CSE 446/546 Winter 2024 Final Exam

March 13, 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.

There are 43 questions on this exam. You will have 1 hour and 50 minutes to complete the exam.

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

- Each question is worth 1 point unless noted otherwise.
- 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 marked [Select All], in which case they are "select all that apply" and 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. One Answer In the context of logistic regression, which of the following statements is true about the interpretation of the model coefficients?

   (a) The coefficients represent the **change in the log odds** of the dependent variable for a one-unit change in the predictor variable, holding all other variables constant.

   (b) The coefficients represent the **change in the dependent variable** for a one-unit change in the predictor variable, holding all other variables constant.

   (c) The coefficients are directly proportional to the **probability of the dependent variable being 1**.

   (d) The coefficients represent the probability that the **predictor variable will be present** when the dependent variable is 1.

2. One Answer You are working on a machine learning project to classify emails as either spam (1) or not spam (0) using logistic regression. The model has been trained based on emails with labels and several features, including the frequency of specific keywords. For a particular new email, the model's output of the log-odds is 0.4.
   Given the model's output, which of the following options best describes its classification of the email?

   (a) The email is classified as not spam because a positive log-odds score indicates a higher likelihood of the email belonging to the negative class (not spam)

   (b) The email is classified as spam because the log-odds score is positive, indicating that the odds of the email being spam are greater than the odds of it not being spam.

   (c) The email is classified as not spam because the probability of being spam is less than 0.5.

   (d) The email is classified as spam because the probability of it being spam is positive.

3. Select All In the context of logistic regression used for binary classification, which of the following statements is true?

   (a) The model directly outputs class labels (0 or 1)

   (b) The model's optimization has a closed-form solution.

   (c) The model produces a linear decision boundary with respect to the features.

   (d) The model uses the softmax function to output class probabilities.

4. Select All Which are key properties of the Radial Basis Function kernel?

(a) It works best when features take on categorical values.

(b) It relies on the distance between points in the original feature space.

(c) It relies on the distance between points in infinite-dimensional space.

(d) It implicitly maps to an infinite-dimensional feature space.

(e) It identifies hyperplanes in an infinite-dimensional space.

5. One Answer Which of the following is **not** a valid kernel?

(a) $K(x, x') = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\|x - x'\|_2^2)$

(b) $K(x, x') = -\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\|x - x'\|_2^2)$

(c) $K(x, x') = x^\top x'$

(d) $K(x, x') = 1$

6. Select All Which of the following statements about the "kernel trick" are true in the context of machine learning algorithms?

(a) It provides an efficient method for computing and representing high-dimensional feature expansions.

(b) It implicitly maps to a high-dimensional feature space.

(c) It eliminates the need for regularization.

(d) It can only be used in regression prediction settings.

7. Select All Consider the kernel ridge regression problem.

$$\hat{w} = \arg\min_{w} \frac{1}{n}\sum_{i=1}^{n}(y_i - \phi(x_i)^\top w)^2 + \lambda\|w\|_2^2 \quad \text{becomes} \quad \hat{\alpha} = \arg\min_{\alpha} \|Y - K\alpha\|_2^2 + \lambda\alpha^\top K\alpha$$

Let $\phi(x) : \mathbb{R}^d \to \mathbb{R}^p$ be the feature mapping the kernel matrix K is with respect to. Let n be the number of samples we have. Which of the following statements is true?

(a) Ridge regression can only be kernelized assuming $\alpha \in \text{span}\{x_1, x_2, ..., x_n\}$ where $x_i \in \mathbb{R}^d$ denotes the ith training sample

(b) When $n \ll p$, the kernel method will be more computationally efficient than using regular ridge regression.

(c) There is no closed-form solution if K is positive definite.

(d) The optimal $\hat{w}$ can be obtained after solving for the optimal $\hat{\alpha}$ even though w is not explicitly included in the optimization problem

8. Select All Assume we have n samples from some distribution P_X, and wish to estimate the variance of P_X, as well as compute a confidence interval on the variance. If $n = 1$ and we draw only a single datapoint $X_1 = 2$ from P_X, which of the following are true?

