CSE 446/546 Autumn 2023 Midterm Exam

November 1, 2023

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

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).

- Write your name and UW NetID (<netID>@uw.edu) in the provided spaces on every page of the exam.
- For each multiple choice and True/False question, clearly indicate your answer by filling in the letter(s) associated with your choice.
- 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.

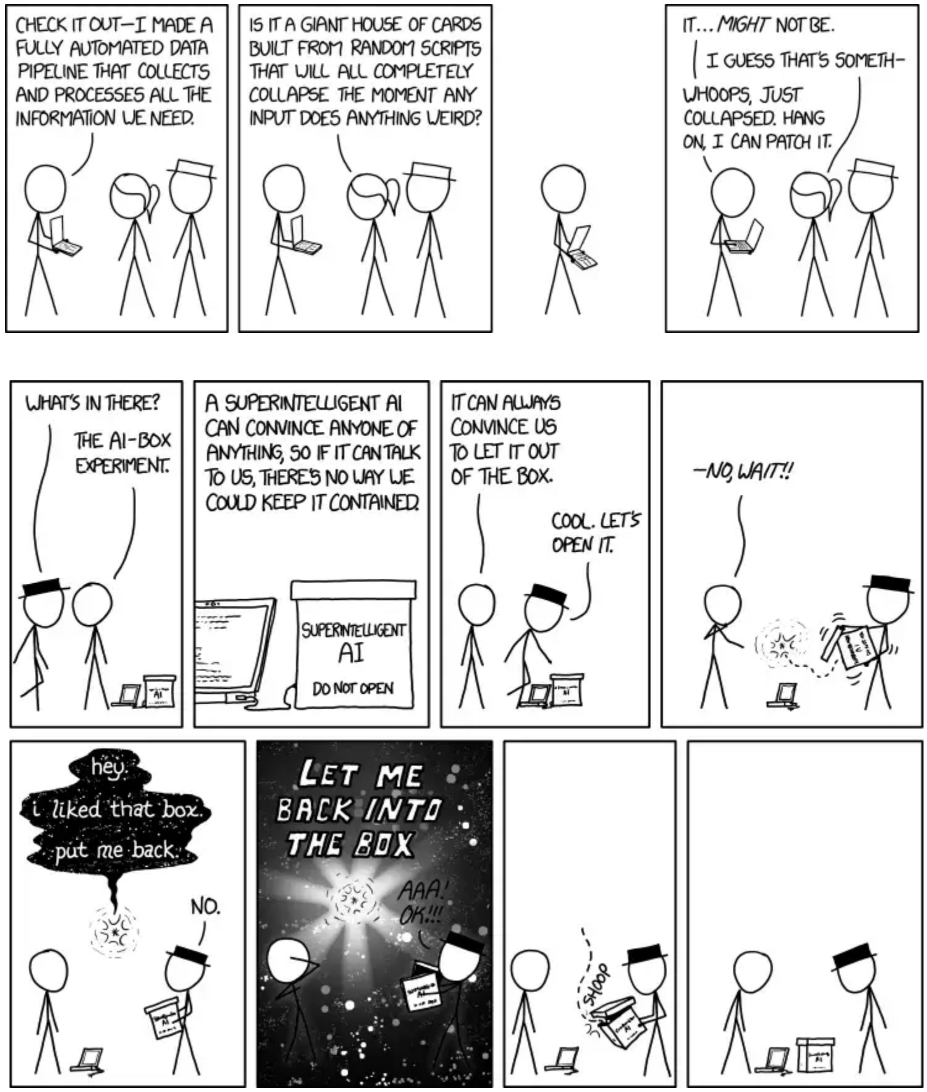


Figure 1: These images are included only to cover the back of this page. They have no relation to the exam.

1. Which of the following is the definition of irreducible error in machine learning?

   (a) The error that cannot be eliminated by any model

   (b) The error that is caused by overfitting to the training data

   (c) The error that is caused by underfitting to the testing data

   (d) All of the above

2. What is the general model for $\mathbb{P}(Y = 1|X = x, \theta)$ in logistic regression, where $X = (X_0, X_1, ..., X_n)$ is the features, Y is the predictions, and θ is the parameters? Assume that a bias term has already been appended to X (i.e., $X_0 = \mathbf{1}$).

