CSE 446/546 Winter 2024 Midterm Exam

February 7, 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 Positive is 3/5. The probability of a correct classification given that the sample is Positive is 4/5, and the probability of a correct classification given that the sample is Negative is 7/10. What is the probability that a randomly selected sample is Positive given that it has been classified as Positive? $\boxed{\text{One Answer}}$

   ○ (a) $\frac{4}{5}$
   ○ (b) $\frac{12}{25}$
   ○ (c) $\frac{3}{5}$
   ○ (d) $\frac{12}{19}$

2. Which of the following statements must be true for a square matrix $\mathbf{A}$ to have an inverse matrix $\mathbf{A}^{-1}$?

   ○ (a) $\mathbf{A}$ must be symmetric.
   ○ (b) The rank of $\mathbf{A}$ is less than its number of columns.
   ○ (c) $\mathbf{A}$ must have at least one column of 0s.
   ○ (d) The determinant of $\mathbf{A}$ is not equal to 0.

3. Consider the following system of linear equations:

$$2x + 3y = 16$$

$$4x + 6y = 32$$

   Which of the following statements are true regarding the system above?

   ○ (a) The system has an infinite number of solutions because the two equations are linearly dependent.
   ○ (b) The system has a unique solution because there are two equations for two unknowns.
   ○ (c) The system has no solution because the determinant of the coefficient matrix is zero.
   ○ (d) The system has no solution because the equations represent parallel lines that never intersect.

4. We define the gradient $\nabla_w f(w) = \begin{bmatrix} \frac{\partial f(w)}{\partial w_1} & \cdots & \frac{\partial f(w)}{\partial w_n} \end{bmatrix}^\top$ for any function $f : \mathbb{R}^n \to \mathbb{R}$. What is the value of $\nabla_w(w^\top Aw + u^\top Bw + w^\top Bv)$? Here $A, B \in \mathbb{R}^{n \times n}$, A is symmetric, and $u, v \in \mathbb{R}^n$. One Answer

   (a) $Aw + Bu + B^\top v$

   (b) $2Aw + B^\top u + Bv$

   (c) $Aw + B^\top u + Bv$

   (d) $2Aw + Bu + B^\top v$

5. Consider the principle of Maximum Likelihood Estimation (MLE) in statistical modeling. MLE is a method used to estimate the parameters of a statistical model. Based on this method, which of the following statements is most accurate? One Answer

   (a) MLE identifies the model parameters that maximize the probability of the observed data under the model.

   (b) MLE is used to directly compute the probability of the parameters being correct, independent of the observed data.

   (c) MLE is primarily concerned with minimizing the variance of the parameter estimates to achieve model stability.

   (d) MLE identifies the model parameters that minimize the squared prediction error over the training data.

6. A machine learning engineer is modeling the number of requests to his website per hour with a Poisson distribution model. If the engineer observed 4, 5, 6, 7, 8, and 9 requests per hour in the past 6 hours, what is the maximum likelihood estimation of the rate parameter λ of the Poisson distribution? Recall that Poisson distribution with parameter λ assigns every non-negative integer x probability $\mathbb{P}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$. One Answer

   (a) $e^{\frac{39}{6}}$

   (b) $\sqrt{\frac{39}{6}}$

   (c) 6

   (d) $\frac{39}{6}$

7. (2 points) Assume a simple linear model $Y = Xw$. For simplicity, no intercept is considered. Given the following dataset:

$$X = \begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix} \qquad\qquad Y = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

(a) (1 point) Compute the least squares estimate of w without any regularization. You may leave your answer as a fraction, if necessary.

Hint: if $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then $A^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

Answer: $\hat{w} =$ ______________________

(b) (1 point) Predict $\hat{Y}$ for $X = \begin{bmatrix} 6 \\ 7 \end{bmatrix}$.

Answer: $\hat{Y} =$ ______________________

8. Assume we have access to data $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Assume that we also have access to weights $\{w_i\}_{i=1}^n$, $w_i \in \mathbb{R}$ and $w_i > 0$, which gives the "importance" of each data point. Our goal is to solve the weighted least squares regression problem:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n w_i (x_i^\top \theta - y_i)^2.$$

Denote $X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$, and $W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix} \in \mathbb{R}^{n \times n}$.

What is $\widehat{\theta}$ in terms of X, Y, and W? One Answer

(a) $(X^\top X)^{-1} X^\top W^{-1} Y$

(b) $W(X^\top X)^{-1} X^\top Y$

(c) $(X^\top X)^{-1} X^\top W Y$

(d) $(X^\top W X)^{-1} X^\top Y$

(e) $(X^\top W X)^{-1} X^\top W Y$

9. In the context of least squares regression, how does the presence of high noise levels in the data impact the reliability of the model's parameter estimates? One Answer

(a) High noise levels predominantly affect the intercept term of the regression model, but leave the slope estimates relatively unaffected.

(b) High noise levels can increase the variability of the parameter estimates, potentially leading to a model that captures random noise rather than the true underlying relationship.

