Exam 2

Time limit. 90 minutes.

Collaboration and materials. You must complete this exam individually. You may not use any materials (physical or electronic) besides both sides of five sheets of 8.5x11-inch paper with 1-inch margins and the equivalent of 10-point font.

Questions. This exam has twenty multiple-choice questions. Some questions require you to select exactly one of the answer choices, while others require you to select all of the answer choices that apply. Questions of the latter kind always end with "Select all that apply."

Scoring. Each question is weighted equally. For questions requiring you to select one of the answer choices, no partial credit will be awarded. For questions requiring you to select all of the answer choices that apply, partial credit will be awarded for each correct answer selected while no points will be awarded if no correct answers are chosen or if any incorrect answers are selected.

Submission. You will receive a bubble sheet for your answers. Please print your full name as it appears on Gradescope (please no cursive), your student ID, and today's date (December 7). You may leave the "Section" box blank. **Your version is A. Please check that this matches the pre-bubbled version number at the top of the bubble sheet.** For each question, please fill in the appropriate bubbles completely using either pencil or blue/black pen. If you have filled in a bubble with pen but have changed your mind, you can cross out that bubble with an X. Note that the answer choices are presented in the order A, B, C, D, E.

1 1 point

Suppose we have data (X_i, Y_i), where $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. We fit a degree-$d$ polynomial regression model to the data. Which values for d might give the lowest expected test error? [You may use the fact that any polynomial regression model with $d \geq 2$ has no bias.]

○ $d = 2$ must give the lowest expected test error.

○ Some value $d \leq 2$ must give the lowest expected test error.

○ Some value $d \geq 2$ must give the lowest expected test error.

○ Any value of $d \geq 0$ might give the lowest expected test error.

○ Some value d among $\{1, 2, 3\}$ must give the lowest expected test error.

2 1 point

What is the minimum number of terminal nodes in a tree with interaction depth 3?

○ 2

○ 3

○ 4

○ 5

○ 6

3 1 point

Which of the following statements about convexity are correct? Select all that apply.

□ Every convex function has at least one local minimum.

□ A function where every local minimum is a global minimum is convex.

□ For a convex function, every local minimum is a global minimum.

□ Gradient descent on a convex function is guaranteed to converge to a global minimum, regardless of the starting point and learning rate.

□ A convex function cannot have any sharp corners.

4 1 point

I trained a predictive model to classify which animal is depicted in a grayscale image. Unfortunately, the images I used for training and testing had their pixels shuffled, compared to the original images. In particular, the pixel at location (i, j) in each of my images came from location $\pi(i, j)$ in the corresponding original image, for some function π and for each location i, j. In other words, the pixels were shuffled in the same way for each image.

For which of the following predictive models would this shuffling of pixels result in a substantially different classification accuracy than if the same model were applied to the non-shuffled images? Select all that apply.

□ Multi-class logistic regression

□ One-hidden-layer fully connected neural network

□ Convolutional neural network

□ Logistic lasso, where each pixel value is a feature

□ Random forest, where each pixel value is a feature

5 1 point

Which of the following increases the model complexity of a multi-layer neural network trained for a fixed number of SGD steps, when other factors are held equal? Select all that apply.

□ Adding dropout

□ Adding weight decay

□ Adding another hidden layer to the network

□ Adding more SGD steps

□ Adding more training data

6 1 point

Which of the following learning tasks require human-labeled data? (Human-labeled data is data (X_i, Y_i) where humans must manually determine what is Y_i for each X_i.)

□ Image colorization

□ Machine translation

□ Language modeling

□ Sentiment analysis

□ Image classification

7 1 point

Which of the following statements is true? Select all that apply.

□ The size of the input images does not impact the number of parameters in a CNN.

□ The lengths of the input sentences does not impact the number of parameters in a basic RNN for sentiment analysis.