(a) The bootstrap estimate of the variance is 0.

(b) The bootstrap estimate of the variance is 2.

(c) The bootstrap cannot be applied when we only have $n = 1$ samples.

(d) The bootstrap is likely to give a very poor estimate of the variance in this setting.

9. Select All Which of the following statements about the bootstrap method are true?

(a) It requires a large sample size to be effective and cannot be used effectively with small datasets.

(b) It involves repeatedly sampling with replacement from a dataset to create samples and then calculating the statistic of interest on each sample.

(c) Bootstrap methods can only be applied to estimate the mean of a dataset and do not apply to other statistics like median or variance.

(d) One of the advantages is that it does not make strong parametric assumptions about the distribution of the data.

(e) It can be used to estimate confidence intervals for almost any statistic, regardless of the original data distribution.

10. Select All Which of the following are advantages of using random forests over decision trees?

(a) The optimal decision tree cannot be efficiently computed, but the optimal random forest can.

(b) Random forests typically have smaller variance than decision trees.

(c) Random forests typically have smaller bias than decision trees.

(d) Random forests are less prone to overfitting compared to decision trees.

11. Select All Which of the following is true about k-nearest neighbors (KNN)?

(a) KNN works best with high dimensional data.

(b) When k=1, the training error is always less than or equal to the test error.

(c) The computational cost of making a prediction on a new test point increases with the size of the training dataset.

(d) The effectiveness of KNN is independent of the distance metric used.

12. One Answer For k-nearest neighbors (KNN), changing k will affect:

(a) Bias
(b) Variance
(c) Both bias and variance
(d) Neither bias nor variance

13. One Answer In k-nearest neighbors (KNN), having higher dimensional features is always more desirable because it provides more dimensions to calculate the distance between two data points.

(a) True
(b) False

14. One Answer The training algorithm for k-means clustering is guaranteed to converge to a local minimum of the k-means objective function.

(a) True
(b) False

15. One Answer The k-means objective function always improves with each successive iteration of the k-means training algorithm until the objective function converges.

(a) True
(b) False

16. One Answer In k-means clustering, choosing a different set of initial centroids always leads to the same final clusters after convergence.

(a) True
(b) False

17. One Answer In k-means clustering, if two points are assigned to the same cluster, any point that is a convex combination of those two points must also be assigned to that same cluster. [Hint: recall that a point x_3 is a convex combination of the points x_1 and x_2 if $x_3 = \alpha x_1 + (1-\alpha)x_2$ for some $0 \leq \alpha \leq 1$]

(a) True

(b) False

18. One Answer Recall that in a Gaussian mixture model (GMM) with $K = 2$ Gaussians, the "responsibilities" γ_{ik} indicate the probability that the i-th data point was generated by the k-th Gaussian. These responsibilities provide soft cluster assignments.

In a GMM, if two points x_1 and x_2 have responsibilities $\gamma_{1k} \geq p$ and $\gamma_{2k} \geq p$, respectively, then any point x_3 that is a convex combination of x_1 and x_2 must also have a responsibility $\gamma_{3k} \geq p$.

(a) True

(b) False

19. Select All What are advantages of using a Gaussian mixture model (GMM) to cluster data over k-means clustering?

(a) GMMs are better at handling non-spherical clusters.

(b) There is a closed-form solution that optimizes the GMM loss function, whereas k-means requires an iterative optimization algorithm.

(c) GMMs are better at modeling uncertainty of cluster assignments.

(d) There are no advantages of using a GMM over k-means clustering.

20. Select All What role(s) does the activation function in a neural network play?

(a) It determines the size of the neural network.

(b) It introduces non-linearity into the network.

(c) It directly minimizes the loss function during training.

(d) It calculates the gradient of the network's weights.

