Practical Machine Learning

2160032\_Park\_Manjun

2017뀈 3썡 22씪

## Data Preprocessing

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(rpart)  
library(rpart.plot)  
library(randomForest)

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(corrplot)  
library(e1071)

### Download the Data

trainUrl <-"https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"  
testUrl <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"  
trainFile <- "./data/pml-training.csv"  
testFile <- "./data/pml-testing.csv"  
if (!file.exists("./data")) {  
 dir.create("./data")  
}  
if (!file.exists(trainFile)) {  
 download.file(trainUrl, destfile=trainFile, method="curl")  
}

## Warning: running command 'curl "https://d396qusza40orc.cloudfront.net/  
## predmachlearn/pml-training.csv" -o "./data/pml-training.csv"' had status  
## 127

## Warning in download.file(trainUrl, destfile = trainFile, method = "curl"):  
## download had nonzero exit status

if (!file.exists(testFile)) {  
 download.file(testUrl, destfile=testFile, method="curl")  
}

## Warning: running command 'curl "https://d396qusza40orc.cloudfront.net/  
## predmachlearn/pml-testing.csv" -o "./data/pml-testing.csv"' had status 127

## Warning in download.file(testUrl, destfile = testFile, method = "curl"):  
## download had nonzero exit status

### Read the Data

After downloading the data from the data source, we can read the two csv files into two data frames.

path <- getwd()  
setwd(path)  
trainRaw <- read.csv("./pml-training.csv")  
testRaw <- read.csv("./pml-testing.csv")  
dim(trainRaw)

## [1] 19622 160

## [1] 19622 160

dim(testRaw)

## [1] 20 160

## [1] 20 160

The training data set contains 19622 observations and 160 variables, while the testing data set contains 20 observations and 160 variables. The "classe" variable in the training set is the outcome to predict.

### Clean the data

In this step, we will clean the data and get rid of observations with missing values as well as some meaningless variables.

sum(complete.cases(trainRaw))

## [1] 406

## [1] 406

First, we remove columns that contain NA missing values.

trainRaw <- trainRaw[, colSums(is.na(trainRaw)) == 0]  
testRaw <- testRaw[, colSums(is.na(testRaw)) == 0]

Next, we get rid of some columns that do not contribute much to the accelerometer measurements.

classe <- trainRaw$classe  
trainRemove <- grepl("^X|timestamp|window", names(trainRaw))  
trainRaw <- trainRaw[, !trainRemove]  
trainCleaned <- trainRaw[, sapply(trainRaw, is.numeric)]  
trainCleaned$classe <- classe  
testRemove <- grepl("^X|timestamp|window", names(testRaw))  
testRaw <- testRaw[, !testRemove]  
testCleaned <- testRaw[, sapply(testRaw, is.numeric)]

Now, the cleaned training data set contains 19622 observations and 53 variables, while the testing data set contains 20 observations and 53 variables. The "classe" variable is still in the cleaned training set.

### Slice the data

Then, we can split the cleaned training set into a pure training data set (70%) and a validation data set (30%). We will use the validation data set to conduct cross validation in future steps.

set.seed(22519) # For reproducibile purpose  
inTrain <- createDataPartition(trainCleaned$classe, p=0.70, list=F)  
trainData <- trainCleaned[inTrain, ]  
testData <- trainCleaned[-inTrain, ]

## Data Modeling

We fit a predictive model for activity recognition using **Random Forest** algorithm because it automatically selects important variables and is robust to correlated covariates & outliers in general. We will use **5-fold cross validation** when applying the algorithm.

controlRf <- trainControl(method="cv", 5)  
modelRf <- train(classe ~ ., data=trainData, method="rf", trControl=controlRf, ntree=250)  
modelRf

## Random Forest   
##   
## 13737 samples  
## 52 predictor  
## 5 classes: 'A', 'B', 'C', 'D', 'E'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 10989, 10989, 10991, 10990, 10989   
## Resampling results across tuning parameters:  
##   
## mtry Accuracy Kappa   
## 2 0.9907550 0.9883040  
## 27 0.9914101 0.9891328  
## 52 0.9850038 0.9810259  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 27.

## Random Forest  
##  
## 13737 samples  
## 52 predictor  
## 5 classes: 'A', 'B', 'C', 'D', 'E'  
##  
## No pre-processing  
## Resampling: Cross-Validated (5 fold)  
##  
## Summary of sample sizes: 10989, 10989, 10991, 10990, 10989  
##  
## Resampling results across tuning parameters:  
##  
## mtry Accuracy Kappa Accuracy SD Kappa SD   
## 2 0.9904636 0.9879361 0.0006534224 0.0008263772  
## 27 0.9913374 0.9890405 0.0015731292 0.0019917940  
## 52 0.9850766 0.9811193 0.0027732182 0.0035098533  
##  
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 27.