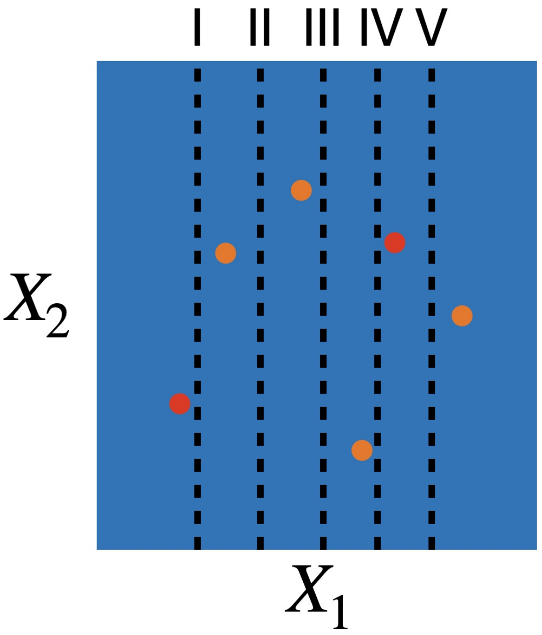
□ The number of features does not impact the number of parameters in a linear regression model.

□ The number of features does not impact the number of parameters in a multi-layer neural network.

□ The lengths of the input sentences does not impact the number of parameters in an RNN with attention for sentiment analysis.

8 1 point

Consider a classification tree based on the training data shown below.



Which of the five initial splits will lead to the lowest misclassification error? Assume that ties are broken in favor of the minority class. Select multiple choices if the lowest misclassification error is achieved by more than one of the splits under consideration.

□ I

□ II

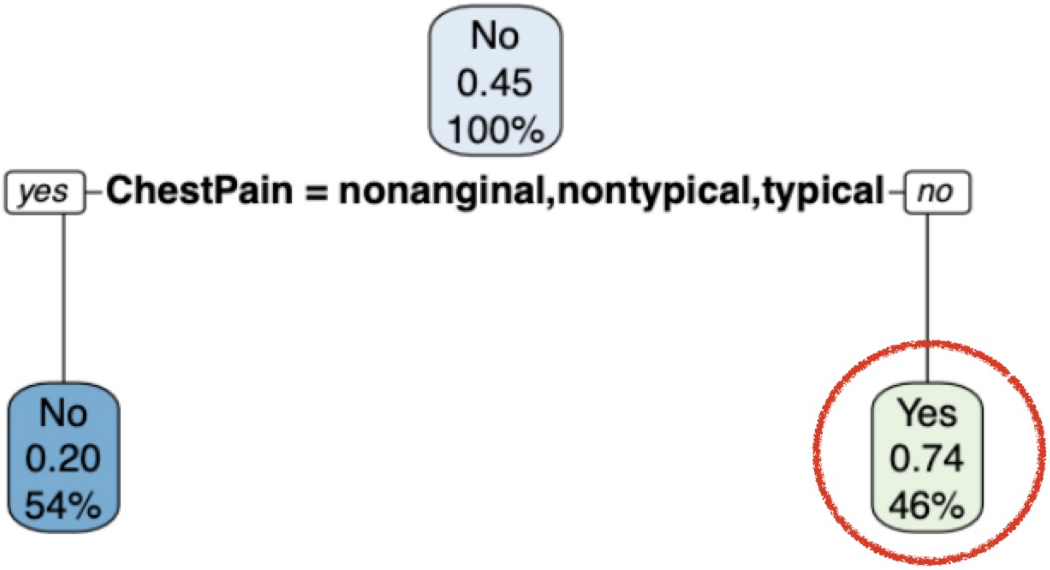
□ III

□ IV

□ V

9 1 point

Consider the following stump tree.



The misclassification error among the training observations falling in the circled node is X%, where X is an integer. What is the sum of the digits of X?