21. One Answer Consider a neural network being trained to minimize a loss function using backpropagation. If the learning rate is set too high, what is the most likely outcome during the training process?

(a) The network efficiently converges to the global minimum of the loss function, resulting in optimal training performance.

(b) The network's weights oscillate or diverge, potentially causing the loss to increase rather than decrease.

(c) The network immediately overfits to the training data, leading to poor generalization on unseen data.

(d) The network's training process significantly slows down, requiring more epochs to reach convergence.

22. One Answer Increasing the number of hidden layers in a neural network will always decrease prediction error on the training data.

(a) True

(b) False

23. One Answer When designing a neural network, the depth of the network (i.e., the number of hidden layers) plays a critical role in its performance. Which of the following statements best describes the impact of increasing the network's depth?

(a) Adding more layers to the network always results in better performance on the training data, regardless of the activation function (e.g., linear, ReLU, sigmoid, etc.).

(b) Increasing the depth of the network linearly improves its performance on both training and unseen data, as deeper networks can represent more complex functions.

(c) Adding more layers to the network can improve its ability to learn hierarchical representations of data, which is beneficial for complex pattern recognition tasks, though it may also increase the computational complexity and the risk of overfitting.

(d) Deeper networks are less likely to overfit compared to shallower ones, as they have a greater capacity to generalize from the training data to unseen data.

24. One Answer In the backpropagation algorithm, after computing the gradient of the loss function with respect to the weights, which step is crucial for updating the weights of a neural network to minimize the loss function?

(a) The weights are updated by setting them directly equal to the negative of the computed gradients to immediately minimize the loss.

(b) The learning rate is applied to the gradients, and this product is then subtracted from the current weights to gradually decrease the loss over iterations.

(c) Each weight's gradient is squared, and this squared gradient is then subtracted from the current weight value to ensure only positive updates.

(d) Gradients are normalized to unit length before being applied to update the weights, ensuring uniform step sizes across all dimensions.

25. One Answer A narrow and deep neural network will always outperform a wide, shallow neural network if the two networks have an approximately equivalent number of parameters.

(a) True

(b) False

26. One Answer During backpropagation, as the gradient flows backward through a sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, the gradient will always:

(a) **Decrease** in magnitude, **preserve** sign polarity

(b) **Increase** in magnitude, **preserve** sign polarity

(c) **Decrease** in magnitude, **reverse** sign polarity

(d) **Increase** in magnitude, **reverse** sign polarity

27. (2 points) Note: This question is significantly more time consuming than the others; you may want to finish other questions first.

A commonly used activation function in neural networks is the ReLU function, defined as

$$\mathsf{ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}.$$

Consider the following neural network with input $x \in \mathbb{R}^d$ and output $y \in \mathbb{R}$ defined as:

$$z = \mathsf{ReLU}(W^{(0)}x + b^{(0)})$$
$$y = W^{(1)}z + b^{(1)}$$

for parameters $W^{(0)} \in \mathbb{R}^{h \times d}, b^{(0)} \in \mathbb{R}^h, W^{(1)} \in \mathbb{R}^{1 \times h}, b^{(1)} \in \mathbb{R}$, and where ReLU is applied element-wise to the vector $W^{(0)}x + b^{(0)}$. Let $W_i^{(1)}$ denote the ith element of $W^{(1)}$, and ${W_i^{(0)}}^\top$ denote the ith row of $W^{(0)}$. For simplicity, assume that each element of $W^{(0)}x + b^{(0)}$ is non-zero.

What is $\frac{dy}{dW_i^{(0)}}$? You may write your answer in terms of $W_i^{(0)}$ and $W_i^{(1)}$.

Answer: _____

What is $\frac{dy}{dW^{(0)}}$? You may write your answer in terms of $\frac{dy}{dW_i^{(0)}}$.

Answer: _____

28. Select All Which of the following statements about PCA are True?

(a) PCA identifies the directions in feature space that minimize the variance of the projected data.

