Exam 2 A - MATH 4322

Cathy Poliak

Fall 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.

Part 1

We would like to predict the `mpg` based on some features. Here are the variables.

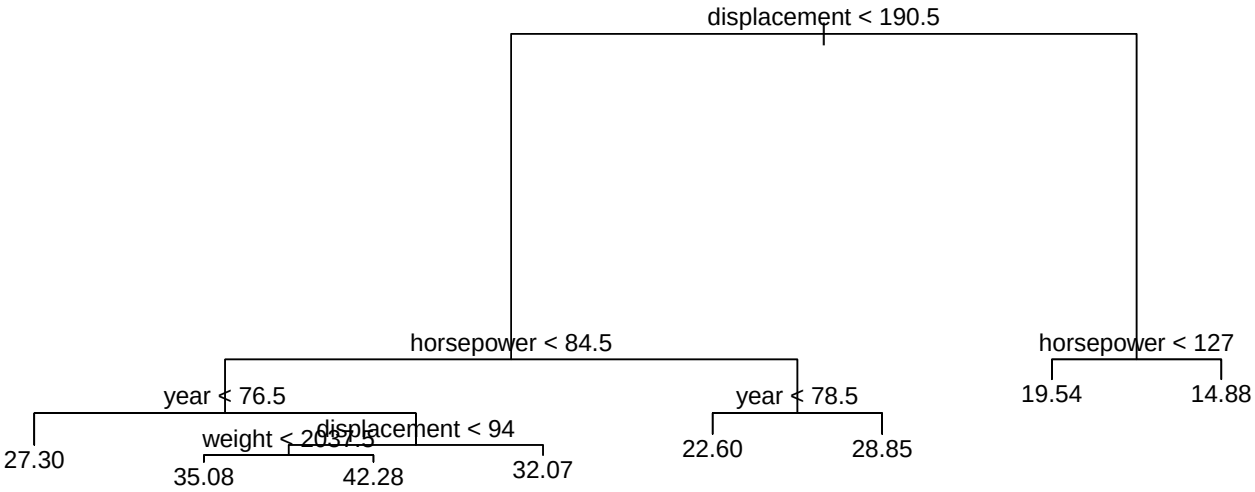
- `mpg`: miles per gallon (response variable)
- `cylinders`: Number of cylinders between 4 and 8
- `displacement`: Engine displacement (cu. inches)
- `horsepower`: Engine horsepower
- `weight`: Vehicle weight (lbs.)
- `acceleration`: Time to accelerate from 0 to 60 mph (sec.)
- `year`: Model year (modulo 100)
- `origin`: Origin of car (1. American, 2. European, 3. Japanese)

it's regression since we are trying to measure a value (mpg) which is quantitative

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

This is a regression problem. Since the response variable "mpg" is quantitative.

2. (8 Possible Points) The following is the output based on the decision tree to predict the average `mpg` based on the predictors listed above.



a. Write down the full `R` code to produce this tree. Including the package that we need to call.

Answer
3 points

add to cheatsheet

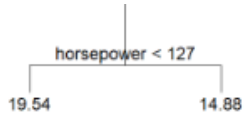
```
library(tree)
tree.auto = tree(mpg ~ . -name, data = Auto2, subset = train)
plot(tree.auto)
text(tree.auto,pretty = 0)
```

b. List all the predictors that appear in the tree.

Answer
2 points - displacement, horsepower, year, and weight

look at the tree, 4 main predictors
displacement
horsepower
year
weight

c. How do we interpret the following terminal nodes? Interpret the numbers at the bottom.



if horsepower < 127, then goto 19.54
if >, then goto 14.88

Answer 3 points - If the horsepower is less than 127,then the mpg is predicted to be 19.54.
If the horsepower is at least 127, then the mpg is predicted to be 14.88.