   (a) $\mathbb{P}(Y = 1|X = x, \theta) = \frac{1}{1+e^{-\theta^\top x}}$

   (b) $\mathbb{P}(Y = 1|X = x, \theta) = \theta^\top x$

   (c) $\mathbb{P}(Y = 1|X = x, \theta) = \log(1 + e^{-\theta^\top x})$

   (d) $\mathbb{P}(Y = 1|X = x, \theta) = \log(1 + e^{\theta^\top x})$

3. Two realtors are creating machine learning models to predict house costs based on house traits (i.e. house size, neighborhood, school district, etc.) trained on the same set of houses, using the same model hyperparameters. Realtor A includes 30 different housing traits in their model. Realtor B includes 5 traits in their model. Which of the following outcomes is most likely?

   (a) Realtor B's model has higher variance and lower bias than Realtor A's model

   (b) Realtor A's model has higher variance than Realtor B's model and without additional information, we cannot know which model has a higher bias

   (c) Realtor A's model has higher variance and lower bias than Realtor B's model

   (d) Realtor A's model has higher variance and higher bias than Realtor B's model

4. When $\mathcal{L}(w, b) = \sum_{i=1}^{n}(y_i - (w^\top x_i + b))^2$ is used as a loss function to train a model, which of the following is true?

   (a) It minimizes the sum of the absolute differences between observed and predicted values.
   (b) It maximizes the correlation coefficient between the independent and dependent variables.
   (c) It requires the use of gradient descent optimization to find the best-fit line.
   (d) It minimizes the sum of the squared difference between observed and predicted values.

5. True/False: As the value of the regularization term coefficient in Ridge Regression increases, the sensitivity of predictions to inputs decreases.

   (a) True
   (b) False

6. Which of the following statements about logistic regression is true?

   (a) The loss function of logistic regression without regularization is convex, and the loss function of logistic regression with L2 regularization is convex.
   (b) Neither the loss function of logistic regression without regularization is convex nor the loss function of logistic regression with L2 regularization is convex.
   (c) The loss function of logistic regression without regularization is convex, but the loss function of logistic regression with L2 regularization is non-convex.
   (d) The loss function of logistic regression without regularization is non-convex, but the loss function of logistic regression with L2 regularization is convex.

7. Which of the following is NOT an assumption of logistic regression?

   (a) The output target is binary.
   (b) The input features can be continuous or categorical.
   (c) The residual errors are normally distributed.

8. Suppose we've split a dataset into train, validation, and test sets; trained a regression model on the train set; and found the optimal value for a regularization constant λ. **Select all** of the regression methods for which adding the validation set into the train set and retraining can change the optimal value for λ.

   (a) LASSO regression

   (b) Ridge regression

9. Suppose that we want to estimate the ideal parameter θ^* for $p(x, y, \theta)$ given a set of observations $\{x_i, y_i\}$. Which of the following is a key assumption made when using $\hat{\theta}_{MLE} = \arg\max_\theta \sum_i \log(p(x_i, y_i|\theta_i))$ for Maximum Likelihood Estimation (MLE) to estimate the model parameter?

   (a) The data is normally distributed.

   (b) The data is independent and identically distributed (i.i.d.).

   (c) The data contains no outliers.

   (d) The data is linearly separable.

10. Provide one advantage and one disadvantage of Stochastic Gradient Descent (SGD) over Gradient Descent (GD).

    Answer: ______________________________

    ____

    ____

11. (2 points) Assume a simple linear model $Y = \beta_1 X$. For simplicity, no intercept is considered. Given the following dataset:

$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad Y = \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}$$

(a) (1 point) Compute the least squares estimate of β_1 without any regularization. You may leave your answer as a fraction, if necessary.

Answer: $\hat{\beta}_1 =$ ______________________

$$L(\beta_1) = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 + \alpha \|\beta_1\|_1 \tag{11}$$

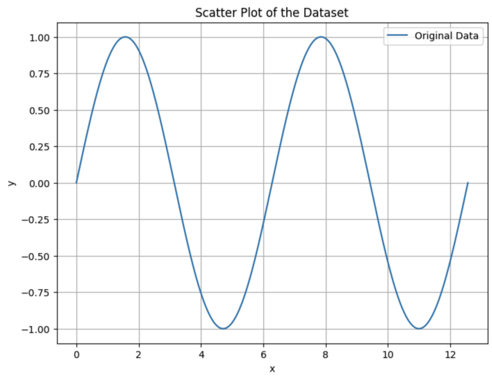
(b) (1 point) Using Lasso Regression (equation 11) with a penalty term $\alpha = 2$, would β_1 increase or decrease? Provide a short explanation.

