Homework 3: Naive Bayes and Decision Trees

Hieu Pham

Instructions

You can submit your homework in one of two ways.

1. You can fill in the missing code blocks directly in this .Rmd file (be sure to change the file name)
2. You can create a new .R file and clearly label you answers.

Getting Started

In this assignment, we will be working with two datasets. One about breast cancer and the other about future car prices.

The packages listed below are simply suggestions, but please edit this list as you see fit.

```
## you can add more, or change...these are suggestions
library(tidyverse)
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(caret)
library(Metrics)
library(randomForest)
```

Problem Set

1. How does the random forest differ from a decision tree?

2. What is the difference between classification and regression?

3. When evaluating a predictive machine learning model, how is using an independent test set different than using cross validation?

4. Suppose for a predictive problem, you used an independent test set and had an 80% accuracy. Then for the same predictive problem, you used cross validation and had an accuracy of 85%. Are these two numbers comparable?

5. suppose you build a predictive algorithm, you had an MAE of 123.53 and an RMSE or 423.51. Are these two numbers meaningful? If yes, they are explain why. If not, explain what other information you need to make these numbers meaningful.

6. Read in the `BreastCancer.csv` dataset, and remove any rows with missing values. Using an independent test set with an 80/20 split, build a 1. Naive Bayes Classifier, 2. Decision Tree Classifier, 3. Random Forest Classifier using accuracy and Recall as your evaluation metric.

- Which algorithm has the best accuracy?
- Which algorithm has the best Recall measure?
- According to the decision tree, what are the top three most important features?

7. Now that you have a predictive model for predicting breast cancer in patients, how can you use this information to make decision? That is, how would you embed this information in a prescriptive analytics pipeline?

8. Read in the `CarPrice.csv` dataset, and remove any rows with missing values. Using cross validation with an 3 folds, build a 1. Decision Tree Regressor, 2. Random Forest Regressor using MAE and RMSE as your evaluation metric.

- Which algorithm has the best MAE?
- Which algorithm has the best RMSE?
- According to the decision tree, what are the top three most important features?

9. Now that you have a predictive model for predicting car prices one year in the future, how can you use this information to make decision? That is, how would you embed this information in a prescriptive analytics pipeline?