the above plot, will the MSE for training be Low or High?

   (a) Training Error: Low. Testing Error: Low.

   (b) Training Error: LOW. Testing Error: High.

   (c) Training Error: High. Testing Error: Low.

   (d) Training Error: High. Testing Error: High.

   Correct answers: (a)

   Explanation: Training Error: Low. Testing Error: Low.

   Both errors are LOW because training on dataset A should produce a straight line which fits both A and B very well.

4. Continuing with the linear relation hypothesis, you now employ datasets **A** and **B** (combined) for **training**, and dataset **C** for **testing**. Judging from the above plot, will the MSE for training be LOW or HIGH? How about for testing?

   (a) Training Error: Low. Testing Error: Low.

   (b) Training Error: Low. Testing Error: High.

   (c) Training Error: High. Testing Error: Low.

   (d) Training Error: High. Testing Error: High.

   Correct answers: (b)

   Explanation: Training Error: Low. Testing Error: High.

   Training error will be LOW because training on dataset A and B should produce a straight line which fits both A and B very well. However, extrapolating forward the straight line produced will not be a good fit for dataset C leading to a HIGH testing MSE.

5. Apply gradient descent with step size $\eta = \frac{1}{2}$ to find the minimizer for function $f : \mathbb{R}^2 \mapsto \mathbb{R}$ defined as

$$f(x) = (x_1 - 1)^2 + (x_2 - 2)^2 + x_1 x_2.$$

Starting at $x^{(0)} = (0, 0)$, we iterate with $x^{(t+1)} = x^{(t)} - \eta \nabla_x f(x^{(t)})$. What is $x^{(2)}$?

(a) $(0, \frac{3}{2})$

(b) $(1, 2)$

(c) $(-1, -2)$

(d) $(-4, -\frac{13}{2})$

Correct answers: (a)

Explanation: $\nabla_x f(x) = (2(x_1 - 1) + x_2, 2(x_2 - 2) + x_1)$. So $\nabla_x f(x^{(0)}) = (-2, -4)$, $x^{(1)} = (1, 2)$, $\nabla_x f(x^{(1)}) = (2, 1)$, $x^{(2)} = (0, \frac{3}{2})$.

6. In the context of stochastic gradient descent, what is the effect of using a very small mini-batch size (e.g., 1)?

(a) It increases the computational efficiency due to vectorization over the mini-batch.

(b) It guarantees that the algorithm will find the global minimum for non-convex problems.

(c) It may lead to high variances in the gradient computed at each step, causing the updates to oscillate.

Correct answers: (c)

Explanation: (a) Gradient of current batch depends on previous model, cannot vectorization. (b) No such guarantee.

7. Define the least square loss to be $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where there are n samples in the data, $\hat{y}_i$ is the predictive value, and y_i is the label in the dataset. We assume $y_i = W^\top x_i + \epsilon_i$ where ϵ_i is the noise. In the context of linear regression, which of the following statements is true? One Answer

   (a) Least squares estimation is equivalent to MLE when the noise is assumed to be uniformly distributed.

   (b) MLE for linear regression doesn't need to assume any distribution of the noise.

   (c) Using least squares loss in linear regression is equivalent to MLE if the noise is assumed to be normally distributed with zero mean and any variance.

   (d) MLE can be applied only when the noise distribution is non-normal, whereas least squares estimation does not assume any specific distribution of the noise.

Correct answers: (c)

Explanation: (C) is correct because the equivalence of least squares and MLE specifically hinges on the assumption that the errors are normally distributed with zero mean. The variance does not need to be known or constant.

8. Consider a dataset containing three observations for a simple linear regression problem, where y is the dependent variable and x is the independent variable. The dataset is given as follows:

x	y
1	2
2	3
3	5

Find the coefficient β_1 of the linear regression (without bias) $\hat{y} = \beta_1 x$ using the least squares as loss.

$\beta_1 =$ ______

Explanation: $\beta_1 = \frac{23}{14}$
$\beta_1 = (X^\top X)^{-1}(X^\top Y) = \frac{1}{14} \times 23$

9. Assume you have a set of independent observations from a normally distributed population: $2, 5, 7, 4$. If the mean (μ) of the population is known to be 4, calculate the Maximum Likelihood Estimation (MLE) of the variance (σ^2).

