Statistics 652 - Midterm-Final

Prof. Eric A. Suess

February 24, 2021

Midterm

For the titanic data set try the following machine learning classification algorithms.

Use the training and test datasets from the **titanic** R package.

You should note that the *titanic_train* has the *Survived* variable and the *titanic_test* does not. So to select your best model you need to use the *titanic_train* dataset to train and test your models. So that means you will need to select a training dataset from *titanic_train* and select a testing dataset (this would be a validation dataset) from *titanic_train* to evaluate the models you try.

I have not demonstrated the use of cross-validation, once you are confortable running all of the models see if you can figure out how to use cross-validation to pick the best model.

Once you have picked the best model you should do the following:

1. Re-run your chosen model on the full *titanic_train* dataset.
2. Then produce predictions for the *titanic_test* dataset. This is what you would submit in a .csv to kaggle in a competition.

Build **classification models** for the *Survived* variable. Pick a model scoring function and determine which model is the best. I would suggest making a confusion matrix and computing the accuracy or kappa.

0. Null Model
1. kNN (the sample code given did not scale or normalize, if you use this model you need to do that.)
2. Boosted C5.0
3. Random Forest
4. Logistic Regression using regularization
5. Naive Bayes

Extra Credit:

Make one plot containing all of the ROC curves for the algorithms trained.

Data

```
library(titanic)

data(titanic_train)
data(titanic_test)

head(titanic_train)
head(titanic_test)
```

Final

For the *Ozone* data from the R package *mlbench* try the following machine learning prediction algorithm that is useful for feature selection.

Read the paper Feature Selection with the Boruta Package and implement the algorithm.

Which features are most important as determined by the Boruta RandomForest Algorithm?