(b) PCA identifies the directions in feature space that minimize the reconstruction error between the original data and its projection onto the principal components.

(c) All principal component directions are orthogonal to one another.

(d) The first principal component direction is the eigenvector of the data covariance matrix that has the smallest eigenvalue.

(e) The principal component directions can be found from a singular value decomposition of the data matrix.

29. Select All Which of the following statements about PCA are True?

(a) For samples in d-dimensions, the top d principal components can fully reconstruct the original samples

(b) For samples in d-dimension, it's impossible to fully reconstruct the samples using top q principal components when $q < d$.

(c) Standard cross-validation techniques can be used to identify q, the optimal dimensionality of the PCA projection.

30. One Answer Throughout this course, we have seen that the solution to linear regression problems can be written as $\widehat{\theta} = (X^\top X)^{-1} X^\top Y$, for data matrices $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^n$. Let $X = U \Sigma V^\top$ denote the SVD of X. What is $\widehat{\theta}$ in terms of U, Σ, V, and Y? Note that this suggests that one way to solve a least squares problem is to simply compute the SVD of X.

(a) $V(\Sigma^\top \Sigma)^{-1} \Sigma^\top U^\top Y$

(b) $V \Sigma^{-1} U^\top Y$

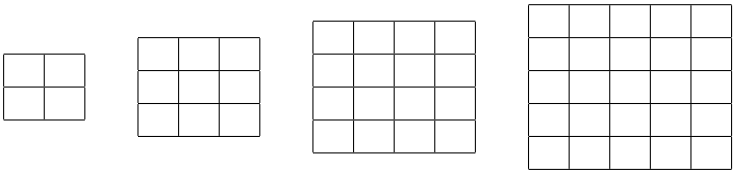
(c) $U \Sigma^\top (\Sigma^\top \Sigma)^{-1} V^\top Y$

(d) $U(\Sigma^\top \Sigma)^{-1} \Sigma^\top V^\top Y$

31. Consider the following matrix X and convolutional neural network (CNN) filter F.

$$X = \begin{array}{|c|c|c|c|}\hline 8 & 17 & 8 & 16 \\ \hline 13 & 7 & 10 & 5 \\ \hline 12 & 0 & 13 & 17 \\ \hline 7 & 11 & 11 & 9 \\ \hline \end{array} \qquad F = \begin{array}{|c|c|}\hline 1 & 0 \\ \hline 0 & 1 \\ \hline \end{array}$$

Apply the filter F to matrix X with Padding = 1 (padding with zeros) and stride = 2. Write the resulting matrix below in the grid of the correct size. Only write answers in one matrix, otherwise the problem will be graded as incorrect.



32. Give one main reason we might use a convolutional neural network over a fully connected one. Briefly explain why the CNN architecture makes that advantage possible.

Answer: _____

_____

_____

_____

33. Select All Which of the following functions are convex?

(a) $f(x) = x^\top Ax$, where A is a symmetric matrix and $x \in \mathbb{R}^d$.

(b) The logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$, where $x \in \mathbb{R}$.

(c) The L2 regularization term $f(x) = \lambda \|x\|_2^2$, where $\lambda > 0$ and $x \in \mathbb{R}^d$.

(d) The L1 regularization term $f(x) = \lambda \|x\|_1$, where $\lambda > 0$ and $x \in \mathbb{R}^d$.

(e) The regularization term $f(x) = \lambda \sum_{i=1}^{d} \sqrt{x_i}$, where $\lambda > 0, x_i > 0$.

34. Select All Which of the following statements are true for a convex function $f(x)$? You may assume that $f(x)$ is defined and is twice differentiable for all values of x.

(a) If you pick any two points on the graph of $f(x)$, the line segment connecting them will not lay underneath the graph.

(b) Every local minimum of $f(x)$ is also a global minimum.

(c) $f(x)$ must have at least one sharp corner or point, like the tip of a triangle.