(a) $\sigma^2 = 3.0$

(b) $\sigma^2 = 3.5$

(c) $\sigma^2 = 4.5$

(d) $\sigma^2 = 5.0$

Correct answers: (b)

Explanation: $\sigma^2 = \frac{(2-4)^2+(5-4)^2+(7-4)^2+(4-4)^2}{4} = 3.5$

10. You are analyzing the time until failure for a set of lightbulbs. The data represents the number of months each bulb lasted before failing and is given as follows: 2, 3, 4, 5. Assuming these times are modeled as being drawn from an exponential distribution, calculate the maximum likelihood estimate (MLE) of the rate parameter (λ) of this distribution.

The probability density function (PDF) for the exponential distribution is given by $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$.

$\lambda =$ ______

Explanation: MLE of $\lambda = \frac{n}{\Sigma_{i=1}^{n} x_i} = \frac{2}{7}$

11. Suppose you train a polynomial regression of degree $d = 2$ to approximate the quadratic function $y = 7x^2 - 3x + \varepsilon$, where ε is a Gaussian random variable $N(0, 1)$. What is the irreducible error? One Answer

(a) 0

(b) 1

(c) -3

(d) 7

(e) 2

Correct answers: (b)

12. The irreducible error is typically: One Answer

(a) Observable and quantifiable during the model training phase.

(b) Reduced by increasing the complexity of the model.

(c) Independent of the model used and represents the noise level in the data.

(d) Eliminated by extensive feature engineering.

Correct answers: (c)

13. Suppose you train a linear regression model to approximate the quadratic function $g(x) = 7x^2 + 3$. What's the most likely outcome? One Answer

(a) The model will have high bias and high variance

(b) The model will have high bias and low variance

(c) The model will have low bias and high variance

(d) The model will have low bias and low variance

Correct answers: (b)

14. Which statement best describes the purpose of using basis functions in linear regression? One Answer

(a) To transform the model into a non-linear regression.

(b) To simplify the relationship between features and the target variable.

(c) To allow the linear model to capture non-linear relationships in the data by transforming the input variables.

(d) To reduce the computational complexity of fitting the model.

Correct answers: (c)

15. What is likely to happen if you increase the regularization parameter significantly beyond the optimal point for LASSO regression?

    (a) The model becomes more complex, including more features in the prediction.

    (b) The model simplifies, potentially ignoring some relevant features, leading to underfitting. model increases its accuracy on the training data but performs poorly on unseen data.

    (c) The model remains unaffected regardless of changes in the regularization parameter.

    Correct answers: (b)

16. In the context of Ridge Regression, how does the application of data standardization on features influence the optimal regularization parameter (i.e., lambda) compared to scenarios where it is not applied?

    (a) The optimal lambda is always increased or unchanged.

    (b) The optimal lambda is always decreased or unchanged.

    (c) The optimal lambda might increase or decrease.

    (d) The optimal lambda is irrelevant to data standardization.

    Correct answers: (c)

17. Which of the following scenarios would be most appropriate to use logistic regression instead of linear regression?

    (a) Predicting a person's height based on their parents' heights.

    (b) Predicting whether a student will pass or fail a class based on study hours.

    (c) Predicting the sale price of a house based on square footage and number of bedrooms.

    (d) Predicting the weight of a package based on its dimensions.

    Correct answers: (b)

18. Which of the following statements about logistic regression is NOT true?

(a) ℓ_2 (i.e., L^2) regularization can be added to the logistic regression cost function to prevent overfitting by penalizing large coefficient values.

(b) The maximum likelihood estimates for the logistic regression coefficients can be found in closed-form.

(c) The logistic sigmoid function is used to model the probability of the positive class in binary logistic regression.

(d) None of the above.

Correct answers: (b)

19. Which of the following sets is always convex?

(a) The intersection of any number of convex sets.

(b) The union of any number of convex sets.

(c) Boundary of a ball of positive radius.

(d) Set of vectors in $\mathbb{R}^d$ $(d > 2)$ which have at most 2 coordinates zero:

$$\{x \in \mathbb{R}^d \mid \text{there exist } i, j \in \{1, 2, \ldots d\} \text{ such that } x_i = x_j = 0\}.$$

Correct answers: (a)

20. True/False: The function $f\colon \mathbb{R}^d \to \mathbb{R}$ defined as $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$, is convex for all matrices $\mathbf{A} \in \mathbb{R}^{d \times d}$.

(a) True

(b) False

Correct answers: (b)