CSE 446/546 Spring 2023 Final Exam

June 7, 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 NetID on your notes page (if you have one).

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 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.

1. If we have n data points and d features, we store nd values in total. We can use principal component analysis to store an approximate version of this dataset in fewer values overall. If we use the first q principal components of this data, how many values do we need to approximate the original demeaned dataset? Justify your answer.

   Answer: _____

   _____

2. Suppose we have a multilayer perceptron (MLP) model with 17 neurons in the input layer, 25 neurons in the hidden layer and 10 neuron in the output layer. What is the size of the weight matrix between the hidden layer and output layer?

   (a) 25×17

   (b) 10×25

   (c) 25×10

   (d) 17×10

3. Recall that a kernel function $K(x, x')$ is a metric of the similarity between two input feature vectors x and x'. In order to be a valid kernel function, $K(x, x') = \phi(x)^T \phi(x')$ for some arbitrary feature mapping function $\phi(x)$. Which of the following is **not** a valid kernel function for input features $x, x' \in \mathbb{R}^2$?

   (a) $(x^T x')^2$

   (b) $3x^T x'$

   (c) $x^T x'$

   (d) All of the above are valid

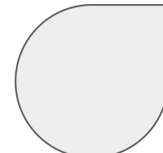4. Consider the following figure. Which shape is not convex?
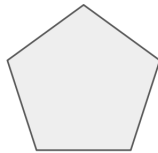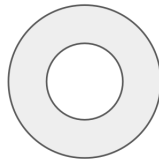
I.   II.   III.   IV.

(a) I.

(b) II.

(c) III.

(d) IV.

5. What is the typical effect of increasing the penalty (λ) in the ridge regression loss function? Select all that apply.

(a) It increases the bias of the model.

(b) It decreases the bias of the model.

(c) It increases the variance of the model.

(d) It decreases the variance of the model.

6. Suppose we are performing linear regression using a non-linear basis expansion Φ. Which of the following statements is true about the learned predictor?

(a) It is a linear function of the inputs and a linear function of the weights.

(b) It is a linear function of the inputs and a non-linear function of the weights.

(c) It is a non-linear function of the inputs and a linear function of the weights.

(d) It is a non-linear function of the inputs and a non-linear function of the weights.

7. Which of the following is true about the k-nearest neighbors (KNN) algorithm?

(a) It is a parametric model.

(b) It learns a nonlinear decision boundary between classes.

(c) It requires a separate training phase and testing phase for prediction.

(d) It typically requires longer training compared to other ML algorithms.

8. Which of the following statements about the first principal component is true?

(a) If we add Gaussian noise to a feature in the input matrix X, the first principal component remains unchanged.

(b) The first principal component is equivalent to the eigenvector corresponding to the largest eigenvalue of the input matrix X.

(c) The first principal component is the vector direction which maximizes the variance of the input.

(d) The first principal component corresponds to the most influential feature for prediction.

9. Leave-one-out cross-validation (LOOCV) is a special case of k-fold cross-validation where:

   (a) The training set contains all but one sample, and the remaining sample is used for testing.

   (b) The training set contains only one sample, and the remaining sample is used for testing.

   (c) The training set contains exactly one sample from each class, and the remaining samples are used for testing.

   (d) The training set contains one sample from each fold, and the remaining folds are used for testing.

10. Which of the following statements accurately compare or contrast bootstrapping and cross-validation? Select all that apply.

    (a) Bootstrapping and cross-validation both train models on subsets of the training data.

    (b) In cross-validation, there is no overlap between the subsets each model trains on.

    (c) Bootstrapping and cross-validation are both methods to estimate prediction error.

    (d) In bootstrapping, each model is trained on the same number of data points as the original training set, unlike cross-validation.

    (e) In cross-validation, each learned model is evaluated on non-overlapping subsections of the original training set, unlike bootstrapping.

11. Which of the following are true about a twice-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$? Select all that apply.

    (a) f is convex if $f(\lambda x + \lambda y) \leq \lambda f(x) + \lambda f(y)$ for all x, y in the domain of f and $\lambda \in [0, 1]$.

    (b) f is convex if $\nabla^2 f(x) \succeq 0$ for all x in the domain of f.

    (c) f is convex if the set $\{(x, t) \in \mathbb{R}^{d+1} : f(x) \leq t\}$ is convex.