○ 8

○ 9

○ 10

○ 11

○ 12

10 1 point

Suppose we build a random forest using n = 5 training observations, B = 4 trees, and p = 10 features. The following are the four bootstrap samples:

- Bootstrap sample 1: Observations 2, 4, 1, 3, 2
- Bootstrap sample 2: Observations 3, 5, 4, 1, 5
- Bootstrap sample 3: Observations 1, 4, 4, 5, 1
- Bootstrap sample 4: Observations 3, 3, 5, 4, 2

Over the course of the calculation of the out-of-bag error, how many times was a fitted tree evaluated on a data point?

○ 1

○ 2

○ 3

○ 4

○ 5

11 1 point

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

[Here, "Parallelized across trees" means that computations for each tree can be performed without knowing the results of computations for other trees.]

□ The training of a boosted tree model can be parallelized across trees.

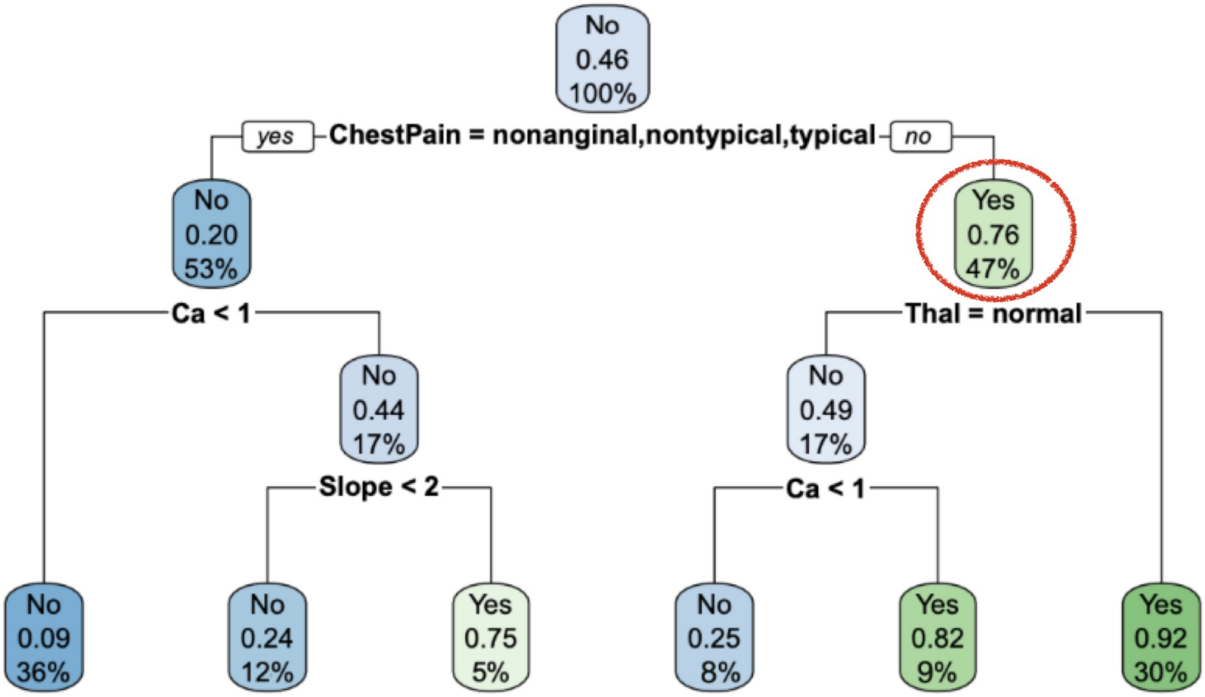
□ The testing of a boosted tree model can be parallelized across trees.

□ The training of a random forest model can be parallelized across trees.

□ The testing of a random forest model can be parallelized across trees.

12 1 point

The fraction of training observations in the circled node below have Thal = normal can be expressed as a simplified fraction A/B. To which of the intervals below does A + B belong?



○ [61, 70]

○ [71, 80]

○ [81, 90]

○ [91, 100]

○ [101, 110]

13 1 point

Consider a version of a random forest where all trees are grown on the original sample, as opposed to on bootstrapped samples. Which of the following statements are true? Select all that apply.

□ If m is tuned appropriately, this version of a random forest may still outperform a single decision tree.

