Exam 2 A - MATH 4322

Cathy Poliak

Spring 2022

Name: _____ **PSID**: _____

Instructions

- Allow one sheet of notes front and back to be turned in for extra credit.
- Allow calculator.
- Total possible points 100.
- For multiple choice circle your answer on this test paper.
- For short answer questions answer fully on this test paper, partial credit will be given.
- Once completed leave at the desk, I will pick up your test.

Part 1

We are trying to find a number (mpg) so it is quantitative, and therefore regression

We want to be able to predict the miles per gallon (mpg) of an automobile based on certain variables.

1. (3 Possible Points) Is this a regression or classification problem? Give the reason for your answer.

This is a regression problem, since the mpg is a quantitative variable.

2. (8 Possible Points) The following is a decision tree resulting in fitting the response mpg to some predictors.

   a. List the predictors used in this tree.

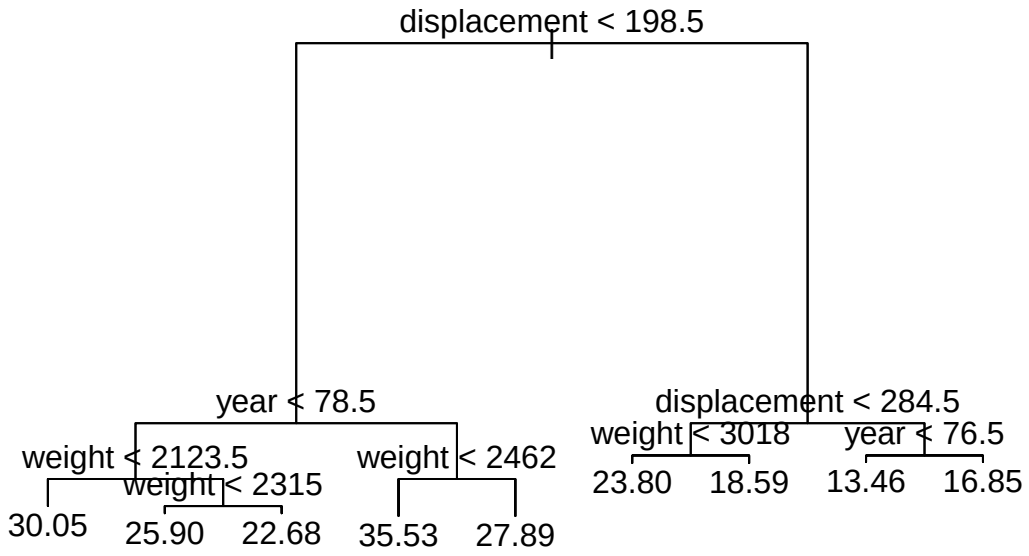
If we look at the tree below, we see only 3 predictors used: displacement, year, and weight.

displacement, year, and weight

b. Interpret what the end notes 30.05 means. Explain fully who we get this value based on the tree.

If the displacemtn is less than 198.5, and year is less than 78.5, and weight is less than 2123.5, then the average predicted mpg is 30.05

If a value is < the value listed on the tree branch, then it goes left otherwise it goes right. To reach 30.5, we need to go left 3 times so it needs to be less than all 3 checks.



3. (10 Possible Points) The following are the mean square errors based on the single tree, random forest and bagging.

   a. Give the formula for the mean squared error (MSE).

just know this lol

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2$$

b. Match the correct MSE with: a. Single tree, b. Random Forest, or c. Bagging.
   i. 7.7099
   ii. 8.0475
   iii. 11.4043

Random Forest is most accurate, so smallest error
Bagging is between RF and Single Tree
Single Tree is least accurate, so biggest error

i is Random Forest, ii. is Bagging and iii. is Single Tree

c. Interpret the MSE value of 11.4043?