12. Which of the following statements explain why random forests are preferable to individual decision trees? Select all that apply.

(a) Random forests reduce overfitting by aggregating predictions from multiple trees.

(b) Random forests reduce overfitting by having each tree in the forest use a subset of all the data features.

(c) Random forests can handle a larger number of features compared to individual decision trees.

(d) Random forests provide better interpretability and understanding of the underlying relationships in the data than individual decision trees.

13. Let P_{XY} represent the distribution of (x_i, y_i) samples in our training dataset, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. Suppose P_X is supported everywhere in $\mathbb{R}^d$ and $P(Y = 1 \mid X = x)$ is smooth everywhere. Which of the following statements is true about 1-nearest neighbor classification as the number of training samples $n \to \infty$?

(a) The error of 1-NN classification approaches infinity.

(b) The error of 1-NN classification is at most twice the Bayes error rate.

(c) The error of 1-NN classification is at most the Bayes error rate.

(d) The error of 1-NN classification approaches zero.

14. Which of the following is true about Pooling layers in convolutional neural networks (CNNs)? Select all that apply.

(a) A 2×2 pooling layer has 4 parameters.

(b) Pooling layers never change the height and width of the output image.

(c) For a max-pooling layer, the gradients w.r.t. some inputs will always be zero.

(d) Pooling layers do not change the depth of the output image.

15. Which of the following statements about SVMs are true? Select all that apply.

(a) SVMs are only applicable to binary classification problems.

(b) SVMs cannot be applied to non-linearly separable data.

(c) SVMs are a form of supervised learning.

(d) SVMs are primarily used for regression tasks.

16. In Gaussian mixture models (GMMs), which of the following statements is **false**?

(a) GMMs assume that the data points within each component follow a Gaussian distribution.

(b) GMMs can be used for clustering.

(c) The number of components in a GMM must be equal to the number of features in the dataset.

17. True/False: If X is a matrix in $\mathbb{R}^{n \times m}$, $X^\top X$ is always invertible.

(a) True

(b) False

18. Note: This question was thrown out during the exam in Spring 2023. Consider the dataset pictured below. The features of each datapoint are given by its position. So the datapoint (0,1) appears at position (0,1). The ground truth label of the datapoint is given by its shape, either a circle or square. You have a test set of datapoints, shown with no fill, and a train set of data, shown with a grey fill.
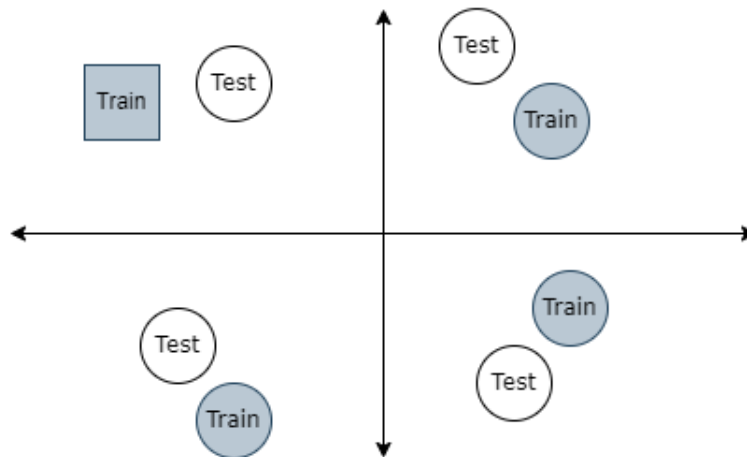
True/False: KNN with $K = 1$ has higher train accuracy than with $K = 4$.

(a) True

(b) False

19. True/False: Consider the dataset from the previous problem. KNN with $K = 1$ has higher test accuracy than with $K = 4$.

(a) True

(b) False

20. Suppose you are training neural networks on 100×100 images to predict 5 classes. Neural network A consists of a single linear layer followed by a softmax output activation. Neural network B consists of a sequence of layers with dimensions 128, 512, and 32, respectively, followed by a softmax output activation. Assuming that both neural networks are trained using an identical procedure (e.g. batch size, learning rate, epochs, etc), and neither contains hidden activations, what can you generally expect about the relative performance of A and B on the test data?

   (a) A will outperform B

   (b) B will outperform A

   (c) A and B will perform roughly the same.

