CSE 446 Winter 2023 Midterm Exam

February 10, 2023

Please WAIT to open the exam until you are instructed to begin. You can write your name on this page.

Please write your name and ID on your notes page (if you have one). We will collect this with your exam.

Please take out your student ID and leave it on the corner of your desk, as we will come around and check them while you work on the exam.

Instructions: This exam consists of a set of True/False and multiple choice questions.

Write your name and ID number in the provided spaces on every page of the exam.

For each question, clearly indicate your answer by filling in the letter associated with your choice.

If you need to change an answer, please very clearly indicate what your final answer is. Responses where we cannot determine the selected option will be marked as incorrect.

1. True/False: Leave-one-out (LOO) and k-fold cross-validation can be used for hyperparameter tuning.
   (a) True
   (b) False

2. Which of the following is most indicative of a model overfitting?
   (a) High bias, low variance
   (b) Low bias, high variance
   (c) Low bias, low variance

3. Which of the following statements about LASSO is true?
   (a) LASSO's objective function has a closed-form solution.
   (b) LASSO has lower bias than ordinary least squares.
   (c) LASSO can be interpreted as least squares regression when the model's weights are regularized with the l_1 norm.
   (d) LASSO can be interpreted as least squares regression when the model's weights are regularized with the l_2 norm.

4. Which of the following is not a convex set?
   (a) The hyperplane given by $H = \{\mathbf{x} \in \mathbb{R}^{\mathbf{n}} : \sum_{\mathbf{i=1}}^{\mathbf{n}} \alpha_{\mathbf{i}} \mathbf{x_i} = \beta_{\mathbf{i}}\}$
   (b) The interval $[a, b]$ where $a, b \in \mathbb{R}$
   (c) The "unit square" $\{x \in \mathbb{R}^2 : ||x||_1 = 1\}$
   (d) The unit ball $\{x \in \mathbb{R}^2 : ||x||_2 \leq 1\}$

5. **Extra Credit:** Consider a data matrix $X \in \mathbb{R}^{n \times m}$, target vector $y \in \mathbb{R}^n$, and the resulting least squares solution $\hat{w} \in \mathbb{R}^m$. Now let y' be the vector that results from squaring every value in the target vector y, and let $\hat{w}'$ be the vector that results from squaring every value in $\hat{w}$.

$$y' = [y_1^2, \ldots, y_n^2] \text{ and } \hat{w}' = [\hat{w}_1^2, \ldots, \hat{w}_m^2]$$

True/False: If we leave the data matrix X unchanged and we use y' as our new target vector, the resulting least squares solution will be $\hat{w}'$.

(a) False

(b) True

6. Reducing the regularization of a model would typically . . .

(a) Decrease its bias and increase its variance

(b) Decrease its bias and decrease its variance

(c) Increase its bias and decrease its variance

(d) Increase its bias and increase its variance

7. How many models must be trained when using k-fold cross-validation to determine which of three possible λ values $(\lambda_1, \lambda_2, \lambda_3)$ is best for ridge regression on training set with n samples (assume n is a multiple of k)?

(a) $3n/k$

(b) k

(c) n

(d) $3k$

8. k-fold cross-validation is equivalent to leave-one-out (LOO) cross-validation on a training set of n samples when k is equal to

(a) k is not computable

(b) $n - 1$

(c) n

(d) 1

9. Let $X \in \mathbb{R}^{m \times n}$, and $Y \in \mathbb{R}^m$. We want to fit a linear regression model. We call a matrix a "short wide" matrix if there are more columns than rows. Which of the following is **NOT** always true when X is a "short wide" matrix (i.e., $n > m$):
   (a) $X^\top X$ is symmetric and positive semidefinite.
   (b) $X^\top X$ is not invertible.
   (c) The columns of X are linearly independent.
   (d) The null space of X is non-empty.

10. Assume you (1) standardized a training set and (2) trained a machine learning model on this standardized training set. Before you use your model to make predictions on a test set, you should do which of the following (choose exactly one answer)
   (a) not standardize the test set.
   (b) use the mean and standard deviation from train set to standardize the test set.
   (c) use the mean and standard deviation from test set to standardize the test set.
   (d) collect new data and use the new data's mean and standard deviation to standardize the test set.

11. Let $x_1, x_2, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is an unknown variable. The PDF of $\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ for any $x \in \mathbb{R}$. Using the log-likelihood, find the maximum likelihood estimation of μ in terms of x_i.
   (a) $\frac{1}{\sum_{i=1}^n x_i}$
   (b) $\frac{1}{n}\sum_{i=1}^n x_i$
   (c) $\sum_{i=1}^n \frac{x_i}{\sigma^2}$
   (d) $\sigma \sum_{i=1}^n x_i$

12. True/False: We can make the irreducible error smaller by using a larger number of training samples.
   (a) True
   (b) False

