Stat 240 - Take Home Final Exam - Spring 2019

Due by 6 pm, Wednesday, May 15

YOUR NAME GOES HERE

Instructions

Your fully reproducible Rmd AND compiled pdf solution is due via Moodle by Wednesday, May 15th by 6 pm. NO LATE SUBMISSIONS will be accepted. Submit both files early, and check that your solution compiles well in advance of the deadline!

The solution you provide for the exam must be your own work.

The take home exam may be completed in multiple sittings. I expect it will take you several hours to complete because you will be supplying your own code for the solution. Some problems are longer than others, so I encourage you to read through them all to get a sense of each.

The exam is open class textbooks (recall these are available via the e-reserves link on Moodle), our class materials, our Moodle site, but CLOSED to all external sources for statistical information.

If needed, you may use the R help menu, files, and official package documentation for code assistance. For example, you can do ?functionname, or google the official package documentation pdf (say, for igraph, for example). All other external code references are forbidden. It is inappropriate, for example, to google for the functionname, find a similar analysis done by someone else, and use that code. The help menu should be more than sufficient - you should have examples of all code needed from class. If you encounter an error and cannot resolve it from the help menu, contact me for assistance.

For your solution:

- Be sure your code is READABLE in the pdf (i.e. that it does not go over the ends of the lines)
- That all necessary packages are listed in the top chunk in the Rmd (the package list itself can be hidden in the pdf)
- Show all code necessary to reproduce your work in the pdf (i.e. ONLY the package list can be hidden in the pdf).
- Be sure to submit BOTH the pdf and Rmd files in Moodle

You may NOT consult anyone except the instructor (Amy Wagaman). If a question is asked where the entire class should be notified of the answer, the instructor will handle that notification.

General guidance: If you have a chunk of code that you are using to do something, it is helpful to state what you are using it for. It is a good idea to explain the steps you are taking for your analysis. If I cannot follow your work, it will not earn you partial credit. You won't need this for every part below, but the idea is that I need to be able to follow what you are doing. Basically, shorter code chunks are recommended, with interspersed comments, rather than ALL code and output (several pages), followed by your response.

There are 5 problems, for a total of 125 points. Point breakdown is shown by problem and by part. Partial credit may be given for problems based on your submitted work.

If desired, you may delete this line and the instructions above in your submission.

Honor Code Statement

Please add your full name as your electronic signature for agreeing to abide by the honor code and honor statement provided below.

I have abided by the Amherst College Honor Code in regards to this exam. In particular, I have neither given nor received any unauthorized assistance with this take home exam.

ELECTRONIC SIGNATURE HERE

Question 1 - 29 points

Letter recognition is a challenging problem with wide-ranging applications. For example, think of all the mail that goes through the post office - numerical recognition of zip code digits is very helpful, and so is letter recognition for addresses. For this problem, you have a filtered (for size) version of a data set from UCI's machine learning repository to work with. A separate pdf of a data dictionary is provided for you. The data set and a few other variants of it are provided for you to use. We are interested in correctly predicting which capital letter we have based on attributes of it when viewed as an image with pixel characteristics.

```
#setup chunk, do not alter
letters <- read_csv("https://awagaman.people.amherst.edu/stat240/letters2.csv")
letters <- mutate(letters, letter = factor(letter))

#create train/test set
letters <- letters %>% mutate(id = row_number())
set.seed(240)
lettertrain <- letters %>% sample_frac(0.75)
lettertest  <- anti_join(letters, lettertrain, by = 'id')

letters <- select(letters, -id) #drop id variable
lettertrain <- select(lettertrain, -id) #drop id variable
lettertest <- select(lettertest, -id) #drop id variable
```

a. (2 points) What are the classes in the filtered data set? How many are there? Is any class rare? Provide supporting output.

SOLUTION:

b. (6 points) Use a classification tree with parameter values chosen by you to optimize the resulting solution to obtain an AER and estimated TER. Report your findings, including your chosen parameters, AER, estimated TER, and what method you used to obtain the estimated TER. Using all default parameter values does not demonstrate that you optimized the solution.

SOLUTION:

c. (6 points) Investigate the applicability of linear discriminant analysis (LDA) to this classification problem. Does an LDA solution seem viable using all possible predictors? Any subset of predictors? Report on your findings, including how you assessed LDA's viability with appropriate support. If LDA is viable, perform an LDA and obtain an AER.

Note: you just need appropriate support, not EVERY piece of support you have. For example, you might show one piece of output and say that this pattern was also present in other instances, instead of showing tons of output.

SOLUTION:

d. (6 points) Apply nearest neighbor methods with parameter values chosen by you to optimize the solution to find an AER and estimated TER. Report your findings, including your chosen parameters, AER, estimated TER, and what method you used to obtain the estimated TER. Using k=1 is not recommended.

