CSE 446/546 Autumn 2024 Midterm Exam

October 30, 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 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 to a TA.

1. [4 points] [Select All That Apply]

If X and Y are independent random variables, which of the following are **true**?

(a) $\text{Cov}(X, Y) = 0$

(b) $E[XY] = E[X]E[Y]$

(c) $\text{Var}(XY) = \text{Var}(X)\text{Var}(Y)$

(d) $P(X, Y) = P(Y|X)P(X|Y)$

2. [4 points] [One Answer]

A certain disease affects **2%** of the population. A diagnostic test for this disease has the following characteristics:

- **Sensitivity (True Positive Rate):** If a person has the disease, the test returns a positive result with probability **0.90**.
- **False Positive Rate:** If a person does not have the disease, the test returns a positive result with probability **0.10**.

If a randomly selected person tests positive, what is the probability that they actually have the disease?

(a) $\frac{11}{58}$

(b) $\frac{9}{58}$

(c) $\frac{9}{50}$

(d) $\frac{49}{58}$

3. [10 points]

The probability mass function of a geometric distribution with unknown parameter $0 < p \leq 1$ is

$$P(X = k|p) = (1-p)^{k-1}p,$$

where $k = 1, 2, 3, \ldots$. The interpretation of X is that it is the number of independent Bernoulli trials needed to get one success, if each trial has success probability p.

Given a set of n observations $\{x_1, x_2, \ldots, x_n\}$ from a geometric distribution, derive the Maximum Likelihood Estimate (MLE) $\hat{p}_{\text{MLE}}$ for the parameter p.

Hint: don't forget about the chain rule: for $h(x) = f(g(x))$, $h'(x) = f'(g(x))g'(x)$.

Answer: _____

4. [5 points] [Select All That Apply]

Which of the following is true about maximum likelihood estimation, in general?

(a) It always produces unbiased parameter estimates.

(b) It can be used for continuous probability distributions.

(c) It can be used for discrete probability distributions.

(d) It maximizes the likelihood of the data given the model parameters.

(e) It maximizes the likelihood of the model parameters given the data.

5. [4 points] [Select All That Apply]

Suppose $A \in \mathbb{R}^{n \times n}$ is a positive semi-definite (PSD) matrix. Which of the following is always **true** about A?

(a) All eigenvalues of A are non-negative.

(b) All elements of A are non-negative.

(c) A is invertible.

(d) $x^T Ax \leq 0$ for all x.

6. [4 points]

Assume we have $X \in \mathbb{R}^{n \times p}$ representing n data points with p features each and $Y \in \mathbb{R}^n$ representing the corresponding outcomes. Using linear regression with no offset/intercept, provide an expression to predict the outcome for a new data point $x_{\text{new}} \in \mathbb{R}^p$ in terms of X and Y.

Answer: $\hat{y}_{\text{new}} =$ _____

7. [4 points]

Suppose you want to use linear regression to fit a weight vector $w \in \mathbb{R}^d$ and an offset/intercept term $b \in \mathbb{R}$ using data points $x_i \in \mathbb{R}^d$. What is the minimum number of data points n required in your training set such that there will be a single unique solution?

Answer: _____

8. [2 points] [One Answer]

In a regression model, what is the primary purpose of using general basis functions?

(a) Transform nonlinear relationships between features and the target variable into a linear form.

(b) Regularize the model to prevent overfitting.

(c) Reduce the number of data samples needed for model training.

(d) Simplify the model by reducing the number of features.

9. [2 points] [One Answer]

In regression, when our prediction model is linear-Gaussian, i.e., $y_i \sim N(x_i^\top w, \sigma^2)$ for target output $y_i \in \mathbb{R}$ and feature vectors $x_i \in \mathbb{R}^d$, finding the w that maximizes the data likelihood is equivalent to minimizing the average absolute difference between the target output and predicted output.

(a) True

(b) False

10. [6 points] [Select All That Apply]

In ridge regression, we obtain $\widehat{w}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$ for $\lambda \geq 0$. Which of the following is **true**?

(a) $X^T X$ is always invertible.

(b) $X^T X + \lambda I$ is always invertible.

(c) Increasing λ typically adds bias to the model.

(d) Increasing λ typically adds variance to the model.

(e) When $\lambda = 0$, ridge regression reduces to ordinary (unregularized) linear regression.

(f) As $\lambda \to \infty$, $\widehat{w}_{\text{ridge}} \to 0$.

11. [2 points] [One Answer]

You have a dataset with many features. You know *a priori* that only a small portion of those features are relevant to your prediction problem, but you don't know which are the relevant features. Is it better to use Ridge regression or Lasso regression?

(a) Ridge regression

(b) Lasso regression

12. [2 points] [One Answer]

Which of the following best explains the effect of Lasso regression on the bias-variance tradeoff?

(a) Lasso regression reduces both bias and variance simultaneously, leading to a more accurate model.

(b) Lasso regression reduces bias by shrinking coefficients, often at the expense of increasing variance.

(c) Lasso regression reduces variance by shrinking coefficients and can increase bias, especially when some features are dropped entirely from the learned predictor.

(d) Lasso regression increases both bias and variance as it enforces sparsity in the learned predictor.

13. [2 points] [One Answer]

In prediction, the total expected prediction error can be decomposed into three components: bias squared, variance, and irreducible error. By optimizing the model complexity and increasing the size of the dataset, it is possible to reduce all three components.