(c) High noise levels decrease the variance of the estimated parameters, making the model more robust.

(d) Noise in the data generally has minimal impact on the least squares estimates since the method inherently separates signal from noise in most scenarios.

10. In linear regression analysis using the least squares method, how might outliers in the dataset impact the resulting regression line?

    (a) Outliers affect only the precision of the prediction intervals, not the regression line itself.
    (b) Outliers enhance the model's accuracy by providing a wider range of data points.
    (c) Outliers can significantly skew the regression line, potentially leading to an inaccurate representation of the overall data trend.
    (d) Outliers have a minimal impact, as the least squares method averages out their effects.

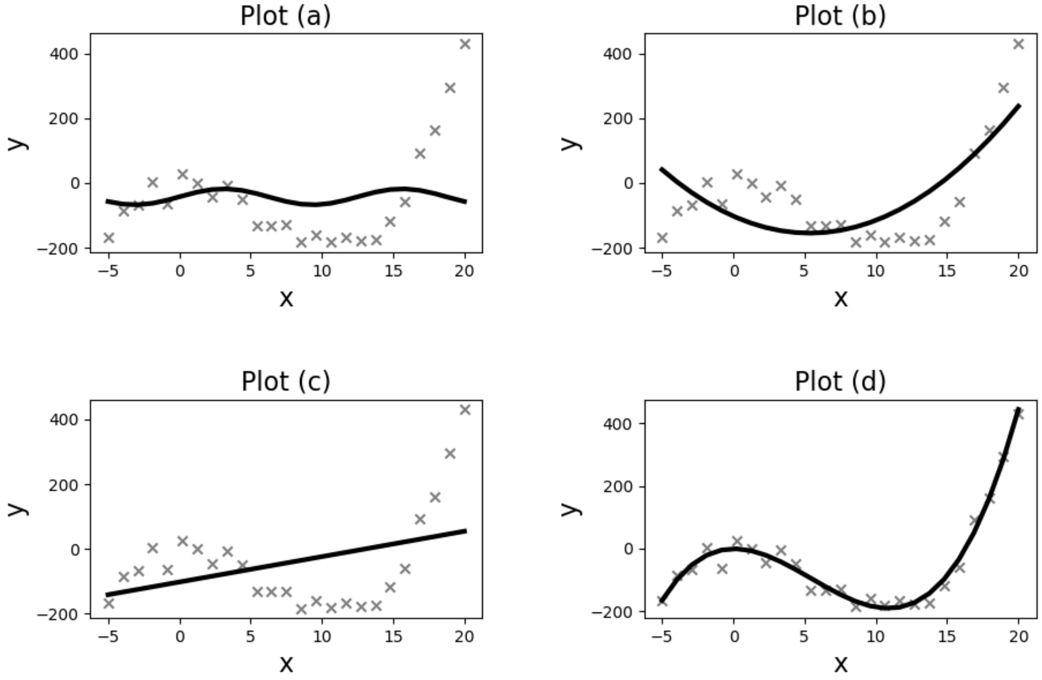
11. How does increasing the complexity of a model typically affect the properties of that model? Select all that apply.

    (a) It tends to decrease bias but increase variance, potentially leading to overfitting.
    (b) It can increase training accuracy.
    (c) It tends to increase both bias and variance.
    (d) It tends to decrease variance but increase bias, potentially leading to underfitting.

12. True/False: A model with high variance tends to perform well on both the training and test data.

    (a) False
    (b) True

13. The plots below show fits (in black) to the data points (“x” symbols in grey), using several different basis functions:



The basis functions used are:

1. $h_1(x) = [1, x]$
2. $h_2(x) = [1, x, x^2]$
3. $h_3(x) = [1, x, x^2, x^3]$
4. $h_4(x) = [1, \sin(\frac{4\pi}{25}x)]$

For each plot, please identify the basis function used:

Plot (a): _______, Plot (b): _______, Plot (c): _______, Plot (d): _______

14. In the context of linear regression, general basis functions are used to:

(a) Increase the speed of convergence in gradient descent optimization.

(b) Encourage sparsity in the learned weights.

(c) Minimize the computational complexity of linear regression models.

(d) Transform the input data into a higher-dimensional space to capture non-linear relationships.

15. Which statement best describes 'irreducible error' of a machine learning predictor? One Answer

(a) Irreducible error is due to the inherent noise in the data that cannot be eliminated by any model, regardless of model complexity.

(b) It is the error that is minimized by performing cross-validation.

(c) It is the error that can be minimized by increasing the size of the training dataset.

(d) It represents the error that arises due to the process of feature engineering and potentially including too many features that aren't relevant.

16. Suppose you train a polynomial regression model of degree $d = 3$ to approximate the quadratic function $g(x) = 7x^2 + \varepsilon$, where ε is Gaussian random variable with $\mu = 0$ and $\sigma^2 = 4$. What is the irreducible error? One Answer

(a) 2

(b) 0

(c) 4

(d) x^3

17. True/False: Increasing the proportion of your dataset allocated to training (as opposed to testing) will guarantee better performance on unseen data. One Answer

(a) False

(b) True

18. Which of the following statements best describes a potential issue that can arise if the test dataset is not properly separated from the training dataset? One Answer

    (a) The model will always underfit, regardless of the algorithm used.