Then, we estimate the performance of the model on the validation data set.

predictRf <- predict(modelRf, testData)  
confusionMatrix(testData$classe, predictRf)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction A B C D E  
## A 1672 1 0 0 1  
## B 7 1128 4 0 0  
## C 0 0 1021 5 0  
## D 0 0 13 950 1  
## E 0 0 1 6 1075  
##   
## Overall Statistics  
##   
## Accuracy : 0.9934   
## 95% CI : (0.991, 0.9953)  
## No Information Rate : 0.2853   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.9916   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D Class: E  
## Sensitivity 0.9958 0.9991 0.9827 0.9886 0.9981  
## Specificity 0.9995 0.9977 0.9990 0.9972 0.9985  
## Pos Pred Value 0.9988 0.9903 0.9951 0.9855 0.9935  
## Neg Pred Value 0.9983 0.9998 0.9963 0.9978 0.9996  
## Prevalence 0.2853 0.1918 0.1766 0.1633 0.1830  
## Detection Rate 0.2841 0.1917 0.1735 0.1614 0.1827  
## Detection Prevalence 0.2845 0.1935 0.1743 0.1638 0.1839  
## Balanced Accuracy 0.9977 0.9984 0.9908 0.9929 0.9983

## Confusion Matrix and Statistics  
##  
## Reference  
## Prediction A B C D E  
## A 1672 1 0 0 1  
## B 8 1127 4 0 0  
## C 0 1 1020 5 0  
## D 0 0 14 949 1  
## E 0 0 0 6 1076  
##  
## Overall Statistics  
##   
## Accuracy : 0.993   
## 95% CI : (0.9906, 0.995)  
## No Information Rate : 0.2855   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.9912   
## Mcnemar's Test P-Value : NA   
##  
## Statistics by Class:  
##  
## Class: A Class: B Class: C Class: D Class: E  
## Sensitivity 0.9952 0.9982 0.9827 0.9885 0.9981  
## Specificity 0.9995 0.9975 0.9988 0.9970 0.9988  
## Pos Pred Value 0.9988 0.9895 0.9942 0.9844 0.9945  
## Neg Pred Value 0.9981 0.9996 0.9963 0.9978 0.9996  
## Prevalence 0.2855 0.1918 0.1764 0.1631 0.1832  
## Detection Rate 0.2841 0.1915 0.1733 0.1613 0.1828  
## Detection Prevalence 0.2845 0.1935 0.1743 0.1638 0.1839  
## Balanced Accuracy 0.9974 0.9979 0.9907 0.9927 0.9984

accuracy <- postResample(predictRf, testData$classe)  
accuracy

## Accuracy Kappa   
## 0.9933730 0.9916171

## Accuracy Kappa  
## 0.9930331 0.9911870

oose <- 1 - as.numeric(confusionMatrix(testData$classe, predictRf)$overall[1])  
oose

## [1] 0.006627018

## [1] 0.006966865

So, the estimated accuracy of the model is 99.30% and the estimated out-of-sample error is 0.70%.

## Predicting for Test Data Set

Now, we apply the model to the original testing data set downloaded from the data source. We remove the problem\_id column first.

result <- predict(modelRf, testCleaned[, -length(names(testCleaned))])  
result

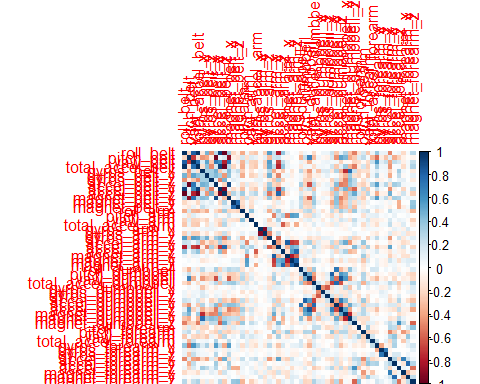
## [1] B A B A A E D B A A B C B A E E A B B B  
## Levels: A B C D E

## [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

## Appendix: Figures

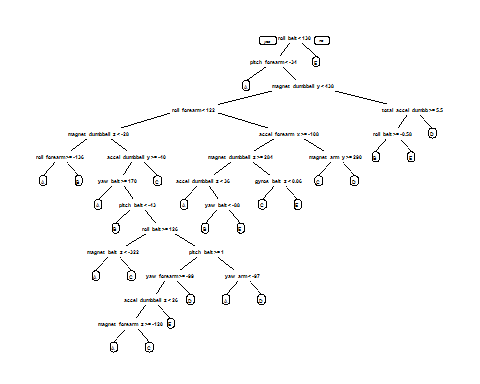
1. Correlation Matrix Visualization

corrPlot <- cor(trainData[, -length(names(trainData))])  
corrplot(corrPlot, method="color")



1. Decision Tree Visualization

treeModel <- rpart(classe ~ ., data=trainData, method="class")  
prp(treeModel) # fast plot



plot of chunk unnamed-chunk-13

plot of chunk unnamed-chunk-13