21. Recall that the probability of seeing data $\mathcal{D}$ from a Gaussian distribution is $\mathbb{P}(\mathcal{D}|\mu, \sigma) = (\frac{1}{\sigma\sqrt{2\pi}})^n \prod_{i=1}^{n} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$. Consider the Maximum Likelihood Estimators $\hat{\mu}_{MLE}$ and $\hat{\sigma^2}_{MLE}$ from this distribution. Which of the following are true? Select all that apply.

   (a) $\hat{\mu}_{MLE}$ is dependent on $\hat{\sigma^2}_{MLE}$

   (b) $\hat{\sigma^2}_{MLE}$ is dependent on $\hat{\mu}_{MLE}$

   (c) $\hat{\mu}_{MLE}$ is a biased estimator

   (d) $\hat{\sigma^2}_{MLE}$ is a biased estimator

22. True/False: The bootstrap method samples a dataset with replacement.

   (a) True

   (b) False

23. What does the PyTorch optimizer's `step()` function do when training neural networks?

    (a) Adjust the network's weights based on the gradients

    (b) Randomly initializing the network's weights.

    (c) Sets all the network's gradients to zero to prepare it for backpropagation

    (d) Compute the gradients of the network based on the error between predicted and actual outputs.

24. Below is a simple computation graph with inputs x and y with an initial computation of $z = xy$ before the unknown path to final loss L. A forward propagation pass has been completed with values $x = 3$ and $y = 4$, and the upstream gradient is given as $\partial L/\partial z = 5$. Complete the backpropagation pass by **filling in** the scalar answers to boxes $\partial L/\partial x$ and $\partial L/\partial y$.
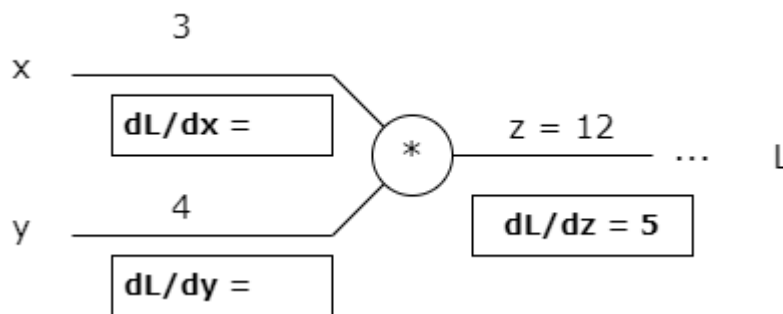
25. What are two possible ways you could reduce overfitting in a neural network?

Answer: _____

_____

_____

26. True/False: Suppose you set up and train a neural network on a classification task and converge to a final loss value. Keeping everything in the training process the exact same (e.g. learning rate, optimizer, epochs). It is possible to reach a lower loss value by ONLY changing the network initialization.

(a) True

(b) False

27. Why should we not use ridge regression to select features by setting a threshold on coefficient magnitude and only consider features with coefficient magnitude larger than the threshold as important features?

