Modern Data Mining - HW 2

Group Member 1
Group Member 2
Group Member 3

Overview / Instructions

This is homework #2 of STAT 471/571/701. It will be **due on 10 October, 2017 by 11:59 PM** on Canvas. You can directly edit this file to add your answers. Submit the Rmd file, a PDF or word or HTML version with only 1 submission per HW team.

Solutions will be posted. Make sure to go through these files to pick up some tips.

R Markdown / Knitr tips

You should think of this R Markdown file as generating a polished report, one that you would be happy to show other people (or your boss). There shouldn't be any extraneous output; all graphs and code run should clearly have a reason to be run. That means that any output in the final file should have explanations.

A few tips:

- Keep each chunk to only output one thing! In R, if you're not doing an assignment (with the `<-` operator), it's probably going to print something.
- If you don't want to print the R code you wrote (but want to run it, and want to show the results), use a chunk declaration like this: `{r, echo=F}`
- If you don't want to show the results of the R code or the original code, use a chunk declaration like: `{r, include=F}`
- If you don't want to show the results, but show the original code, use a chunk declaration like: `{r, results='hide'}`.
- If you don't want to run the R code at all use `{r, eval = F}`.
- We have shown examples in our lectures files.
- For more details about these R Markdown options, see the documentation.
- **Delete the instructions and this R Markdown section, since they're not part of your overall report.**

Problem 0

Review the code and concepts covered during lecture: model selection and penalized regression through elastic net.

Problem 1: Model Selection

Do ISLR, page 262, problem 8, and write up the answer here. This question is designed to help understanding of model selection through simulations.

Problem 2: Regularization

Crime data continuation: We use a subset of the crime data discussed in class, but only look at Florida and California. `crimedata` is available on Canvas; we show the code to clean here.

```
crime <- read.csv("CrimeData.csv", stringsAsFactors = F, na.strings = c("?"))
crime <- dplyr::filter(crime, state %in% c("FL", "CA"))
```

Our goal is to find the factors which relate to violent crime. This variable is included in crime as `crime$violentcrimes.perpop`.

A) EDA

- Clean the data first
- Prepare a set of sensible factors/variables that you may use to build a model
- Show the heatmap with mean violent crime by state. You may also show a couple of your favorate summary statistics by state through the heatmaps.
- Write a brief summary based on your EDA

B) Use LASSO to choose a reasonable, small model. Fit an OLS model with the variables obtained. The final model should only include variables with p-values < 0.05. Note: you may choose to use lambda 1se or lambda min to answer the following questions where apply.

1. What is the model reported by LASSO?
2. What is the model after running OLS?
3. What is your final model, after excluding high p-value variables? You will need to use model selection method to obtain this final model. Make it clear what criterion/criteria you have used and justify why they are appropriate.

C) Now, instead of Lasso, we want to consider how changing the value of alpha (i.e. mixing between Lasso and Ridge) will affect the model. Cross-validate between alpha and lambda, instead of just lambda. Note that the final model may have variables with p-values higher than 0.05; this is because we are optimizing for accuracy rather than parsimoniousness.

1. What is your final elastic net model? What were the alpha and lambda values? What is the prediction error?
2. Use the elastic net variables in an OLS model. What is the equation, and what is the prediction error.
3. Summarize your findings, with particular focus on the difference between the two equations.

B+) Repeat similar stepts as that of **B)** but start with the set of variables that also include all two way interactions

1. How many variables do you have now?
2. Comparing the final models with the ones from **B)**, which one would you use? Commenting on your choice.