Prediction Assignment Project

Diego Mora

6/13/2019

Summary

In this report we want to construct a model for predict the correct exceution of a particular excercise using the data of some sensors applied to 6 athletes. We are going to load and transform the data and then fit a random forest model.

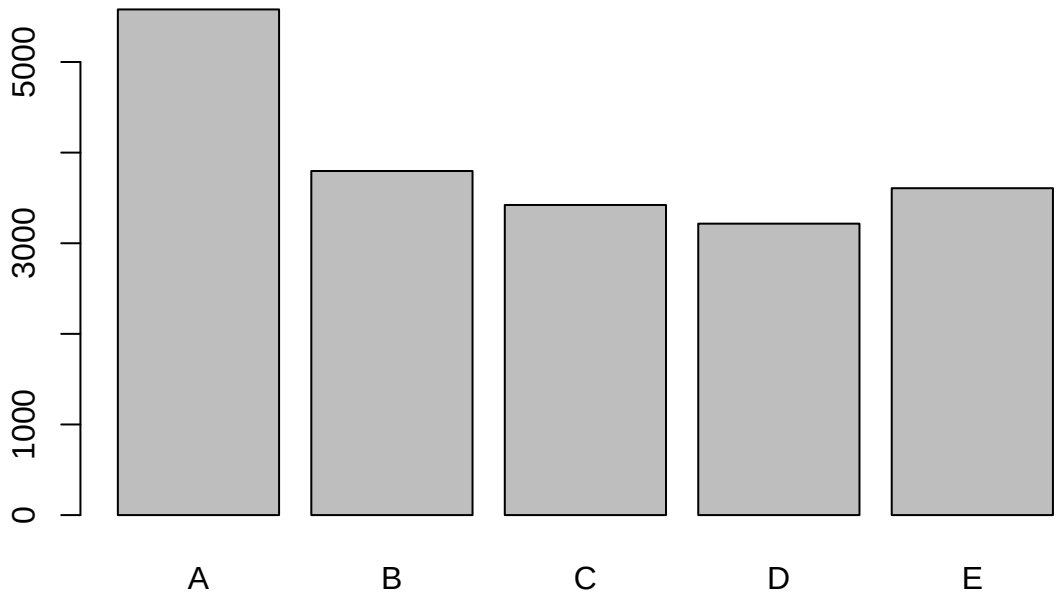
Load the Data

```
urltraining <- c("https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv")
urltesting <- c("https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv")

download.file(urltraining,paste(getwd(),"/training.csv",sep = ""))
download.file(urltesting,paste(getwd(),"/testing.csv",sep = ""))

training <- read.csv("training.csv", na.strings = c("NA","","#DIV/0!"))
testing <- read.csv("testing.csv", na.strings = c("NA","","#DIV/0!"))
```

```
plot(training$classe)
```



Transform

Select just the variables that have relevant data, it means the variables with less than 70% of NA values.

```
training <- training %>% select(which(colMeans(is.na(.)) < 0.7))
```

As we don't want that our model depends on a particular participant we can take off the name variable, the time stamps and the "x" variable that is nothing more than the index.

```
training <- training %>% select(6:60)
```

Prediction Model

Data Partition:

```
set.seed(1919)
inTrain <- createDataPartition(y=training$classe, p=0.7, list=FALSE)
trainingMod <- training[inTrain,]
testingMod <- training[-inTrain,]
```

First step is preprocessing the data using PCA and setting the threshold to find how many components we need. We want threshold or 95%, to find how many number of components are needed to capture the 95% of the variance. Note: we exclude the "classe" variable in the preprocess.

```
preProc <- preProcess(trainingMod[,1:54],method = "pca",thresh = 0.95)
preProc
```

```
## Created from 13737 samples and 54 variables
##
## Pre-processing:
##   - centered (53)
##   - ignored (1)
##   - principal component signal extraction (53)
##   - scaled (53)
##
## PCA needed 25 components to capture 95 percent of the variance
```

Now we can use our preprocessing to set a new training data with our Principal Components

```
trainPC <- predict(preProc, trainingMod[,1:54])
```

We can now fit a model, we want to use Random Forest

```
modelFit <- randomForest(trainingMod$classe~.,data = trainPC)
```

Our PCA for the testing set and then the predictions. We can meassure our model with the confusion matrix.

```
testPC <- predict(preProc,testingMod[,1:54])
pred <- predict(modelFit,testPC)
confusionMatrix(testingMod$classe,predict(modelFit,testPC))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1657   10    6    1    0
##          B   15 1100   20    0    4
##          C    2   20  989   14    1
##          D    1    0   39  921    3
##          E    0    2    6    3 1071
##
## Overall Statistics
##
##                Accuracy : 0.975
##                  95% CI : (0.9707, 0.9789)
##     No Information Rate : 0.2846
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9684
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
```

```
## Sensitivity            0.9893   0.9717   0.9330   0.9808   0.9926
## Specificity            0.9960   0.9918   0.9923   0.9913   0.9977
## Pos Pred Value         0.9898   0.9658   0.9639   0.9554   0.9898
## Neg Pred Value         0.9957   0.9933   0.9854   0.9963   0.9983
## Prevalence             0.2846   0.1924   0.1801   0.1596   0.1833
## Detection Rate         0.2816   0.1869   0.1681   0.1565   0.1820
## Detection Prevalence   0.2845   0.1935   0.1743   0.1638   0.1839
## Balanced Accuracy      0.9926   0.9818   0.9627   0.9861   0.9951
```

Conclusion

This random forest model has a 97.5% of Accuracy with a 95% confidence interval of 97% to 97.9%. we need 25 Principal Components to describe the 95% of the variance of the data.

References

The data used to this project was taken from:

Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) . Stuttgart, Germany: ACM SIGCHI, 2013.