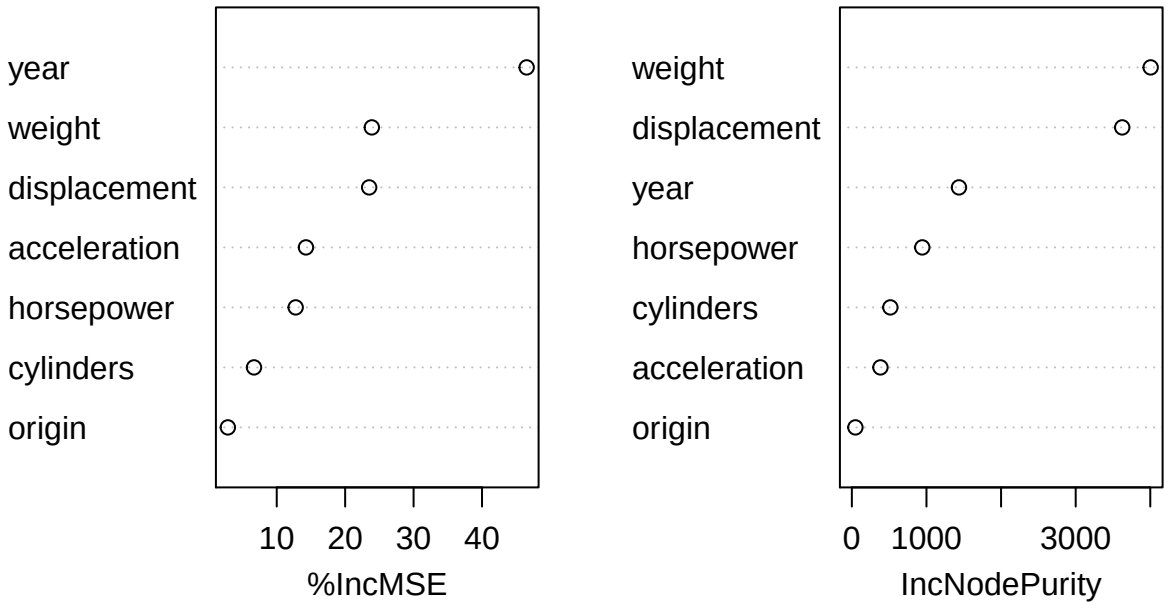
if we look at the formula (yhat - y) is squared, so sqrt to get 3.377

The is the squared ammount that we may be off by. For this example we on average will be off by 3.377.

4. (3 Possible Points) The variable of importance plot is below from the bagging method. What are the three most important variables? Compare that to the single tree in problem 1.

The three most important are year, weight and displacement. This is the three variables used in the single tree.

for the plot below, look at the 3 biggest values (from left to right, smallest to biggest)
year, weight, and displacement are the 3 biggest values (right-most) so they are
most important



Part 2

We want to know if customer will default on their credit card debt. The variables are:

- default - A factor with levels `No` and `Yes` indicating whether the customer defaulted on their debt. This is our output (response variable).
- student - A factor with levels `No` and `Yes` indicating whether the customer is a student.
- balance - The average balance that the customer has remaining on their credit card after making their monthly payment.
- income - Income of the customer.

1. (3 Possible Points) Is this a regression or classification problem? Give the reason for your answer.

This is a classification problem, since the variable default is a categorical variable

Two categories:
either they default or don't
categorical, classification

2. (10 Possible Points) The following is an excerpt of the single tree.

   a. In node 7), how many observations are in this node?

102

Nodes are formatted like so: node), split, n, deviance, yval, (yproblow yprobhigh) with a * if its terminal
we need to look at the n in this case, which for node 7 is 102, therefore our observation is 102

b. In node 7), what is the range of the 'balance', that is what are the possible values of the varaibl

The values of balance are greater than 1874.98

The balance is the split, defined in the formatting from the previous question. For node 7, its balance > 1874.98, so thats our answer

c. In node 7), what percent did default on their credit card?

70.5%

if we look at the formatting defined two questions ago, we look at yprobhigh
why yprobhigh? because we are trying to find who defaulted (yes category)
if it was no category we would look at yproblow

d. Is node 14) a terminal node?

Yes

Yeah because we see a *, and as the format dictates 3 questions ago, that
indicates its a terminal node

e. What is the overall prediction for the default ('No' or 'Yes') for node 14)?

Yes, they are predicted to default on the credit card.

```
 7) balance > 1874.98 102  123.60 Yes ( 0.294118 0.705882 )
  14) balance < 2152.31 77  103.00 Yes ( 0.389610 0.610390 ) *
  15) balance > 2152.31 25    0.00 Yes ( 0.000000 1.000000 ) *
```

3. (3 Possible Points) Below is the confusion matrix for the tree. What is the test error rate?