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

   a. (3 points) Give the formula for the mean squared error (MSE).

please know this

Answer

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

```
b.  (2 points) Match the correct MSE with: a. Single tree or b. Random Forest.
     i.  6.759509  **Random Forest**
     ii.  12.33831 **Single Tree**
```

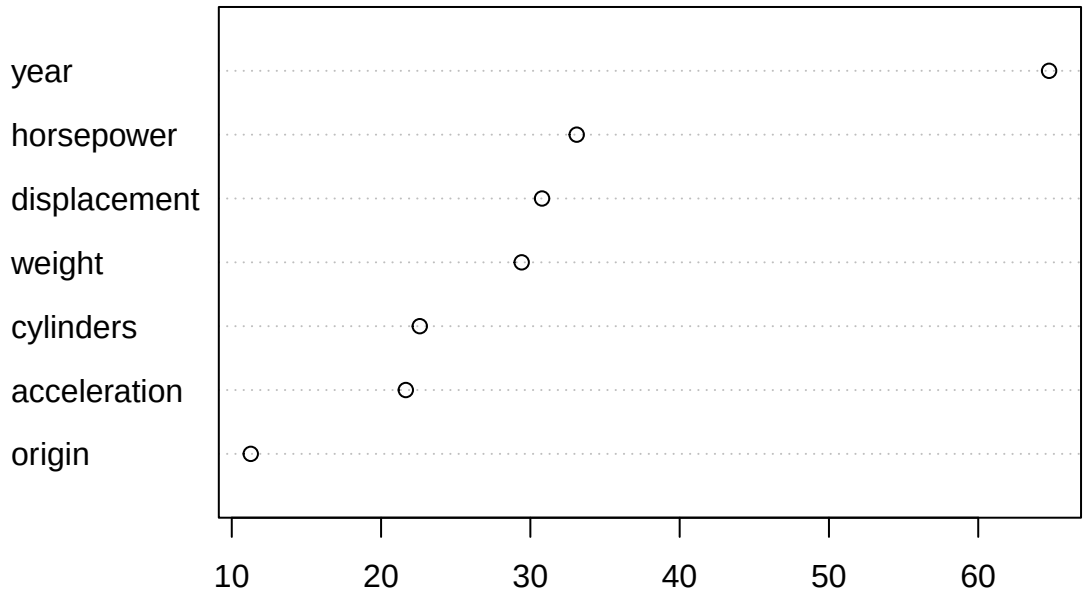
smallest is always RF
then bagging, then single tree
no bagging here but just so you know

```
c.  (5 points) Interpret the MSE value of 12.33831?
```

Answer
With this tree produced, the predicted mpg will be off by $\sqrt{12.33831} = 3.512594$ on average.

like shown, take the square root of 12.33831
thats how much we will be off by on average

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.



its asking for 3 most important so look at 3 biggest features

Answer
The three most important features are year, horsepower and displacement. This does not really match up with the original tree because displacement is at the top of the tree.

5. (3 Possible Points) Give the full `R` code to get the plot above. Including the code to get the model.

Answer
Need to call the library.
Have mpg as the response variable and make sure that importance = TRUE is in the function.
Call the plot.

add to cheatsheet

```
library(randomForest)
rf.auto = randomForest(mpg ~ .-name, data = Auto2,
                           mtry = (ncol(Auto2)-2)/3,
                           importance = TRUE)
varImpPlot(rf.auto)
```

Part 2

We are wanting to predict survival on the Titantic and determine which features help us determine if a person survived or not. The following features are used.