Answer: ____

____

____

12. Suppose you're given a scatter plot of a dataset, and the pattern appears to be a periodic wave-like curve that repeats itself at regular intervals.



Which of the following basis functions might be most appropriate to capture the relationship between x and y for this dataset?

(a) Polynomial basis functions: $\phi(x) = \{1, x, x^2, x^3, ...\}$

(b) Radial basis functions: $\phi(x) = \exp(-\lambda||x - c||^2)$

(c) Fourier basis functions: $\phi(x) = \{1, \sin(\omega x), \cos(\omega x), \sin(2\omega x), \cos(2\omega x), ...\}$

(d) Logarithmic basis function: $\phi(x) = \log(x)$

(e) Exponential basis function: $\phi(x) = \exp(\lambda x)$

13. Which of the following statements about convexity is true?

(a) If $f(x)$ is convex, then $g(x) = \frac{1}{3}f(x)$ is also convex

(b) If $f(x)$ is convex, then gradient descent on minimizing $f(x)$ will always reach global minimum

(c) If $f(x)$ is convex, then $f(x)$ is everywhere differentiable

14. Suppose you are provided with a dataset of n independently sampled, 1-dimensional data points $X = \{x_1, \dots x_n\}$ that you believe follows a univariate Gaussian distribution. You compute the sample mean $\bar{x}$. What are the unbiased maximum likelihood estimates (MLE) for the parameters (μ, σ) of the univariate Gaussian?

    (a) $\hat{\mu}_{MLE} = \bar{x}$, $\hat{\sigma}^2_{MLE} = \frac{1}{n}(\Sigma_{i=1}^n x_i)$
    (b) $\hat{\mu}_{MLE} = \bar{x}$, $\hat{\sigma}^2_{MLE} = \frac{1}{n}(\Sigma_{i=1}^n (x_i - \hat{\mu}_{MLE})^2)$
    (c) $\hat{\mu}_{MLE} = \bar{x}$, $\hat{\sigma}^2_{MLE} = \frac{1}{n-1}(\Sigma_{i=1}^n (x_i - \hat{\mu}_{MLE})^2)$
    (d) $\hat{\mu}_{MLE} = \frac{1}{n}\bar{x}$, $\hat{\sigma}^2_{MLE} = \frac{1}{n-1}(\Sigma_{i=1}^n (x_i - \hat{\mu}_{MLE})^2)$

15. True/False: When performing gradient descent, decreasing the learning rate enough will slow down convergence but will eventually guarantee you arrive at the global minimum.

    (a) True
    (b) False

16. Which of the following functions is strictly convex over its entire domain?

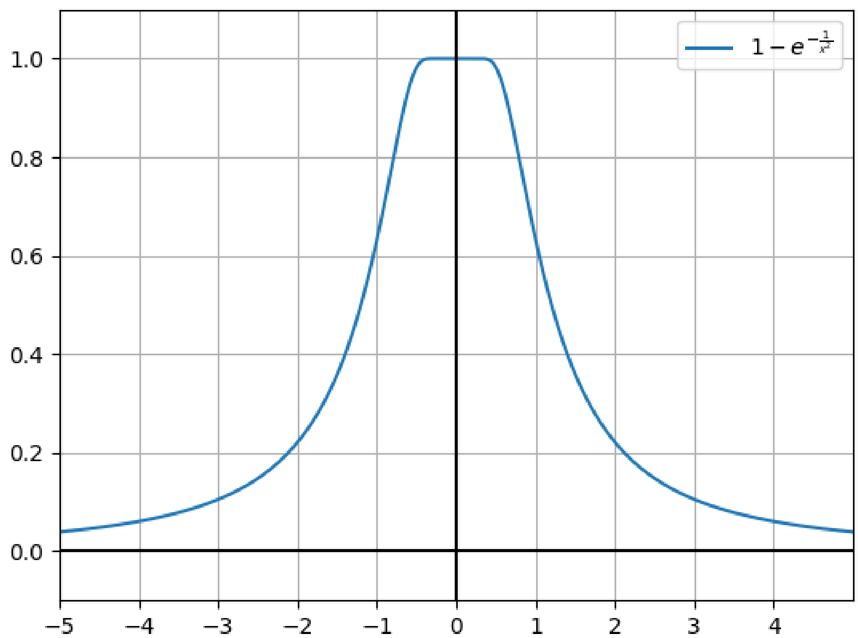
    (a) $f(x) = -x^2$
    (b) $f(x) = x^3$
    (c) $f(x) = \ln(x)$
    (d) $f(x) = e^x$

