Machine_Learning

Aaron Chandler

Sunday, September 27, 2015

```
## Warning: package 'randomForest' was built under R version 3.2.2
```

```
## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.
```

Model Estimation

This analysis predicts the “classe” variable from the Weight Lifting Exercises Dataset from the Human Activity Recognition project.

The final model is a random forest model that uses all variables in the dataset that were empirically collected through the motion sensors. It is specified as: classe ~ `r vars`

The reasoning for selecting these variables is that misperformance of exercise can lead to a wide variety of extraneous motion. Specific misperformance more than likely yields specific extraneous motion. Anticipating the wide variety of specific range for each of the 5 misperformances, as well as the efficient motion of the correct exercise performance would be difficult. This difficulty of qualitatively assessing activity performance is described in the write-up accompanying the dataset. Selecting all of the empirical variable provides the model with as much information as possible to identify the specific patterns associated with each classe.

Final Model The estimated model and in the in sample performance are below:

```
modFit
```

```
## 
## Call:
##  randomForest(formula = classe ~ ., data = train5, mrty = 6) 
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 6
## 
##         OOB estimate of  error rate: 0.36%
## Confusion matrix:
##      A    B    C    D    E  class.error
## A 5575    4    0    0    1 0.0008960573
## B   10 3778    9    0    0 0.0050039505
## C    0   10 3407    5    0 0.0043834015
## D    0    0   21 3193    2 0.0071517413
## E    0    0    3    5 3599 0.0022179096
```

Cross-Validation

K-Folds validation using 3 folds was used to cross validate the model. The training dataset was randomly divided into 3 subsets. Three additional model estimations were done excluding one of the three subsets to be used for testing. The results for each validation are below. The results are displayed in proportion tables.

```
## [1] 0.9914156
```

```
## [1] 0.9919536
```

```
## [1] 0.9933354
```

Cross Validation With Subset 1

```
prop1
```

```
##    pred1
##                A            B            C            D            E
##   A 0.9951923077 0.0041017227 0.0017777778 0.0000000000 0.0016207455
##   B 0.0048076923 0.9893355209 0.0044444444 0.0000000000 0.0000000000
##   C 0.0000000000 0.0065627564 0.9813333333 0.0027223230 0.0000000000
##   D 0.0000000000 0.0000000000 0.0106666667 0.9936479129 0.0008103728
##   E 0.0000000000 0.0000000000 0.0017777778 0.0036297641 0.9975688817
```

The mean error rate for this subset is `r 1-mean(diag(prop.table(testtable1,2)))` .

Cross Validation With Subset 2

```
prop2
```

```
##    pred2
##                A            B            C            D            E
##   A 0.9947340706 0.0007686395 0.0008424600 0.0000000000 0.0000000000
##   B 0.0052659294 0.9923136049 0.0050547599 0.0000000000 0.0000000000
##   C 0.0000000000 0.0069177556 0.9806234204 0.0030241935 0.0000000000
##   D 0.0000000000 0.0000000000 0.0134793597 0.9929435484 0.0008467401
##   E 0.0000000000 0.0000000000 0.0000000000 0.0040322581 0.9991532599
```

The mean error rate for this subset is `r 1-mean(diag(prop.table(testtable2,2)))` .

Cross Validation With Subset 3

```
prop3
```

```
##    pred3
##                A            B            C            D            E
##   A 0.9983516484 0.0000000000 0.0000000000 0.0000000000 0.0000000000
##   B 0.0016483516 0.9944794953 0.0052219321 0.0000000000 0.0000000000
##   C 0.0000000000 0.0055205047 0.9773716275 0.0009182736 0.0000000000
##   D 0.0000000000 0.0000000000 0.0147954743 0.9981634527 0.0016891892
##   E 0.0000000000 0.0000000000 0.0026109661 0.0009182736 0.9983108108
```

The mean error rate for this subset is `r 1-mean(diag(prop.table(testtable3,2)))`.

Expected Error for Out-of-Sample Tests The expected error rate for this model is .711%, which is the mean of the error rates for each of the subset used for testing during the cross-validation. The range for the error rate is expected to be between .57% and .8%, which are the upper and lower bounds of the error rates of the cross validation results.

Prediction Using the Model The final task of this assignment was to predict the classifier of 20 test cases. This model resulted in 100% accuracy for the test cases.

```
url2<-"https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"

final<-read.csv(url2,header=TRUE,stringsAsFactors=FALSE,sep=",")
final2<-final[vars]
final2$sample<-1

fpred<-predict(modFit,newdata=final2)
```