(a) True
(b) False

14. [2 points] [One Answer]

Which strategy is most effective for reducing variance in a high-variance, low-bias model?

(a) Increasing the number of training examples.
(b) Increasing the model complexity.
(c) Decreasing regularization.
(d) Removing the features that exhibit high variance across training examples.

15. [2 points] [One Answer]

If your model has high validation loss and high training loss, which action is most appropriate to improve the model?

(a) Increase the model complexity.
(b) Increase k in k-fold cross-validation.
(c) Increase the number of training examples.
(d) Apply regularization to reduce overfitting.

16. [4 points]

In a project using a customer purchase history dataset with a 60/20/20 train, validation, and test split, the validation accuracy remains consistently lower than the training accuracy. What could be a reason for this?

Answer: _____

17. [2 points] [One Answer]

A consortium of 10 hospitals have pooled together their Electronic Health Records data and want to build a machine learning model to predict patient prognosis based on patient records in their hospitals. They want to maximize the accuracy of their model across all 10 hospitals and do not plan to deploy their model in other hospitals. How should they split the data into train / validation / test sets?

(a) Leave out data from 1 hospital for the validation set, data from another hospital for the test set, and use the rest for train set.

(b) Leave out data from 1 hospital for the validation set, data from another hospital for the test set, and use the rest for train set. After training, add the validation data to the train set and re-train the model on the combined data.

(c) Use data from 8 hospitals with the most number of records for training, and use data from the other 2 hospitals for validation and test sets.

(d) Mix data from all hospitals, randomly shuffle all the records, and then do the 80/10/10 train/validation/test split.

18. [2 points]

Given the task of determining loan approval for applicants using a predictive model given applicant features such as race, salary, education, etc., is it always best practice to allow the model to use all of the given features? Why or why not?

Answer: _____

19. [2 points] [One Answer]

You are building a predictive model about users of a website. Suppose that after you train your model on historical user data, the distribution of users shifts dramatically. What can happen if you deploy your machine learning system without addressing this distribution shift?

(a) The model will automatically adapt to new data distributions.

(b) The model will generate more diverse predictions, increasing its overall accuracy.

(c) The model will maintain its original performance indefinitely regardless of data changes.

(d) The model's predictions may become unreliable or biased.

20. [2 points] [One Answer]

For a possibly non-convex optimization problem, gradient descent on the full dataset always finds a better solution than stochastic gradient descent.

(a) True

(b) False

21. [4 points] [Select All That Apply]

Given the gradient descent algorithm, $w_{t+1} = w_t - \eta \frac{df(w)}{dw}\Big|_{w=w_t}$, which of the following statement is correct regarding the hyperparameter η?

(a) η controls the magnitude of each step.

(b) η determines the initial value of w.

(c) A larger η guarantees faster convergence to the global minimum.

(d) A smaller η guarantees faster convergence to the global minimum.

22. [4 points] [Select All That Apply]

Which of the following functions are convex?

(a) $f(x) = x^3$

(b) $f(x) = \frac{3x(x-1)}{2}$

(c) $f(x) = \sin x$, for $x \in [\pi, 2\pi]$

(d) $f(x) = \log_{10}(x)$

23. [4 points] [Select All That Apply]

Which of the following are **true** about a convex function $f(x) : \mathbb{R}^d \to \mathbb{R}$?

(a) $f(x)$ must be differentiable across its entire domain.

(b) $f(x)$ has a unique global minimum.

(c) $g(x) = -f(x)$ is also convex.

(d) If $f(x)$ is twice differentiable, then $z^\top \nabla^2 f(x) z \geq 0$ for all $z \in \mathbb{R}^d$.

24. [5 points] [Select All That Apply]

Which of the following have convex objective functions?

(a) Linear regression

(b) Linear regression with arbitrary nonlinear basis functions

(c) Ridge regression

(d) Lasso regression

(e) Logistic regression

25. [5 points] [Select All That Apply]

Which of the following scenarios are better suited for a logistic regression model over a linear regression model?

(a) Forecasting the price of stocks for the next year, given the price of stocks for the past year.

(b) Diagnosing the presence or absence of a rare disease, given a medical x-ray.

(c) Predicting what the average global temperature will be in the next year, given historical climate data.

(d) Predicting how likely a student is to successfully complete a 4-year college degree, given their high school grades.

(e) Predicting the hardness of a material on a scale of 1-10 given the molecular structure of the material.

26. [4 points] [Select All That Apply]

Which of the following statements about classification are true?

Recall that the softmax function $\sigma : \mathbb{R}^k \to (0,1)^k$ takes a vector $z \in \mathbb{R}^k$ and returns a vector $\sigma(z) \in (0,1)^k$ with

$$\sigma(z)_i = \frac{\exp(z_i)}{\sum_{j=1}^{k} \exp(z_j)}.$$

(a) Consider a binary classification setting. If the data is linearly separable, we can use a logistic regression model with quadratic features to avoid overfitting.

(b) Because binary logistic regression is a convex optimization problem, it has a closed form solution.

(c) The softmax function is used when we are classifying $k > 2$ classes. When we are classifying only $k = 2$ classes, softmax regression will overfit, so we use binary logistic regression instead.

(d) We can use linear regression to solve classification problems, though the model we learn might not be as accurate compared to using logistic/softmax regression.