SOLUTION:

```
#if you get an error about unused arguments, try class:: before the function call
```

e. (6 points) Use boosting with parameter values chosen by you to optimize the resulting solution to obtain an AER and estimated TER. Report your findings, including your chosen parameters, AER,

estimated TER, and what method you used to obtain the estimated TER. Using all default parameter values does not demonstrate that you optimized the solution.

SOLUTION:

f. (3 points) Consider the models you fit above. Choose one as your optimal model for predicting the capital letters, and state your rationale/supporting evidence for your choice.

SOLUTION:

Question 2 - 20 points

You previously saw part of the avocado data set used on midterm 2. In this problem, you will be working with a smaller subset (only 500 of the roughly 18000 observations) in order to help avocado growers with some visualizations. The original data set is from Kaggle, and a data dictionary has been provided for you in a separate pdf document.

```
#load the subset of the data available
avocado <- read_csv("https://awagaman.people.amherst.edu/stat240/avocadosubset.csv")
```

Your task for this problem is to create two multivariate (not just bivariate) visualizations to share with avocado growers that point out interesting items of note in the data set. Present your well-constructed visuals as well as your "story" for what you found that is interesting in a few paragraphs. You may use univariate and bivariate visuals as additional support as needed for your story.

SOLUTION:

```
#add chunks and reorganize as needed
```

Question 3 - 28 points

For this problem, we consider data on dolphins. The data is presented as an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand. The data set is from D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, Behavioral Ecology and Sociobiology 54, 396-405 (2003).

```
dolphins <- read_graph("https://awagaman.people.amherst.edu/stat240/dolphins.gml", format = "gml")
```

a. (6 points) Obtain some appropriate descriptive statistics to describe basic properties of the network. Use your statistics to write a paragraph description of the network that could be used as an introduction to it for the data section of an article.

SOLUTION:

b. (6 points) Which dolphins are most central to the network? Use at least two centrality measures to determine which dolphins are most central. Are there clear "central" dolphins here? Report on your findings, including whether the centrality measures agree.

SOLUTION:

c. (4 points) How dense and connected is the network? Use at least 2 appropriate descriptive statistics to write sentences to address this question.

SOLUTION:

d. (8 points) What communities are present in the dolphin network? Use at least two algorithms to address this question. Present your findings and compare the solutions the two algorithms give you.

SOLUTION:

e. (4 points) Which community structure from part d do you prefer for the dolphin network? Why? Explain using at least one appropriate statistic as support for your choice.

SOLUTION:

Question 4 - 26 points

In this question, we will examine course evaluation data from the University of Texas. A separate pdf with a data dictionary has been provided for you to review. Analysts have a few questions they want you to address using the data.

```
courseeval <- read_csv("https://awagaman.people.amherst.edu/stat240/courseeval.csv")
```

a. (6 points) In what ways are student evaluations typically the most different? In other words, the analysts want to know how student evaluations are most likely to vary from one to the next. Perform a PCA to help address this question. Report your findings, list choices you made in performing the PCA such as the number of PCs kept, and associated rationale.

SOLUTION:

b. (4 points) Interpret the PCs you have chosen to keep.

SOLUTION:

c. (6 points) Are students picking up on some underlying instructor or course characteristic in their evaluations? Perform a factor analysis to help address this question. Report your findings, list choices you made in performing the factor analysis including the final number of factors and chosen rotation, and associated rationale.

SOLUTION:

d. (4 points) Interpret the factors you have chosen to keep. If you felt no solution was appropriate above, chose one to use for practice interpreting.

SOLUTION:

e. (6 points) Compare and contrast your PCA and FA solutions. What do you learn from each? Which do you think provides more insights into this data set? Explain your response.

SOLUTION:

Question 5 - 22 points

This problem uses the letters data set from question 1. Investigators want to know which letters, if any, naturally group together based on their pixel-based characteristics.

a. (10 points) Use two different algorithms to find natural groupings of the letters in the data set. Do you find strong natural groupings? Report your findings, including what algorithms you used, parameter values for those algorithms (and how you chose them), how many groups you find, and compare the solutions from the two algorithms. Which solution do you prefer? Why?

SOLUTION:

b. (4 points) Do either of your solutions seem to recover specific letters or letter groupings? If so, what letters are grouped together?

SOLUTION:

c. (8 points) Are your preferred natural groups really present? We can assess this by running a different algorithm to see if the groups you found are well-recovered. Use a random forest with parameter values

chosen by you to optimize the resulting solution to assess how well your groups are recovered. Report your findings, including AER and estimated TER. How well are your natural groups recovered?

SOLUTION:

```
#you don't want a regression random forest
```