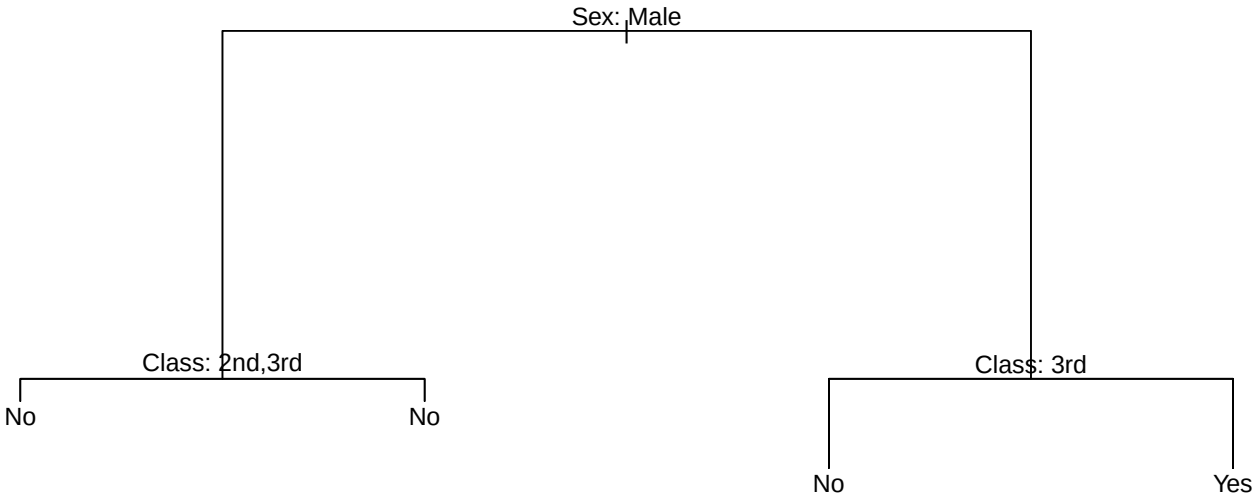
2 categories, either youre a woman or you die

- `Class`: 1st, 2nd, 3rd, or Crew
- `Sex`: Male or Female
- `Age`: Child or Adult
- `Survived`: No or Yes (response variable)

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

Answer
This is a classification problem. Our response variable Survived is categorical.

2. (8 Possible Points) The following is the output based on the decision tree to predict the `Survival` of a passenger.



a. Write down the full `R` code to produce this tree, including what package we need to call.

Answer
2 points

add to cheatsheet

```
library(tree)
tree.titanic = tree(Survived ~ . , data = titanic, subset = train)
plot(tree.titanic)
text(tree.titanic,pretty = 0)
```

b. List all the predictors that appear in the tree.

Answer
look at the tree, only 2 predictors
1 point
Sex and Class.

c. Describe from the tree what type of passenger is predicted to survive.

Answer
1 point
A female who is not in 3rd class is the only type of passenger that is predicted to survive.

d. (1 points for each question) The following is the output of the decision tree. Interpret node 3) by answering the following questions.

i. How many passengers are in this node? **339**
ii. How is this node separated? **Sex: Female**

format: node), split, n, deviance, yval, (yproblow yprobhigh)
look at the n for node 3, 339
look at the split for node 3

look at yprobhigh for node 3 (yprobhigh bc we want YES
if it was no, we would look at yproblow
we look at yval for the survival prediction which is yes

iii. What is the proportion of passengers that survived (`Yes`) in this node? **74.63%**
iv. What is the overall `Survival` prediction for this node *Yes* or *No*? **Yes**

```
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 1541 1974.0 No ( 0.6606 0.3394 )
  2) Sex: Male 1202 1281.0 No ( 0.7754 0.2246 )
    4) Class: 2nd,3rd 487  428.5 No ( 0.8398 0.1602 ) *
    5) Class: 1st,Crew 715  832.0 No ( 0.7315 0.2685 ) *
  3) Sex: Female 339  384.0 Yes ( 0.2537 0.7463 )
    6) Class: 3rd 139  192.7 No ( 0.5036 0.4964 ) *
    7) Class: 1st,2nd,Crew 200  111.5 Yes ( 0.0800 0.9200 ) *
```