□ This version of a random forest is the same as a single decision tree.

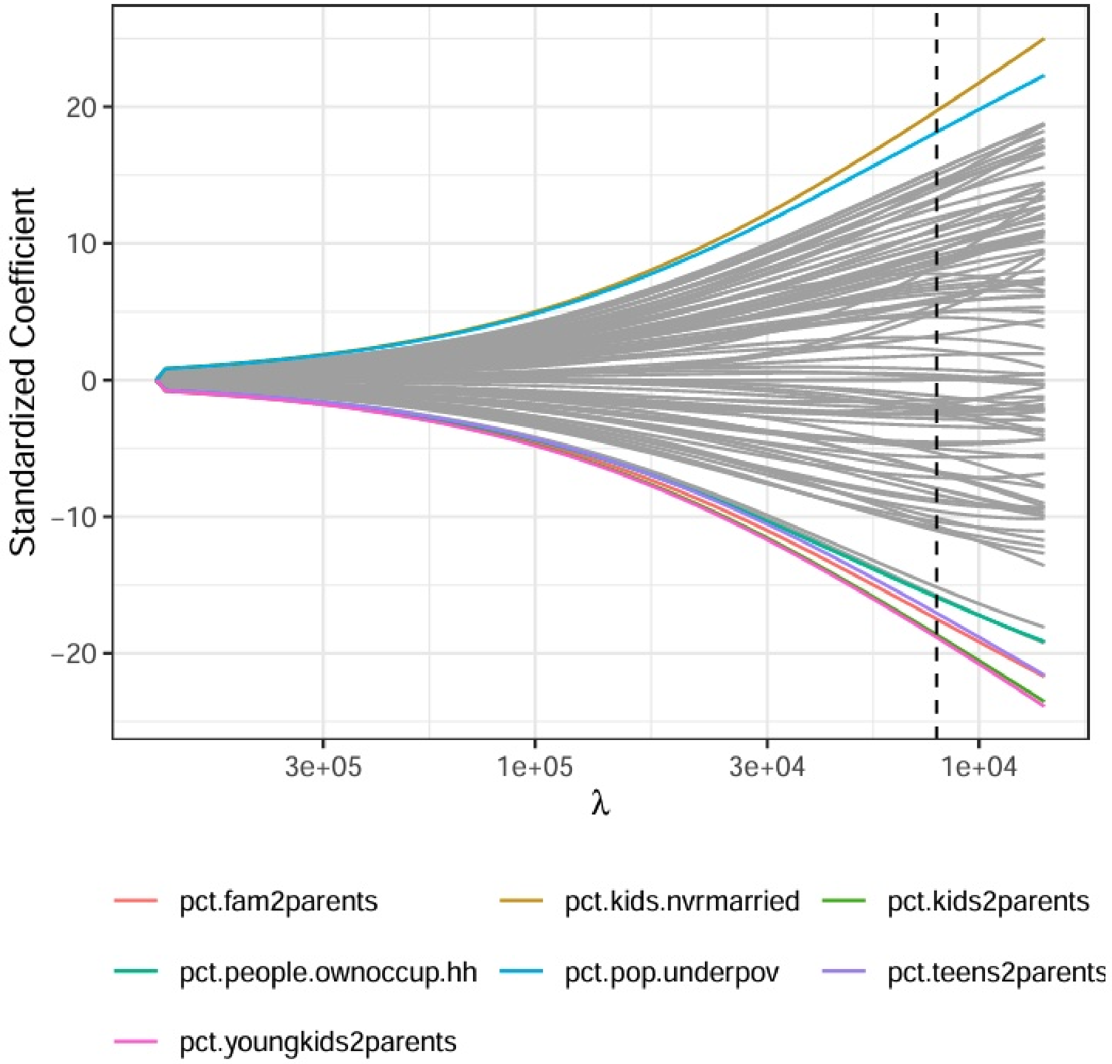
□ For a fixed value of m, this version of a random forest has lower mean variance (i.e. less variable predictions) than the usual random forest.

