Arjan van Vuuren
02/09/2020

Practical Machine Learning Assignment

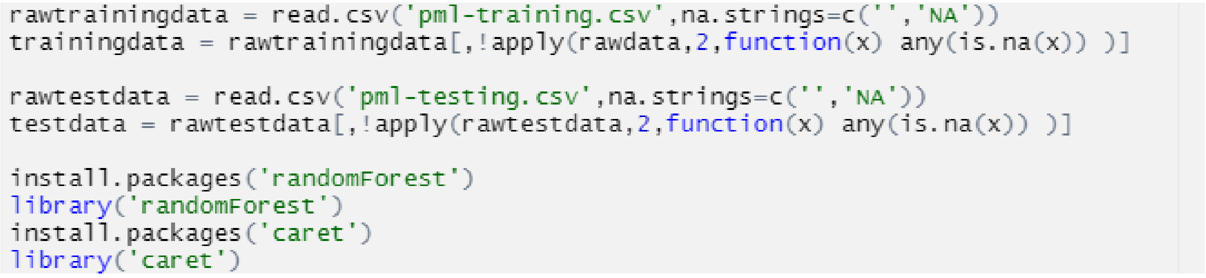
In this assignment the goal is to develop a model, using machine learning, to predict the type of exercise performed based on given data.

Data preparation

In this part, the data is extracted from the websites mentioned in the assignment:

The training data: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv
The testing data: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

The training data is used to develop the model and the testing data is used to make the predictions asked in the assignment.
Before creating the model, first the training and testing data will be extracted and R-packages are installed. Before using the data in our analyses, preparation have to be made. Random forest is used to make my model because it is one of the best performing, according to the lectures. In addition, caret is used for its diverse applications.

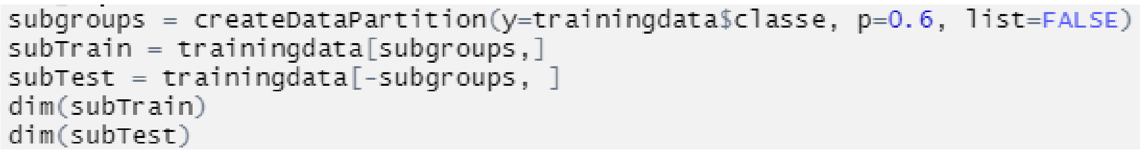


```
rawtrainingdata = read.csv('pml-training.csv',na.strings=c('','NA'))
trainingdata = rawtrainingdata[,!apply(rawdata,2,function(x) any(is.na(x)) )]

rawtestdata = read.csv('pml-testing.csv',na.strings=c('','NA'))
testdata = rawtestdata[,!apply(rawtestdata,2,function(x) any(is.na(x)) )]

install.packages('randomForest')
library('randomForest')
install.packages('caret')
library('caret')
```

For cross validation, the data is split into sub groups, with the ratio 60:40.

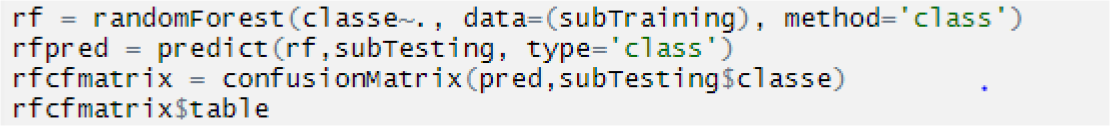


```
subgroups = createDataPartition(y=trainingdata$classe, p=0.6, list=FALSE)
subTrain = trainingdata[subgroups,]
subTest = trainingdata[-subgroups, ]
dim(subTrain)
dim(subTest)
```

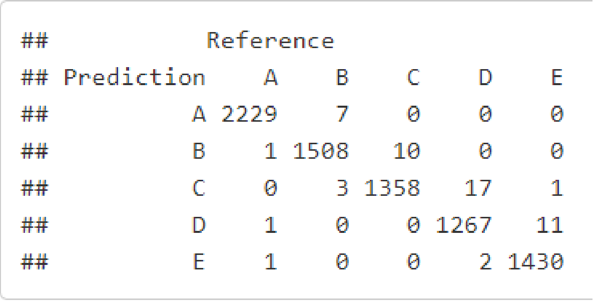
The amount of data in the training and test set are 11776 and 7846 respectively.

Model making

Hereafter, the random forest model is created using the training data. When the model is created, the test set will be used for cross validation. Therefore, a confusion matrix is made.



```
rf = randomForest(classe~., data=(subTraining), method='class')
rfpred = predict(rf,subTesting, type='class')
rfcfmatrix = confusionMatrix(pred,subTesting$classe)
rfcfmatrix$table
```

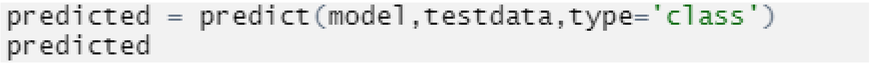


```
##            Reference
## Prediction    A    B    C    D    E
##          A 2229    7    0    0    0
##          B    1 1508   10    0    0
##          C    0    3 1358   17    1
##          D    1    0    0 1267   11
##          E    1    0    0    2 1430
```

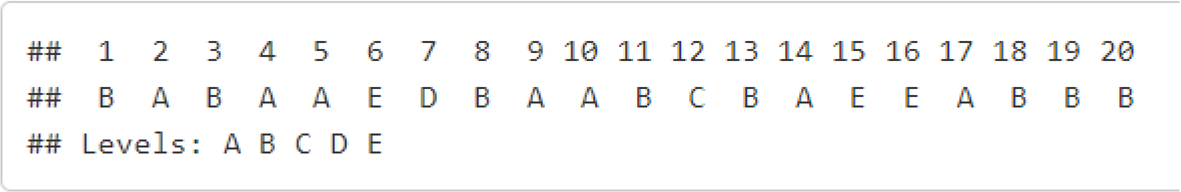
The overall accuracy is 99.31% and the corresponding out of sample error is only 0.69%, which is pretty decent.

Test set calculations

At last, the final predictions can be made using the test data and the answers to the assignment will be given.



```
predicted = predict(model,testdata,type='class')
predicted
```



```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 
##  B  A  B  A  A  E  D  B  A  A  B  C  B  A  E  E  A  B  B  B 
## Levels: A B C D E
```