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

		Observed Class	
		No	Yes
Predicted Class	1	468	118
	2	4	70

test error rate is the only thing we need for confusion matrix from test 1

test error rate = false positive + false negative over total

false positive and false negative are the mismatched diagonals (where yes and no vertically dont match up with yes and no horizontally), where 1 is no and 2 is yes

even if we didnt know what 1 and 2 were error rate is below 1 usually, so take the smaller diagonal as the mismatched diagonal

Answer

$\frac{122}{660} = 0.1848$

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

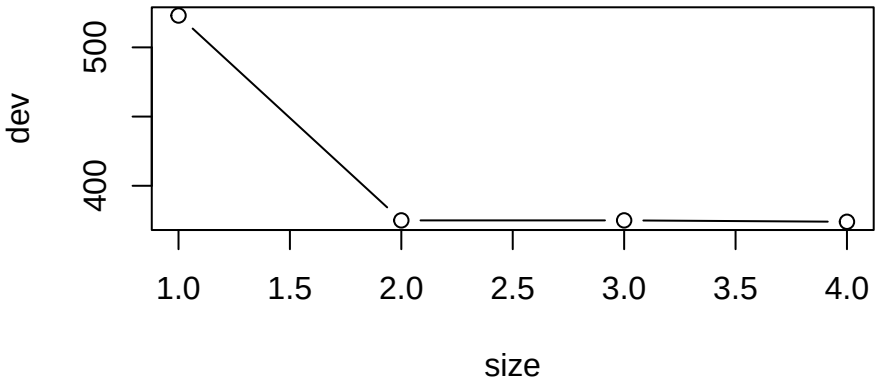
   a. (4 points) Write down the full R code to get this plot.

Answer

add to cheatsheet

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

b. (4 points) According to this plot what should be the number of nodes we can prune for the tree?



we look at the lowest left-most value

it slumps from 2.0 to 4.0 but we take only the 2.0 since its the leftmost value

Answer It looks like 2 nodes would be the best number of terminal nodes.

5. (3 Possible Points) Now we can prune the tree. What is the full R code to get a pruned tree?

add to cheatsheet

```
prune.titanic = prune.misclass(tree.titanic,best = 2)
plot(prune.titanic)
text(prune.titanic,pretty = 0)
```

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	432	97
	Yes	40	91

error rate again, but we know what 1 is and what 2 is

Answer

remember to compare to the previous error rate

$\frac{137}{660} = 0.2076$

This is a little bit higher than the miss classified error from the full tree.

Part 3

Given for a linear neural network model with input variables x_1, x_2, x_3, and two hidden layers h_1, h_2 you are given the following parameter values:

input & **hidden** layer:

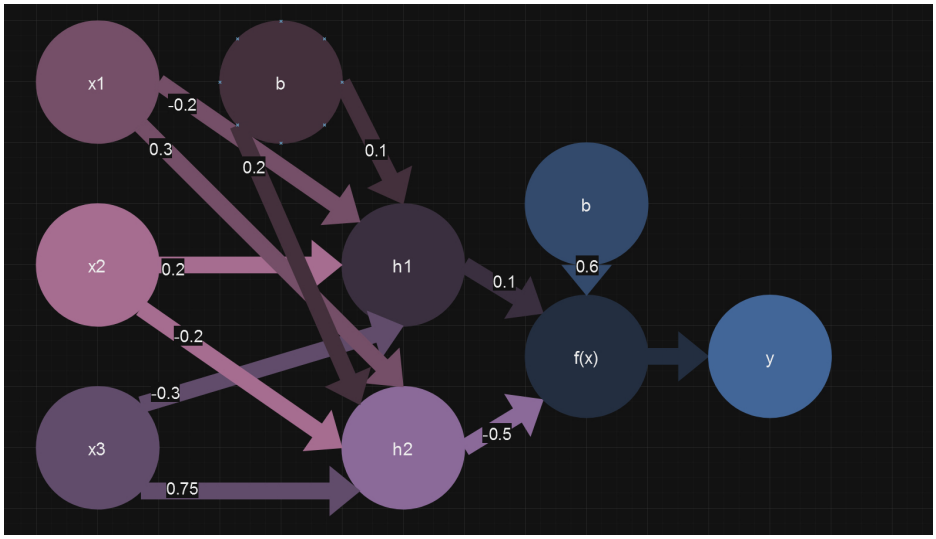
Input \ Hidden	h_1	h_2
1 (bias)	0.1	0.2
x_1	-0.2	0.3
x_2	0.2	-0.2
x_3	-0.3	0.75

