Practical Machine Learning

Prediction Modelling Assignment

The purpose of this assignment is to predict the manner in which exercise is performed by using quantified data collected from the wearable devices like Jawbone, Nike Fuel band and Fitbit by using machine learning algorithms.

The data for this project come was acquired from <http://groupware.les.inf.puc-rio.br/har>.

The training and test data sets were downloaded from the respective links mentioned below:

<https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv>

<https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv>

I have changed the working directory to my current working directory where my datasets have been saved.

setwd("C:/Assignment/COURSERA/PRACTICAL MACHINE LEARNING")

Since there is a lot of missing data in the training and testing data sets it was essential to clean the data sets during the process of loading it in R this has been done by using the na.strings parameter option in the read.csv command.

trainingDs<-read.csv("pml-training.csv", header=TRUE, na.strings=c("NA", "#DIV/0!" ,""))

testDs<-read.csv("pml-testing.csv", header= TRUE, na.strings=c("NA", "#DIV/0!" ,""))

The training data has then been partitioned and preprocessed to eliminate variables with atleast one “NA” value.

trainingDsFiltered <- trainingDs [, apply(trainingDs, 2, function(x) !any(is.na(x)))]

trainingDsTrimmed <- trainingDsFiltered[,-c(1:8)]

testDsTrimmed<- testDs [,names(trainingDsTrimmed [,-52])]

The cleaned data set was subset in order to generate a test set independent from the 20 cases provided set. Partitioning was performed to obtain a 75% training set and a 25% test set.

library(caret)

partitionedTrain<-createDataPartition(y= trainingDsTrimmed$classe, p=0.75,list=F)

trainingData<- trainingDsTrimmed [partitionedTrain,]

testData <- trainingDsTrimmed [-partitionedTrain,]

Random forest trees were generated for the training dataset using cross-validation. Then the generated algorithm was examined under the partitioned training set to examine the accuracy and estimated error of prediction. By using 51 predictors for five classes using cross-validation at a 5-fold an accuracy of 99.2% with a 95% CI [0.989-0.994] was achieved accompanied by a Kappa value of 0.99.

set.seed(42)

fitControl<-trainControl(method="cv", number=5, allowParallel=T)

rffit<-train(classe~.,data= trainingData, method="rf", trControl=fitControl)

predrf<-predict(rffit, newdata= testData)

confusionMatrix(predrf, testData$classe)

prediction\_20<-predict(rffit, newdata= testDsTrimmed)

*# Output for the prediction of the 20 cases provided*

prediction\_20

Once, the predictions were obtained for the 20 test cases provided, the below shown script was used to obtain single text files to be uploaded to the courses web site to comply with the submission assignment. 20 out of 20 hits also confirmed the accuracy of the obtained models.

getwd()

pml\_write\_files = function(yy)

{

n = length(yy)

for (i in 1:n)

{

filename = paste0("problem\_id\_", i,".txt")

write.table(yy[i],file=filename,quote=FALSE,row.names=FALSE,col.names=FALSE)

}

}

pml\_write\_files(prediction\_20)