17. Which of the following is true about a validation set and how it is used?

    (a) The validation set allows us to estimate how a model would perform on unseen data
    (b) When deciding to use a validation set, you do not need a separate test set
    (c) After hyperparameter tuning, the validation set is always added back into the training set before training the final model
    (d) The validation set allows us to train a model quicker by decreasing the size of our training data set

18. (2 points) Suppose we have the function

$$f(x) = \begin{cases} 1 - e^{-\frac{1}{x^2}} & x \neq 0 \\ 1 & x = 0 \end{cases}$$



(a) (1 point) Suppose that we perform gradient descent starting at $x_0 = 0$ with step size $\eta = 1$. what is the asymptotic behavior of gradient descent given by Equation 12?

$$x_{n+1} = x_n - \eta f'(x_n) \tag{12}$$

Answer: ____

(b) (1 point) Now suppose that $x_0 \sim \mathcal{N}(0, \epsilon)$ for some small ϵ. What is the behavior then?

Answer: ____

19. A bag contains 4 red balls and 3 green balls. We draw 3 balls from the bag without replacement. What is the probability that all 3 balls are red? Express your result as a fraction, or as a percentage rounded to the integer percentage (e.g. 77%).

    Answer: ______________________

20. True/False: For a matrix $X \in \mathbb{R}^{n \times d}$ of rank d, there exists an orthogonal matrix V and diagonal matrix D such that $X^\top X = VDV^\top$.

    (a) True

    (b) False

21. You have built a spam detection classifier to help you clean up your email inbox. Your system has uncovered that 90% of all spam emails contain the word "discount". If you assume that the overall probability of an email being spam is 5% and 15% of all incoming emails contain the word "discount", what is the probability that an email containing "discount" is actually spam?

    (a) 0.9

    (b) 0.135

    (c) 0.3

    (d) 0.045

22. Determine if the following two statements are true or false.

    (1) True/False: For large datasets with n samples, it is recommended to use k-fold cross-validation with a value of k that is close to n.
    (2) True/False: In k-fold cross-validation, a larger value of k results in a more computationally efficient process, as it requires fewer model training.

    (a) (1) False, (2) True
    (b) (1) False, (2) False
    (c) (1) True, (2) True
    (d) (1) True, (2) False

23. In Lasso regression, what does the L1 regularization term primarily encourage?

    (a) Encourages the model to fit the training data more closely.
    (b) Encourages the model to have large coefficients for all features.
    (c) Encourages the model to have small but non-zero coefficients for all features.
    (d) Encourages sparsity by driving some feature coefficients to zero.

24. You are provided with a training dataset of i.i.d. input-output pairs $\{(x_i, y_i)\}_{i=1}^n$ and you choose to fit a linear model by minimizing the least squares objective $\hat{w} = \arg\min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$ .Which of the following statements is true?

    (a) The least squares objective is equivalent to maximizing the likelihood function of the observed data assuming Gaussian noise.
    (b) The least squares objective is equivalent to minimizing the likelihood function of the observed data assuming Gaussian noise.
    (c) The least squares objective is equivalent to maximizing the likelihood function of the observed data assuming Laplace noise.
    (d) The least squares objective is equivalent to minimizing the likelihood function of the observed data assuming Laplace noise.

25. Consider a matrix $A \in \mathbb{R}^{n\times n}$ that is symmetric and has orthonormal columns. Which of the following statements is true?

   (a) All eigenvalues of A are real.

   (b) At least one eigenvalue of A is complex.

   (c) All eigenvalues of A are either 0 or 1.

   (d) The eigenvalues of A cannot be determined from the given information.

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

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

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

   Show that when $\lambda > 0$, the matrix $X^\top X + \lambda I$ is invertible.

   Answer:

This page is intentionally left blank as scratch paper.