Sowmya Vijayakumar
ID number: 15233539
MSc Data Analytics

Machine Learning Assignment 2

I am using caret package in R to construct the learning curve on kNN and Random Forest algorithm. In the data preparation step, I am dividing the data into two parts for training and testing data using createDataPartition function. p is the percentage of data that goes to training and the rest goes to the test data.

Code: inTrainingSet <- **createDataPartition**(illness$test_result, p = 0.05, list = FALSE)

I varied p from 0.05 to 0.95 in steps of 10, thereby training on 5% of data and testing on 95% of data. Similarly repeating for various values of the dataset from 5% to 95%. The accuracy value is obtained by getting the confusion matrix using the function confusionMatrix in caret package. This is done for every iteration of training value of p from 0.05 to 0.95 and multiplied by 100 to get the percentage.

```
Trainingvalue<- c(0.05,0.15,0.25,0.35,0.45,0.55,0.65,0.75,0.85,0.95)
Trainingvalue_percentage = Trainingvalue * 100
```

Learning curve is obtained by plotting Accuracy vs Training value using qplot.

Code: **confusionMatrix**(knnPredict, illnessTest$test_result)

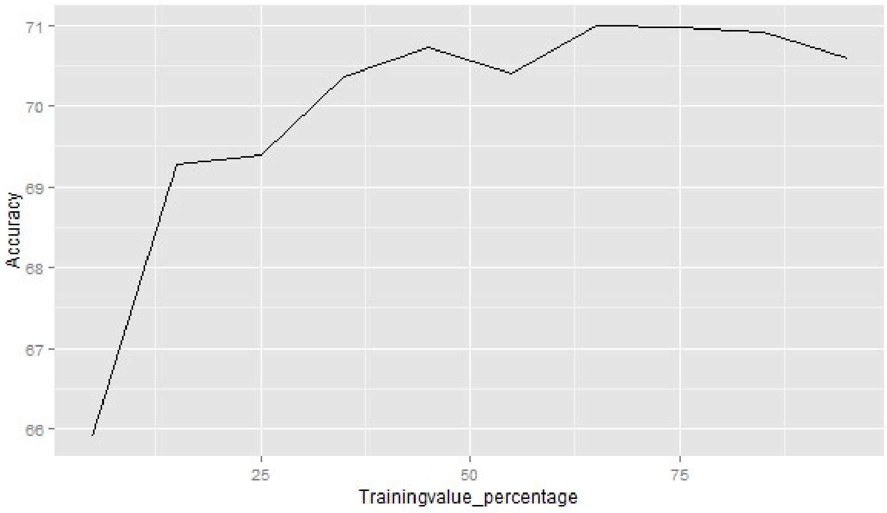
qplot(y=Accuracy,x=Trainingvalue_percentage,geom = "line")

Averaging the points to fit the curve as shown in the class.

The above procedure is repeated for both kNN algorithm and Random Forest algorithm.

Below are the graphs of the learning curves for kNN and Random Forest algorithm.

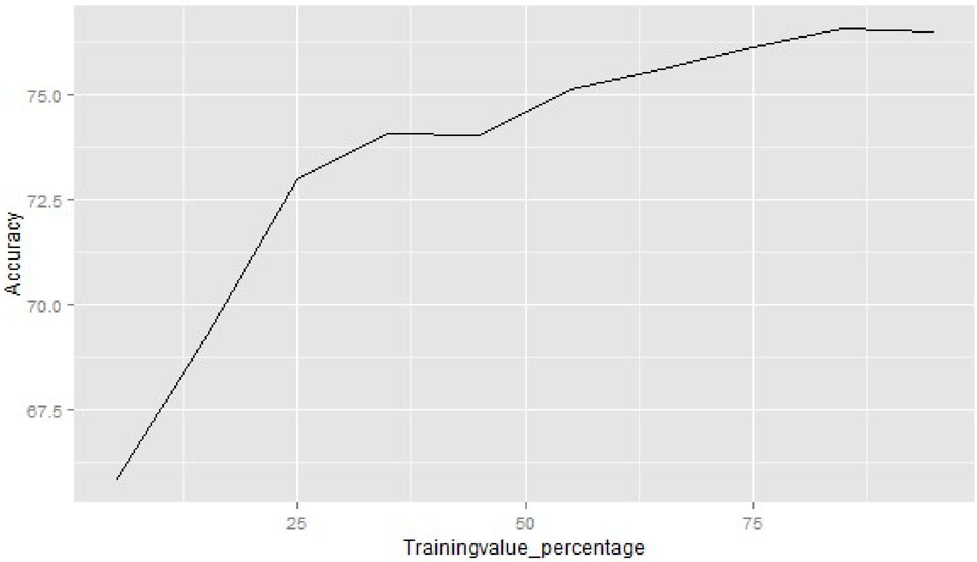
<u>kNN Learning curve:</u>



Date:29/10/15

Sowmya Vijayakumar
ID number: 15233539
MSc Data Analytics

Random Forest Learning Curve



Observations:

- By observing the pattern of the kNN learning curve and Random Forest learning curve, both perform in a similar fashion.
- Maximum value of Random Forest is 76% that is higher than kNN which is just above 70% indicating that random Forest performs better than kNN.
- For more percentage of data going to training set, Accuracy is high. The algorithm predicts better with large data set in training.
- At certain point, above 50% of the data going to the training set, there is very minimal increase in the accuracy value.
- Error rate is high initially in both algorithms, suggesting to add more training data.
- Few dips observed in both the models in the middle range showing sensitivity for high variance and after averaging the points, the lines smoothened.