Answer: _____

_____

_____

28. 4 Deep Neural Network models are trained on a classification task, and below are the plots of their losses:
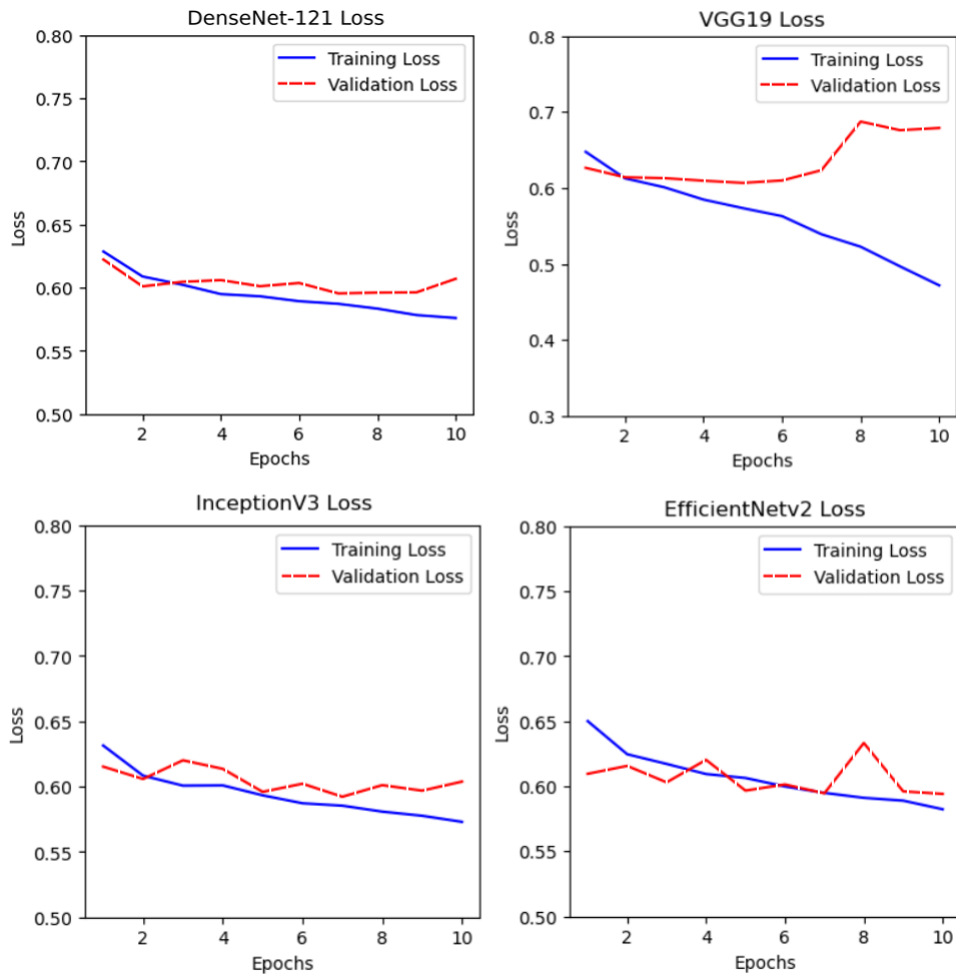
Based on these plots, which model is overfitting?

(a) DenseNet-121

(b) VGG19

(c) InceptionV3

(d) EfficientNetv2

29. True/False: For k-means clustering, the number of clusters k should be that which minimizes the loss function.

(a) True

(b) False

30. We have a convolutional neural network that takes in input of images with dimensions $(3, 128, 128)$. The first convolutional layer (depicted below) has 32 filters each of size $(3, 3, 3)$ and uses a stride of 2 and padding of 1. One of these filters is shown in the diagram at a specific region of the input as well as the corresponding region of the output. After applying this convolutional layer, what must be the value of a, b, and c? *Note that the diagram is not drawn to scale.*
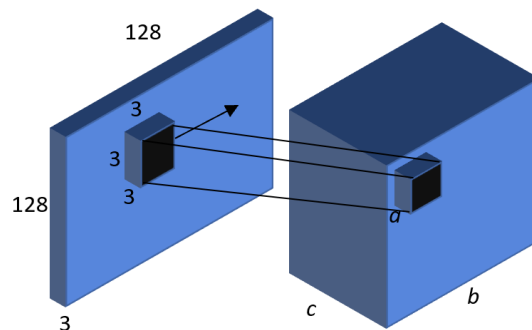
(a) $a = 1$, $b = 64$, $c = 32$

(b) $a = 3$, $b = 64$, $c = 32$

(c) $a = 1$, $b = 32$, $c = 64$

(d) $a = 3$, $b = 32$, $c = 64$

31. In which of the following situations can logistic regression be used? Select all that apply.

    (a) Predicting whether an email is a spam email or not based on its contents.

    (b) Predicting the rainfall depth for a given day in a certain city based on the city's
    historical weather data.

    (c) Predicting the cost of a house based on features of the house.

    (d) Predicting if a patient has a disease or not based on the patient's symptoms and
    medical history.

32. We train a model on some data using LASSO regression. Which of the following
    solutions offers the lowest bias and why?

    (a) The weights $\widehat{w}$ after running the LASSO regression; because non-smooth loss func-
    tions tend to produce lower bias

    (b) The weights $\widehat{w}$ after running the LASSO regression; because sparse solutions have
    lower bias

    (c) The weights $\widehat{w}$ after running unregularized regression again on just the features
    with nonzero weights in the output of LASSO regression; because running multiple
    models tends to reduce bias

    (d) The weights $\widehat{w}$ after running unregularized regression again on just the features
    with nonzero weights in the output of LASSO regression; because regularization
    introduces some bias into the model

33. Which of the following can result from choosing a smaller value of k in k-nearest
    neighbors (KNN)? Select all that apply.