hidden & **output** layer:

Hidden \ Output	y
1 (bias)	0.6
h_1	0.1
h_2	-0.5

1. (10 points) Draw a mathematical model of this linear neural network.

Answer
Make sure there is nodes for each x, hidden node and 2 biases and y - 5 points
The weights are on the graph - 5 points



kinda messy but lemme explain

we have 3 input variables x1, x2, x3
we have 2 hidden layers h1, h2
as we approach each layer, each input variable has a value associated
x1 -> h1 = -0.2, x1 -> h2 = 0.3
x2 -> h1 = 0.2, x2 -> h2 = -0.2
x3 -> h1 = -0.3, x3 -> h2 = 0.75

as hidden layers approach output, they too have values associated
h1 -> f(x) = 0.1, h2 -> f(x) = -0.5

bias should be self explanatory

2. (10 points) Calculate the output for the case of $x_1 = 2$, $x_2 = -4$, and $x_3 = 5$. Show your work.

$$\hat{h}_1 = 0.1 - 0.2(2) + 0.2(-4) - 0.3(5) = -2.6$$
$$\hat{h}_2 = 0.2 + 0.3(2) - 0.2(-4) + 0.75(5) = 5.35$$
$$\hat{y} = 0.6 + 0.1(-2.6) - 0.5(5.35) = -2.335$$

equation format for the nth hidden layer is hn = bias_n + x1_n + x2_n + x3_n where n is the hidden layer value(for example, n is

Part 4: Multiple Choice

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

1. When a given method yields a small training MSE but a large test MSE, we are said to be

   a. Underfitting the data

   b. **Overfitting the data**

   c. Exactly right

   d. Biased of the data

   e. All of these are true.

   overfitting is when the model memorizes the dataset used to train so its good at the training data but not other data

2. What method is to select a subtree from a very large tree that leads to a best lowest test error rate?

   a. Random Forest

   b. Bagging

   c. Boosting

   d. **Pruning**

   e. Cross-Validation

   Pruning solves the problem of overfitting, yielding the lowest test error rate

3. Below is an output for estimating the median based on the bootstrap method. What is the estimate of the median based on all of the bootstrap samples.

   a. 24.3

   b. -0.02

   c. 0.6719

   d. **24.28**

   e. 24.97

   add original + bias under bootstrap statistics
   |
   |
   |
   |
   |
   |
   v

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = s.data, statistic = median.fun, R = 1000)


Bootstrap Statistics :
    original      bias    std. error
t1* 24.29888 -0.01934575   0.6718975
```

4. Suppose we have a data frame with $n = 200$ observations. We want to do a 10-fold Cross-validation to determine the test MSE, how many observations are used in the test set each time?

   a. **20**
   b. 100
   c. 150
   d. 200
   e. 2000

n / k where k is the amount of folds in the k-fold cross validation, in this case 10 and n = size of observations, in this case 200

200 / 10 = 20

5. In the case of the neural networks, the *sigmoid* function, $g(z) = \frac{1}{1+e^{-z}}$ is an example of:

   a. **activation function**
   b. tree function
   c. input funciton
   d. linear function
   e. random forest function

an activation function is something we use to determine whether other neurons in the network activate or not
sigmoid and RELu are two tyes of activation functions that are used in neural networks - someone smarter than me