Predicting How People Exercise

Victor Valdevite Pinto

Monday, July 20, 2015

Summary

The goal of this analysis is to predict the manner in which people did the exercise. This is the "classe" variable in the training set. This report contains how to read the data, subset it, apply cross validation and predict the Classe variable fot the test dataset.

Analysis

Data Preparation

Required Librarys

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.1.3
```

```
## Loading required package: lattice
## Loading required package: ggplot2
```

Reading the Data

```
training <- read.csv("C:\\Users\\Victor\\Box Sync\\Documentos\\Cursos\\Repo\\JH\\Course8_Week3_Project\\
```

Remove unused columns

As a cleanup strategy, we will remove the columns with more than 10% of NA values and columns that represents names, timestamps, dates and other variables that are not important to the prediction.

```
##remove the columns with more than 10% of NA data
training_ss <- training[ , sapply(training, function(x) !mean(is.na(x))>.1)]

##remove irrelevant columns from the data, like date, names, etc...
training_ss <- training_ss[,-c(1,2,3,4,5,6,7)]

training_ss <- training_ss[,c("classe", names(training_ss)[grepl("^(accel_|gyros_)", names(training_ss)
```

Cross Validation strategy

Since the training dataset is big enought, we can use cross validation, creating a new training (60% of the data) and a new testing (40% of the data) datasets. We will create these new datasets using the caret package.

```
intrain <- createDataPartition(y=training_ss$classe, p=0.6, list=FALSE)
cross_training <- training_ss[intrain,]
cross_testing <- training_ss[-intrain,]
```

Prediction Algorithm

The selected algorithm to train the model is Linear Discriminant Analysis in the caret package.

```
fit <- train(classe ~ ., method="lda", data=cross_training)
```

```
## Loading required package: MASS
```

Out of sample error

We will estimate the Out Of Sample Error by predicting the Testing dataset created with the croos validation and taking the 1 - Accuracy. The summary of the prediction is:

```
cm <- confusionMatrix(predict(fit, cross_testing), cross_testing$classe)
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1520  406  668  203  246
##          B  132  712  111  141  250
##          C  186  212  425  124   79
##          D  352  108  150  705  170
##          E   42   80   14  113  697
##
## Overall Statistics
##
##                Accuracy : 0.5173
##                  95% CI : (0.5062, 0.5284)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3816
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.6810  0.46904  0.31067  0.54821  0.48336
## Specificity            0.7287  0.89981  0.90722  0.88110  0.96112
## Pos Pred Value         0.4995  0.52897  0.41423  0.47475  0.73679
```

```
## Neg Pred Value         0.8518  0.87600  0.86173  0.90866  0.89203
## Prevalence             0.2845  0.19347  0.17436  0.16391  0.18379
## Detection Rate         0.1937  0.09075  0.05417  0.08985  0.08884
## Detection Prevalence   0.3878  0.17155  0.13077  0.18927  0.12057
## Balanced Accuracy      0.7049  0.68442  0.60895  0.71465  0.72224
```

The estimated OOSE is: 0.4826663