		Observed Class	
		No	Yes
Predicted Class	No	4804	103
	Yes	38	55

We will only be asked to find test error rate on this exam, which is false positive + false negative over the total of all values in the matrix
False positive and false negatives are the "mismatched diagonals" where the no and yes don't match up

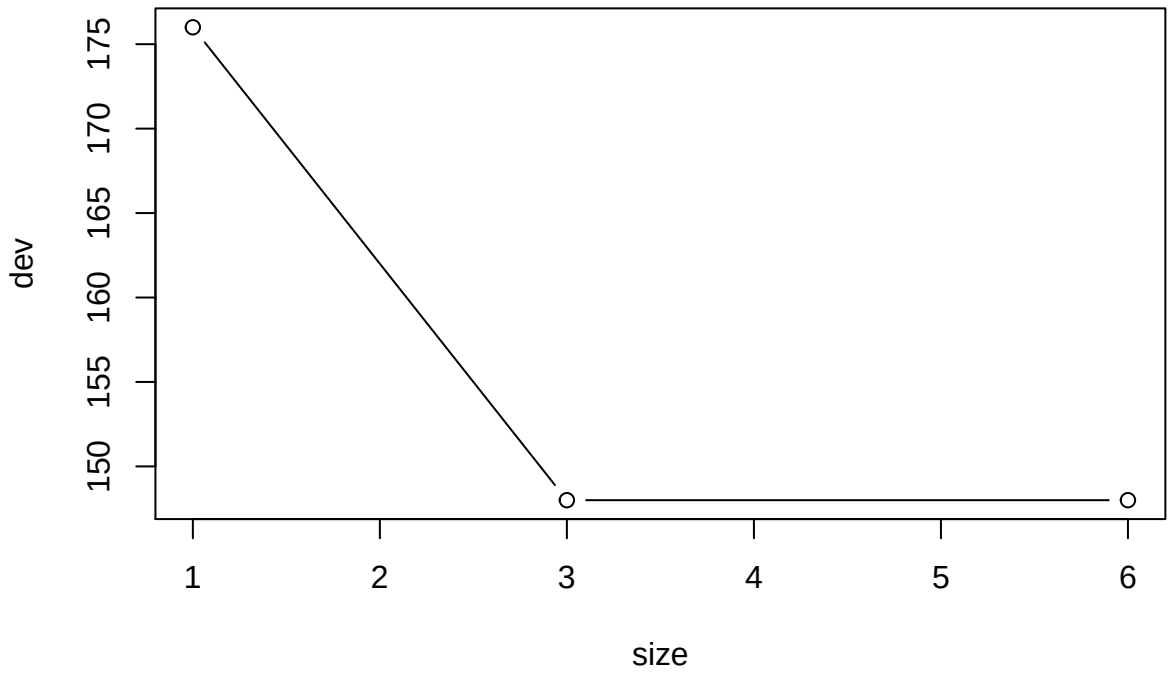
test error rate $= \frac{103+38}{4804+103+38+55} = 0.0282$ or 2.82%

4. (8 Possible Points) The following is a plot of the cross-validation error based on the number of nodes.

   a. Write down the code to get this plot.

```
cv.default = cv.tree(default.tree,FUN = prune.misclass)
plot(cv.default$size,cv.default$dev,type = "b",
     xlab = "size", ylab = "dev")
```

Write this down on your cheatsheet don't bother memorizing this



b. According to this plot what should be the number of nodes we can prune for the tree?