    (b) The model will always overfit, regardless of the algorithm used.

    (c) The evaluation metrics will tend to overestimate the prediction error on unseen data.

    (d) The test data will influence the training process, leading to an overly optimistic estimate of the model's performance on new, unseen data.

    (e) The model's computational complexity will significantly increase, resulting in longer training times.

19. How should data preprocessing be applied when using k-fold cross-validation? Select the most accurate answer. One Answer

    (a) Preprocess the entire dataset before splitting into folds to maintain consistency.

    (b) Avoid preprocessing as it can bias the cross-validation results.

    (c) Only preprocess the test folds and train our model on raw (unprocessed) data.

    (d) Apply preprocessing separately on each iteration of k-fold validation to avoid data leakage

20. What is the main advantage of using k-fold cross-validation? One Answer

    (a) It guarantees improvement in model accuracy on unseen data.

    (b) It provides an estimate of model performance for given hyperparameters.

    (c) It significantly reduces the training time of the model by dividing the dataset into smaller parts.

    (d) It eliminates the need for a separate test dataset.

21. In Lasso regression, how does the regularization parameter λ influence the risk of overfitting? Select all that apply.

    (a) Increasing λ always increases the risk of overfitting as it leads to higher model complexity.
    (b) Decreasing λ to zero may increase the risk of overfitting.
    (c) Increasing λ typically reduces the risk of overfitting by increasing sparsity.
    (d) The choice of λ in Ridge regression has no impact on the risk of overfitting.

22. When comparing Lasso regression to Ridge regression, which of the following properties are true about Lasso regression? Select all that apply.

    (a) Lasso regression can be used to select the most important features of a dataset.
    (b) Lasso regression tends to retain all features but with smaller coefficients.
    (c) Lasso regression is always better suited for handling high-dimensional data with a large number of features.
    (d) Lasso regression has fewer hyperparameters to tune.

23. An enterprising 446/546 student is using ridge regression to predict housing prices based on features such as square footage, number of bedrooms, and proximity to schools. They notice that increasing the regularization strength improves model performance on the validation set but worsens it on the training set. What does this observation suggest about their model before adjusting the regularization strength?

    (a) The choice of features was inappropriate for predicting housing prices.
    (b) The model was underfitting the training data.
    (c) The regularization strength was too high, leading to loss of critical information.
    (d) The model was overfitting the training data.

24. Consider some convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, and assume that f is twice-differentiable. What can we say about $\nabla_x^2 f(x) \in \mathbb{R}^{d \times d}$, the Hessian of $f(x)$? (Hint: think about the case when $d = 1$, and what the second derivative of f says about its shape.) $\boxed{\text{One Answer}}$

(a) $\nabla_x^2 f(x)$ is negative semi-definite.

(b) $\nabla_x^2 f(x)$ is negative definite.

(c) $\nabla_x^2 f(x)$ is positive definite.

(d) $\nabla_x^2 f(x)$ is positive semi-definite.

25. True/False: A solution to a convex optimization problem is guaranteed to be a global minimum. $\boxed{\text{One Answer}}$

(a) True

(b) False

26. True/False: A solution to a convex optimization problem is guaranteed to be unique. $\boxed{\text{One Answer}}$

(a) True

(b) False

27. True/False: A convex optimization problem is guaranteed to have a closed-form solution. $\boxed{\text{One Answer}}$

(a) True

(b) False

28. Briefly explain the main difference between Mini Batch Gradient Descent and Stochastic Gradient Descent.

Then, describe one main advantage of using Mini Batch Gradient Descent over SGD.

Answer: ____

____

____

29. True/False: Stochastic gradient descent provides biased estimates of the true gradient at each step. One Answer

(a) True

(b) False

30. Consider some function $f(x) : \mathbb{R}^d \to \mathbb{R}$, and assume that we want to run an iterative algorithm to find the *maximizer* of f. Which update rule should we use to do this (for some $\eta > 0$)? One Answer

(a) $x_{t+1} \leftarrow -x_t + \eta \cdot \nabla_x f(x_t)$.

(b) $x_{t+1} \leftarrow x_t - \eta \cdot \nabla_x f(x_t)$.

(c) $x_{t+1} \leftarrow -x_t - \eta \cdot \nabla_x f(x_t)$.

(d) $x_{t+1} \leftarrow x_t + \eta \cdot \nabla_x f(x_t)$.

31. You run a social media platform and are planning to implement a system to combat the spread of misinformation by detecting fake news articles. To keep things simple, the system only needs to identify articles as one of two classes: (1) being fake news, or (2) not being fake news. Of the model types we have learned in class so far, which would be the best choice to implement this system?

Answer: ____________________________