□ If m is set to p, this version of a random forest may still outperform a single decision tree.

□ OOB error would no longer be applicable to this version of a random forest.

14 1 point

You ran a ridge regression of the violent crime rate on a number of socioeconomic variables across many U.S. cities, and chose your penalty parameter based on the dashed vertical line in the ridge trace plot below. There are two cities (called A and B) in your test data that have the same features, except city B has `pct.kids.nvrmarried = 40` while city A has `pct.kids.nvrmarried = 30`. By how much does your fitted model predict the violent crime rate in city B exceeds that of city A? Note that the mean and standard deviation of `pct.kids.nvrmarried` are 30 and 5, respectively.



○ 5

○ 10

○ 20

○ 30

○ 40

15 1 point

We fit an intercept-only model to $n = 50$ training data points (X_i, Y_i) drawn from $Y_i = \sin(3 \cdot X_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ for $\sigma^2 = 1$. Doubling σ and n will have what effect on the mean variance of the fit?

○ The mean variance will increase by a factor of four.

○ The mean variance will increase by a factor of two.

○ The mean variance will stay the same.

○ The mean variance will decrease by a factor of two.

○ The mean variance will decrease by a factor of four.

16 1 point

You fit the following logistic regression model of default on balance (the latter measured in dollars):

$$\mathbb{P}[\text{default} \mid \text{balance}] = \text{logistic}(2 - 0.1 \cdot \text{balance})$$

As balance increases by $100, the odds of default is multiplied by what factor?

○ $e^{0.1}$

○ e^{-8}

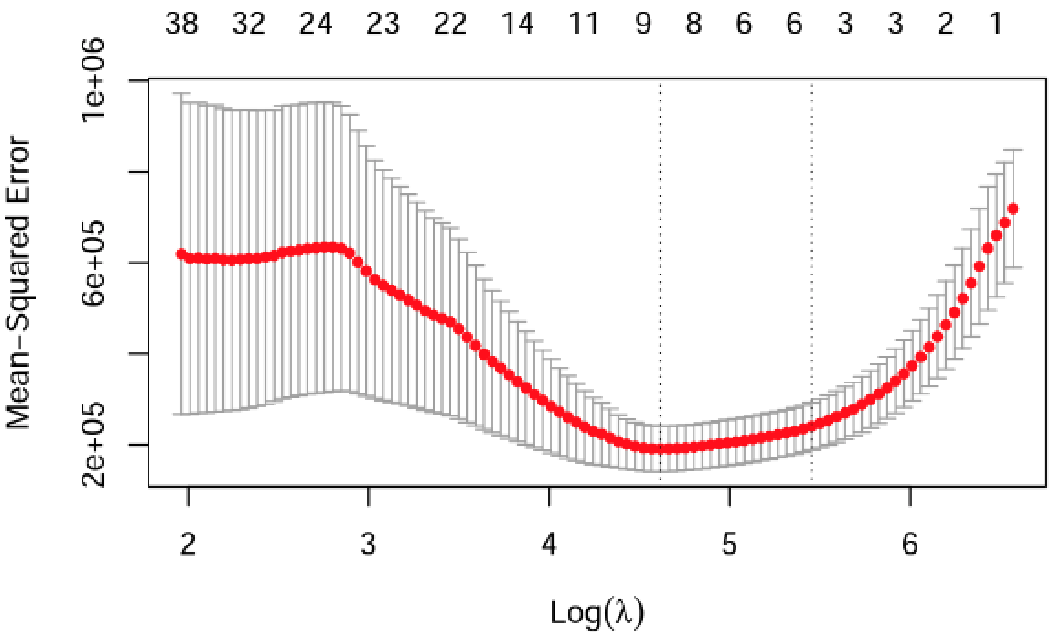
○ e^{-10}

○ e^{10}

○ $e^{-0.1}$

17 1 point

Consider the following lasso CV plot:



Which of the following values of $\text{Log}(\lambda)$ gives a CV error that is within one standard error of the minimum CV error? If there are multiple such values, choose the one that gives the sparsest model.

