Practical Machine Learning Course Project

Li Xu

Summary

This report studies the personal activity data collected by devices such as Jawbone Up, Nike FeulBand and Fitbit. We analyze the activity data and try to predict the manner in which they did the exercise. Here, we use cross-validation method and random forest algorithm to train the training data set and predict the "classe" of the samples in the test data set.

Read Data Sets

The training and test data can be downloaded from the following website

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

By typing

```
TrainingData<-read.csv("pml-training.csv",sep=",",header=T)
TestData<-read.csv("pml-testing.csv",sep=",",header=T)
```

we read both data sets. The training data contains 19622 observations on 160 variables, and test data includes only 20 observations.

Tidy Data Sets

We observe that in some columns of the training data (e.g. max_roll_belt, max_picth_belt), most of the data are missing. These columns are meaningless and sometimes harmful for our prediction. Typing

```
TrainingData<-TrainingData[,colSums(is.na(TrainingData))<0.1*nrow(TrainingData)]
TrainingData<-na.omit(TrainingData)
TrainingData<-TrainingData[,-c(1:7)]
TrainingData<-TrainingData[,-nearZeroVar(TrainingData)]
```

we remove the columns with at least 90% missing data, and then remove all rows with missing data. Also we remove the first 7 columns, which are not irrelevant to predict. And we remove all features with near zero variance. We then finish tidying the data set.

Summary of "classe" in the Training Set

We aim to predict the value of "classe" in the test data. By typing

```
summary(TrainingData$classe)
```

we count the number of samples with each kind of "classe" as follows:

```
   A    B    C    D    E
5580 3797 3422 3216 3607
```

Use Cross Validation

We first split the training set into a training set (60%) and a validation set (40%) by typing

```
library(caret)
library(randomForest)
set.seed(12341234)
inTrain<-createDataPartition(TrainingData$classe,p=0.6,list=F)
TrainingSet<-TrainingData[inTrain,]
ValidationSet<-TrainingData[-inTrain,]
```

Build a Random Forest Model

Here we use the random forest algorithm, which is regarded as one of the best statistical learning algorithms with high accuracy. We fit a random forest model by

```
RFModel<-randomForest(classe~.,data=TrainingSet,ntree=1234,importance=T,proximity=T)
```

Out-of-Sample Prediction

We use this model to predict the validation set and check the confusion matrix by typing

```
predictions<-predict(RFModel,ValidationSet)
confusionMatrix(predictions,ValidationSet$classe)
```

It is obtained that the accuracy of the random forest model is **0.9957** with the 95% confidence interval **(0.9939, 0.997)**. And the out-of-sample error is 0.43%.

Prediction of the Test Data

We then use the model to predict the test data set with 20 samples by

```
predicttest<-predict(RFModel,TestingData)
```

and obtain the prediction results are

```
 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
```

Finally, we create one file for each submission by the following code

```r
answers <- as.character(predicttest)
pml_write_files = function(x){
        n = length(x)
        for(i in 1:n){
                filename = paste0("problem_id_",i,".txt")
                write.table(x[i],file=filename,quote=FALSE,row.names=FALSE,col.names=FALSE)
        }
}
pml_write_files(answers)
```

and then upload the files to Coursera.

References

[1] L. Breiman, “Random Forest”, *Machine Learning*, pp. 5-32, 2001.