13. Let $f(x_1, x_2, x_3) = x_1x_2 - x_2^3 + x_1x_3$. What is $\nabla_{x_1,x_2,x_3} f$?
    (a) $x_2 - 3x_2^2 + x_1$
    (b) $[x_2 + x_3, x_1 - 3x_2^2, x_1]$
    (c) $x_2 + x_3$
    (d) $[x_2, -3x_2^2, x_1]$

14. True/False: Convex optimization problems are attractive because they always have exactly one global minimum.
    (a) True
    (b) False

15. Ridge regression
    (a) reduces variance at the expense of bias
    (b) adds an l_1 penalty norm to the cost function
    (c) often sets many of the weights to 0 when the regularization parameter λ is very large
    (d) is more sensitive to outliers than least squares

16. For a linear regression model, start with random values for each coefficient. The sum of the squared errors is calculated for each pair of input and output values. A learning rate is used as a scale factor, and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible. What is this process called?
    (a) LASSO
    (b) Gradient Descent
    (c) Least squares
    (d) Regularization

17. Let $X \in \mathbb{R}^{m \times n}$, $w \in \mathbb{R}^n$, and $Y \in \mathbb{R}^m$. Consider mean squared error $L(w) = \|Xw - Y\|_2^2$. What is $\nabla_w L(w)$?

(a) $2Y^T(X^\top Xw - Y)$

(b) $2X^T(X^\top Xw - Y)$

(c) $2Y^T(Xw - Y)$

(d) $2X^T(Xw - Y)$

18. Write down a closed-form solution for the optimal parameters w that minimize the loss function

$$L(w) = \sum_{i=1}^{n} (y_i - x_i^T w)^2 + \lambda ||w||_2^2$$

in terms of (1) the $n \times d$ matrix X whose i-th row is a $1 \times n$ vector x_i^T (a sample), (2) the $n \times 1$ vector y whose i-th entry is y_i, and (3) the scalar λ. (You may assume that any relevant matrix is invertible.)

(a) $\widehat{w} = (X^T X)^{-1} Xy + \lambda I$

(b) $\widehat{w} = 2(X^T X + \lambda I)^{-1} X^T y$

(c) $\widehat{w} = \lambda (X^T X)^{-1} X^T y$

(d) $\widehat{w} = (X^T X + \lambda I)^{-1} X^T y$

19. How can overfitting be reduced in polynomial regression?

(a) By decreasing the size of the validation set during hyperparameter tuning.

(b) By increasing the degree of the polynomial.

(c) By using regularization techniques such as l_1 or l_2 regularization.

(d) By reducing the size of your training set.

20. True/False: Linear least squares has a nonconvex loss function.

(a) True

(b) False

21. True/False: It is possible to apply gradient descent method on linear least squares loss.
    (a) True
    (b) False

22. Let $x_1, x_2 \in \mathbb{R}_+$ be sampled from the distribution $\text{Exp}(\lambda)$, where $\lambda \in \mathbb{R}_+$ is an unknown variable. Remember that the PDF of the exponential distribution is $f(x) = \lambda e^{-\lambda x}$ for any $x > 0$ and $f(x) = 0$ otherwise. Using the log-likelihood, find the maximum likelihood estimation of λ in terms of x_1, x_2. Hint: $\frac{d}{dx} e^x = e^x$.
    (a) $\frac{\log(x_1)+\log(x_2)}{2}$
    (b) $\log\left(\frac{e^{x_1}+e^{x_2}}{2}\right)$
    (c) $\frac{x_1+x_2}{2}$
    (d) $\frac{2}{x_1+x_2}$

23. **Extra Credit:** You are taking a multiple-choice exam that has 4 answers for each question. You are a smart student, so in answering a question on this exam, the probability that you know the correct answer is p, and you always choose the correct answer when you know it. If you don't know the answer, you choose one (uniformly) at random. What is the probability that you knew the correct answer to a question, given that you answered it correctly?
    (a) $\frac{p+\frac{1-p}{4}}{p}$
    (b) $\frac{p}{1+p}$
    (c) $\frac{p}{p+\frac{1-p}{4}}$
    (d) $\frac{p}{\frac{p}{4}+1}$

24. **Select all** of the following statements that are **False**. When training a machine learning model you should
    (a) manually select samples from your data to form a test set.
    (b) use a test set to help choose hyperparameter values.
    (c) never use the test set to make changes to the model.
    (d) split your data into training and test sets.

25. Let $X \sim \text{Uniform}[0, 3]$ and $Y \sim \mathcal{N}(2, 2)$ be independent random variables. Then compute $E[XY^2] - E[X]E[Y]^2$.
    (a) 4
    (b) 6
    (c) 0
    (d) 3

26. (True/False:) Stochastic Gradient Descent (SGD) will always be at least as computationally expensive as Gradient Descent (GD) and (True/False:) the number of update steps in SGD is greater than or equal to the number of update steps in GD.
    (a) True, True
    (b) True, False
    (c) False, True
    (d) False, False

27. Which technique is most likely to reduce the variance of a model, holding all else fixed?
    (a) Reducing the complexity of the model
    (b) Using a smaller number of training samples
    (c) Increasing the number of features