○ 2

○ 3

○ 4

○ 5

○ 6

18 1 point

How many columns will the tibble resulting from the operation below have?

```
diamonds |> summarize(max_price= max(price),  .by= c(cut,clarity))
```

○ 1

○ 2

○ 3

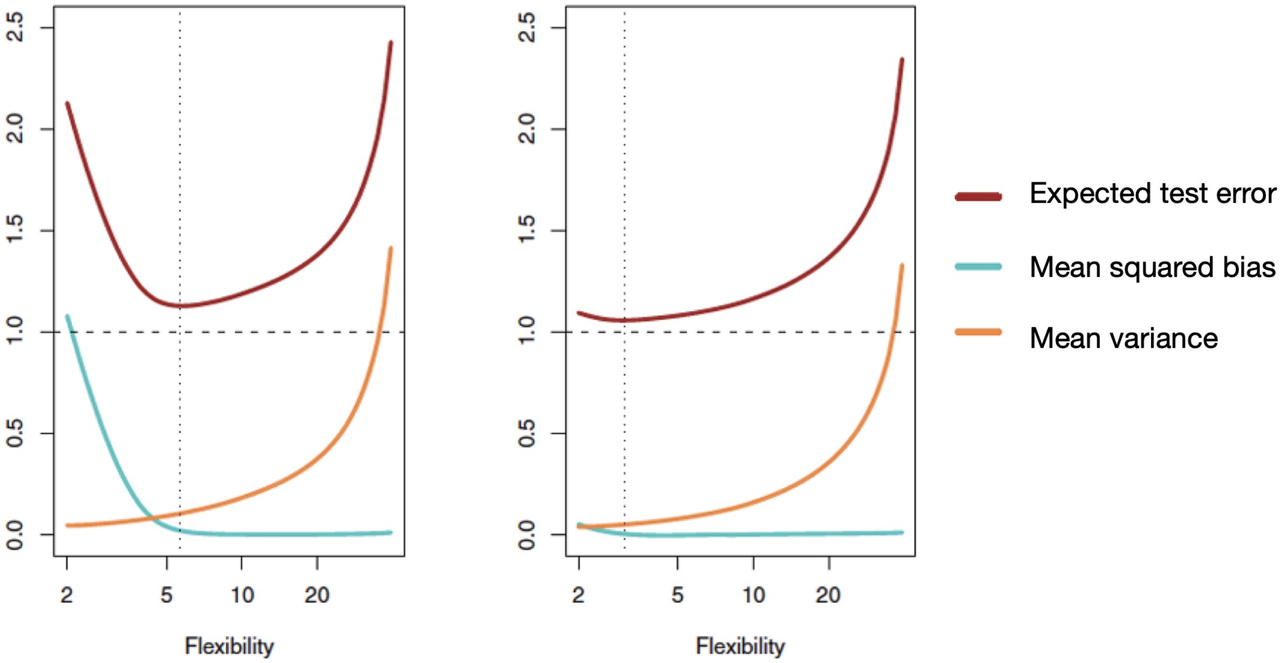
○ 4

○ Not enough information given

19 1 point

Shown below are the ETE, mean squared bias, and mean variance of a predictive model fit to n data points from the distribution $Y = f(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. The only potential differences between the plots are the training sample size n, the noise level σ, and the true trend f

.



Compared to the scenario on the left, the scenario on the right might have which of the following properties? Select all that apply.

[You may assume that the mean variance curves match exactly and that the right-hand edges of the ETE and mean squared bias match exactly between the two plots.]

□ A smaller sample size n

□ A larger sample size n

□ A less complex true trend f

□ A more complex true trend f

□ A smaller noise level σ

20 1 point

What properties do RNNs with attention and transformers share?

□ Attention

□ Recurrence

□ Ability to handle variable-length inputs and outputs

□ Ability to autoregressively generate text

□ Highly parallelizable training