Analysis Exercise Manner

C. Zhang

5/17/2020

Overview

The goal of this project is to predict the manner in which participants did the exercise. This is the "classe" variable in the datasets. At the end, I will predict 20 different test cases using my prediction model.

Loading and Exploring Data

Download training and testing data. Load the libraries we need in this project. Remove the rows of data which has 'NA' and the variables which are not required for the predictions from both the training and testing set.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
```

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

```
## [1] 19622   160
```

```
head(names(training))
```

```
## [1] "X"                    "user_name"            "raw_timestamp_part_1"
## [4] "raw_timestamp_part_2" "cvtd_timestamp"       "new_window"
```

```
grep('classe', names(training))
```

```
## [1] 160
```

```
head(training[, 160])
```

```
## [1] A A A A A A
## Levels: A B C D E
```

```
training = training[,!apply(training,2,function(x) any(is.na(x)) )]
training = training[,-c(1:7)]
```

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

```
## [1]  20 160
```

```
testing = testing[,!apply(testing,2,function(x) any(is.na(x)) )]
testing = testing[,-c(1:7)]
```

Training Data

For cross validation, we split the training data into two groups. The ratio is 60:40.

```
set.seed(336699)
t <- createDataPartition(y = training$classe, p = .6, list = FALSE)
trainingCross <- training[t,]
testingCross <- training[-t,]
dim(trainingCross)
```

```
## [1] 11776    53
```

```
dim(testingCross)
```

```
## [1] 7846   53
```

Firstly I training with the trainingCross data by using the random forest model. Then I predict the outcome with the testingCross data. At last, I examine the confusion matrix to see how well the predictive model performed

```
aTrain <- train(classe ~ ., data = trainingCross, method = 'rf')
aPred <- predict(aTrain, testingCross)
crossT <- confusionMatrix(aPred, testingCross$classe)
crossT$table
```

```
##           Reference
## Prediction    A    B    C    D    E
##          A 2229   15    0    0    0
##          B    1 1500    6    0    0
##          C    0    3 1358   22    2
##          D    0    0    4 1263    4
##          E    2    0    0    1 1436
```

```
crossT$overall[1]
```

```
##  Accuracy
## 0.9923528
```

From the result, we can find that the accuracy is 99.15% which means pretty good to predict classe.

Prediction

Now we use the model to predict the test data and show the prediction.

```
bPred = predict(aTrain, testing)
bPred
```

```
##  [1] 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
```

Conclusion

We build a prediction model which based on Random Forest and trained with the training data. The accuracy is 99.15% which is pretty good. At last, we use the model to predict the test data.