    (a) Increased underfitting.

    (b) Increased overfitting.

    (c) No impact on model fit.

34. True/False: Suppose we are doing ordinary least-squares linear regression with a bias term. Projecting the sample points onto a lower-dimensional subspace with Principal Component Analysis (PCA) and performing regression on the projected points can make the training data loss smaller.

    (a) True

    (b) False

35. What is the purpose of the sigmoid function in logistic regression?

    (a) It converts continuous input into categorical data.

    (b) It standardizes the input to have zero mean and variance 1.

    (c) It optimizes the weights to reduce loss.

    (d) It transforms the output to a probability.

36. Consider the general least squares regression objective function $L(w) = \|Xw - Y\|_2^2 + \lambda w^\top D w$ whose gradient is $\nabla_w L(w) = 2(X^\top X + \lambda D)w - 2X^\top y$. Which conditions must be true for there to be a unique solution to minimizing the objective function? Select all that apply.

    (a) $X^\top X + \lambda D$ must be invertible.

    (b) $X^\top X + \lambda D$ must have a non-trivial nullspace.

    (c) $X^\top X + \lambda D$ must have full rank.

37. Which of the following are true about bagging and boosting? Select all that apply.

    (a) Bagging is a technique that predicts the average of the predictions outputted by independently trained models.

    (b) Random forests are an example of bagging.

    (c) Boosting is a technique that iteratively learns new models that correct for the error produced by previous models it learned.

    (d) Boosting weights the predictions of the different models it learns equally when computing the final prediction.

38. Suppose that after solving a soft margin SVM problem we obtain that the best separating hyperplane is $w^T x + b = 0$ for $w = [1, -2]$ and $b = 3$. Consider the following points $x_1 = [2, 1]$, $x_2 = [-0.5, 1.5]$, $x_3 = [-1.75, 0.5]$. What are the labels ($+1$ or -1) assigned by our model to the three points?

    Answer: $y_1 =$ _____, $y_2 =$ _____, $y_3 =$ _____

39. Recall the RBF kernel: $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$ where $\gamma = \frac{1}{2\sigma^2}$.

    Which of the following statements about the RBF kernel is true?

    (a) The RBF kernel is only applicable to binary classification.

    (b) Increasing the γ hyperparameter reduces overfitting.

    (c) The RBF kernel is positive semi-definite (where $\gamma > 0$).

    (d) The RBF kernel is invariant to feature scaling.

40. For a linear regression with bias term, we want to find $\hat{w}_{LS}, \hat{b}_{LS}$ such that $\hat{w}_{LS}, \hat{b}_{LS} = \text{argmin}_{w,b} \|y - (\mathbf{X}w + \mathbf{1}b)\|_2^2$ for $\mathbf{X} \in \mathbb{R}^{n \times d}, w \in \mathbb{R}^d, y \in \mathbb{R}^n$. $\mathbf{1}$ indicates a vector of all ones.

    If $\mathbf{X}^T \mathbf{1} = 0$, what is $\hat{b}_{LS}$?

    Answer: $\hat{b}_{LS} =$_____

41. What is a commonly used optimization algorithm when training neural networks?

    Answer: _____

42. In a linear regression model with normally distributed errors, which of the following is the likelihood function?

(a) $L(y|X, w) = \sum_{i=1}^{n}(y_i - X_iw)^2$

(b) $L(y|X, w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - X_iw)^2}{2\sigma^2}\right)$

(c) $L(y|X, w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - X_iw)^2}{2}\right)$

(d) $L(y|X, w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - X_iw)}{2}\right)$

43. What is the role of nonlinear activation functions in neural networks? Briefly describe in 1-2 sentences **and** provide an example of an activation function that adds nonlinearity.

Answer: Activation Function: _____

Answer: Role: _____

_____

44. What is the key reason why backpropagation is so important?

(a) Backpropogation allows us to compute the gradient of any differentiable function.

(b) Backpropogation is the only algorithm that enables us to upate the weights of a Neural Network.

(c) Backpropogation is an efficient dynamic program that enables us to compute the gradient of a function at the same time-complexity it takes to compute the function.

(d) Backpropagation introduced Chain Rule into the world of mathematics, enabling significant advances in the field.

45. What is a commonly used loss function when training a neural network for a multi-class classification problem?

Answer: _____