(d) The second derivative of $f(x)$ is always negative.

(e) There is a unique value of x that minimizes $f(x)$.

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

(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

36. One Answer Suppose you train a polynomial regression model of degree $d = 4$ to approximate the linear function $g(x) = 3x + 2$. What is the most likely outcome?

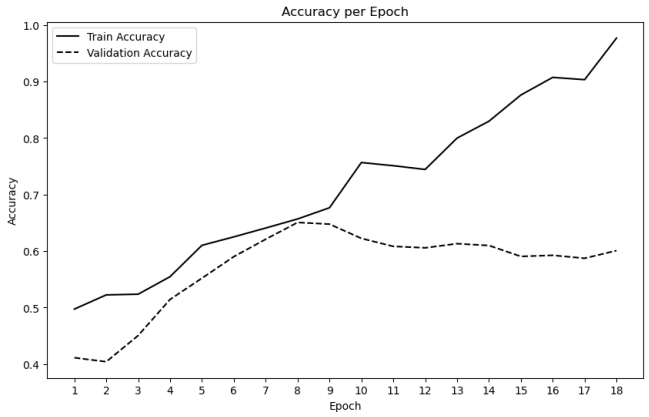
(a) The model will have high bias and high variance

(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

37. Consider the following accuracy plot generated while training a neural network.



Which **one** of these options would be most likely to improve the validation accuracy of your model?

1. Increase the learning rate
2. Decrease the learning rate
3. Increase regularization
4. Decrease regularization
5. Increase the number of epochs
6. Decrease the number of epochs
7. None of the above

Choose one option and explain how it would help improve the model (2 sentences). If none of these options are likely to change the model's performance, explain why (2 sentences).

Answer: _____

_____

_____

_____

38. Select All Which of the following statements about Maximum Likelihood Estimation (MLE) are true?

(a) MLE requires probabilistic assumptions to be made about the data.

(b) Linear regression can be viewed as a MLE problem.

(c) Logistic regression can be viewed as a MLE problem.

(d) MLE guarantees that the estimated parameters are unbiased.

(e) MLE is always equivalent to minimizing the squared reconstruction error.

39. In the context of machine learning, how does k-fold cross-validation enhance the reliability of a model's performance evaluation compared to using a single train-test split, and what is one potential downside of using this method?

Answer: _____

_____

_____

40. Select All Which of the statements about cross-validation are true?

(a) The goal of cross-validation is to estimate the training error.

(b) Leave-one-out cross-validation is equivalent to k-fold cross-validation when k is equal to the total number of training data points n.

(c) Leave-one-out cross-validation is always faster than k-fold cross-validation.

(d) k-fold cross-validation will always produce the same estimate of error, regardless of the choice of k.

41. Select All In which setting(s) might logistic regression be more suitable than k-nearest neighbors (KNN)?

(a) If your targets y_i take on continuous values.

(b) If you want to understand the relationship between your features x_i and your targets y_i.

(c) If you care most about minimizing the time required to train your model.

(d) If you care most about minimizing the time required to make predictions with your model.

42. Select All Ridge regression reduces overfitting by:

(a) Penalizing the L1 norm of the model parameters.

(b) Penalizing the L2 norm of the model parameters.

(c) Increasing model complexity to better fit the training data.

(d) Encouraging sparsity in the model parameters.

43. Select All Which of the following statements about ridge regression are **incorrect**?

(a) Ridge regression introduces a penalty term to the linear regression cost function by controlling the magnitudes of the coefficient values.

(b) The regularization term in ridge regression helps in reducing model complexity by setting some coefficients to exactly zero.

(c) The ridge regression parameter estimate will be equivalent to the Ordinary Least Squares parameter estimate when $\lambda = 0$ in the case that $X^T X$ is full rank (where X is the training data matrix).

(d) The choice of the regularization hyperparameter (λ) in ridge regression can significantly impact the model's bias-variance trade-off.