Final Exam - Take Home

BSAN 450 (Spring 2022)

This Final Exam is due on May 13, 2022 at 6 PM Central. The total points possible are 100 and there are four (4) questions, each carrying equal points. You can discuss the exam with your friends or group members but you will submit your own solutions to Canvas, either in word or pdf. You are expected to write your own codes and produce your own report. Copying other student's code or report is not allowed and will constitute cheating.

Considerations in Grading:

I am not just interested in the final output and R code, but also in the process you go through to arrive at your model. A significant amount of the grade for the exam will be determined by the process you use to come up with your final model. If you have a poor process or do not clearly describe that process your grade will be reduced significantly.

---

1. The time series of the monthly sales for a souvenir shop on the wharf at a beach resort town in Queensland, Australia is in the file named `souvenir.csv` and the name of the time series is `Sales`. Find a time series model that you believe is appropriate for this data. Document the steps you used to find this model and justify your choice of model.

   In developing this model you need to do the following: *For each step of your development, clearly describe what you are doing and your reasons for taking this step. If you do not describe your reasoning you will lose points because I cannot read your mind. I need to know the rationale for what you are doing in developing the model.*

2. The time series of the number of males that are employed in non-agricultural industries is in the file named `emales.csv` and the time series name is `Employed`. Find a time series model that you believe is appropriate for this data. Document the steps you used to find this model and justify your choice of model.

   In developing this model you need to do the following: *For each step of your development, clearly describe what you are doing and your reasons for taking this step. If you do not describe your reasoning you will lose points because I cannot read your mind. I need to know the rationale for what you are doing in developing the model.*

3. The data for this example is the Ames housing data available in `ames.csv`. The goal is to predict the `Sale_Price` using all independent variables. The data description is available at https://cran.r-project.org/web/packages/AmesHousing/AmesHousing.pdf. It is a good practice to make sure that your data has no missing values.

a. Split this data into two equal parts: one for training and another for testing.

b. Fit a regression tree to the training data for predicting `Sale_Price` using all available independent variables. Plot the tree, discuss the size and interpret the tree that you constructed.

c. Prune this tree using cross validation as discussed in class. Plot the pruned tree, discuss the size and interpret the pruned tree that you constructed. Now compare the MSE of the pruned tree and the unpruned tree on your test data. Which one has a lower MSE? Which tree would you recommend?

d. In this sub-part, you will construct a Bagged Regression tree on the training data and compute the MSE of your Bagged regression tree on the test data. Rememeber to carefully specify the `mtry` parameter.

e. In this sub-part, you will construct a Random Forest Regression tree on the training data and compute the MSE on the test data. Rememeber to carefully specify the `mtry` parameter.

f. Finally, here you will develop a Boosted Regression tree on the training data and compute the MSE on the test data. You may try tweaking the `n.tree` and `shrinkage` parameters to see how it affects your test set MSE.

g. Which tree amongst the four different regression trees that you constructed has the lowest MSE on the test set?

4. The data `nci.data` has 64 rows and 6,830 columns. Each row is a cell line for which we have gene expressions recorded for 6,830 genes. The `nci.labs` data has the cancer type for each of these 64 cell lines. Your analysis will rely primarily on `nci.data` and not on `nci.labs` as the latter are just labels.

```
library(ISLR)
nci.labs=NCI60$labs
nci.data=NCI60$data
```

a. Perform PCA on `nci.data` with `scale=TRUE` in the `prcomp` function.

b. Plot the first two principal components. You will notice from your plot that observations belonging to a single cancer type tend to lie near each other in this low-dimensional space.

c. As demonstrated in the lecture, plot the PVE of each principal component (i.e. a scree plot) and the cumulative PVE of each principal component. What is the % of variance explained by the first seven principal components? Based on the `scree plot`, how many principal components should we choose?

d. Scale the `nci.data` such that the features have zero mean and 1 standard deviation. You will have to use the `scale()` function in R for this purpose. Then, on the scaled data, perform K-means clustering with $K = 4$. What are the cluster sizes that you obtain? Calculate the Between Cluster sum of squares as a % of total sum of squares.