Homework 7

STAT 430, Spring 2017

Due: Friday, March 31 by 11:59 PM

Please see the homework instructions document for detailed instructions and some grading notes. Failure to follow instructions will result in point reductions.

Exercise 1

[**25 points**] For this assignment, we will use the `College` data from the `ISLR` package. Familiarize yourself with this dataset before performing analyses. We will attempt to predict the `Outstate` variable.

Test-train split the data using this code.

```
set.seed(42)
library(caret)
library(ISLR)
index = createDataPartition(College$Outstate, p = 0.80, list = FALSE)
college_trn = College[index, ]
college_tst = College[-index, ]
```

Train a total of **six** models using five-fold cross validation.

- An additive linear model.
- An elastic net model using additive predictors. Use a `tuneLength` of `10`.
- An elastic net model that also uses all two-way interactions. Use a `tuneLength` of `10`.
- A well-tuned KNN model.
- A well-tuned KNN model that also uses all two-way interactions. (Should this work?)
- A well-tuned GAM, trained using `method = gamSpline` with `caret`.

Before training the models, set a seed equal to your UIN.

```
uin = 123456789
set.seed(uin)
```

Also answer the following:

- Create a table which reports CV and Test RMSE for each.
- Based on the table, which model do you prefer? Justify your answer.
- For both of the elastic net models, report the best tuning parameters from `caret`. For each, is this ridge, lasso, or somewhere in between? If in between, closer to which?
- Did you scale the predictors when you used KNN? Should you have scaled the predictors when you used KNN?
- Of the two KNN models which works better? Can you explain why?
- For both of the KNN models, plot the CV results against the tuning parameters. Does this plot verify that you used an appropriate tuning grid?
- For the GAM, plot the CV results against the tuning parameters. Does this plot verify that you used an appropriate tuning grid?
- What was the best tuning parameter for the GAMs? Does this suggest non-linearity?
- What year is this dataset from? What was out-of-state tuition at UIUC at that time?

Exercise 2

[**5 points**] Continue using the `College` data. Now use `Private` as the response variable. Fit Regularized Discriminant Analysis trained using five-fold cross-validation and a tuning length of `5` with `train()`. Use the seed below.

```
set.seed(42)
```

Report the tuning parameters and CV-Accuracy of the chosen model. Is this LDA, QDA, or something else? Also report test accuracy