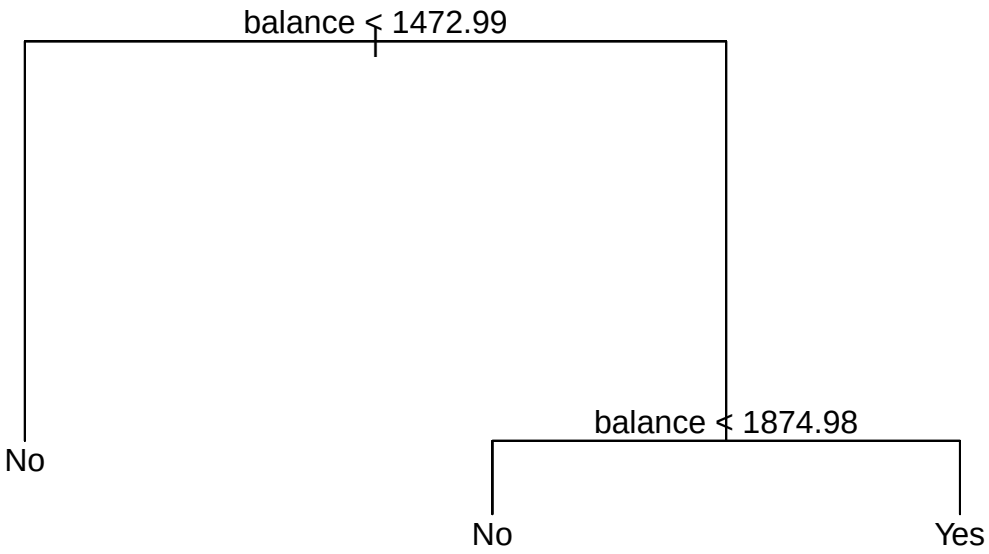
We can use 3 nodes

Look at the lowest left-most value. It slumps from 3 to 6, but since we look at the leftmost value, we take 3 as our answer.

5. (4 Possible Points) Now we can prune the tree. What is the R code to get the following pruned tree?

```
prune.default = prune.misclass(default.tree,best = 3)
plot(prune.default)
text(prune.default,pretty = 0)
```

Write it down



6. (3 Possible Points) The following is the confusion matrix based on the pruned tree. What is the test error rate? Compare this to the test error rate of the unpruned tree.

		Observed Class	
		No	Yes
Predicted Class	No	4804	103
	Yes	38	55

Look at the previous error rate if you're still struggling

test error rate $= \frac{103+38}{4804+103+38+55} = 0.0282$ or 2.82%, this is the same as the unpruned tree.

For comparison, just say whether its the same, lower, or greater

Part 3

1. (15 Possible Points) Suppose we want to estimate the median value of a population. Outline the steps needed to derive a bootstrap estimate of the median from a sample of 100 observations.
   Step 1: From a random sample of 100 observations, we resample 100 times **with replacement**.
   Step 2: Calculate the median of that resample.
   Step 3: Repeat steps 1 and 2 B times. Step 4: Find the mean of the B medians. Step 5: Calculate the standard deviation of the B medians. This is the standard error.

   just have this on your cheatsheet

2. (5 Possible Points) The following is a output from estimating the median based on the bootstrap method.

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = s.data, statistic = median.fun, R = 1000)
##
##
## Bootstrap Statistics :
##     original     bias    std. error
## t1* 25.17468 0.02272656    0.296521
```

add original + bias to get bootstrap estimate
you may be off by the std. error

a. What is the bootstrap estimate of the median value?

25.17468 + 0.02273 = 25.1974

This is the standard deviation of the B medians. We may be off by 0.2965.

Part 4: Multiple Choice

Circle the best answer. Each question is worth 5 points, for a total of 25 points for this part.

1. In R what is the package to create a single a classification tree?

   a. tree

   know this please

   b. randomForest

   c. mass

   d. boot

   e. None of these

2. What approach randomly divides the set observations into two parts, a training set and **one** hold-out set? Then the model is fit on the training set, and the fitted model is used to predict the responses for the observations in the hold-out set.

   a. Validation set approach

   if its randomly divided into two sets, its validation set approach the two sets are training and validation

   b. Leave-one-out cross validation

   c. K-fold cross validation

   d. Bootstrap method

   e. All of these methods

3. What method helps us select the best number of terminal nodes for a pruning a decision tree?

   a. Validation set approach

   k-fold cross validation is best of its bias-variance tradeoff which is both accurate and variable

   b. Leave-one-out cross validation

   c. K-fold cross validation

   d. Bootstrap method

   e. All of these methods

4. Which statement is **not true** about the K-fold cross validation?

you're getting the average in k-fold
so its always changing

   a. Data is randomly divided into K subests.

   b. There are K mean squared errors calculated based on a different group of observations treated as the validation set.

   c. This method results in not overestimating the test error rate as much as the validation set approach.

   d. The results of the mean squared error are always the same, regardless of the size of K.

   e. The cross-validation estimate of the test error rate is the average of the K MSEs.

5. Which method grows B large un-pruned trees with a random sub set of predictors (either $\sqrt{p}$ or $p/3$) for each tree, then averages the resulting predictions from the B trees?

   a. Bagging

   b. Random forest

   c. Boosting

   d. Pruning

   e. Cross-Validation

if you still are not sure, then
"random subset of predictors (either sqrt(p) or p/3)"
if you still are not sure, then look at the lecture slides