Practical Machine Learning

Rehab Fathi

April 16, 2017

Background

Using devices such as Jawbone Up, Nike FuelBand, and Fitbit it is now possible to collect a large amount of data about personal activity relatively inexpensively. These type of devices are part of the quantified self movement - a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. One thing that people regularly do is quantify how much of a particular activity they do, but they rarely quantify how well they do it. In this project, your goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways.

Data Preprocessing

Remove columns with missing values

The data has 160 variables and it's hard to do any data exploration so I'll remove all columns with NAs

```r
train <- read.csv("pml-training.csv", na.strings = c("NA", ""))
traincomplete <- train[,colSums(is.na(train)) == 0]
dim(traincomplete)
```

```
## [1] 19622    60
```

The new dataset contains 19622 observations of 60 variables

Remove almost constant predictors

```r
library(caret)
```

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

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

```r
library(lattice)
library(ggplot2)
nzv <- nearZeroVar(traincomplete)
trainvar <- traincomplete[,-nzv]
dim(trainvar)
```

```
## [1] 19622    59
```

Only one predictor has near zero variance and it is removed

Remove id variable

```r
trainvar <- trainvar[,-1]
```

Constructing trees with caret package

I will use the caret package to construct decision trees to predict the classe variable. The algorithm will use cross validation with 10 folds to minimize the out of sample error

```
fitControl <- trainControl(method = "repeatedcv",
                           number = 10,
                           repeats = 10,
                           classProbs = TRUE)
model<-train(classe~.,
             data=trainvar,
             method="rpart",
             tuneLength=20,
             trControl = fitControl)
```

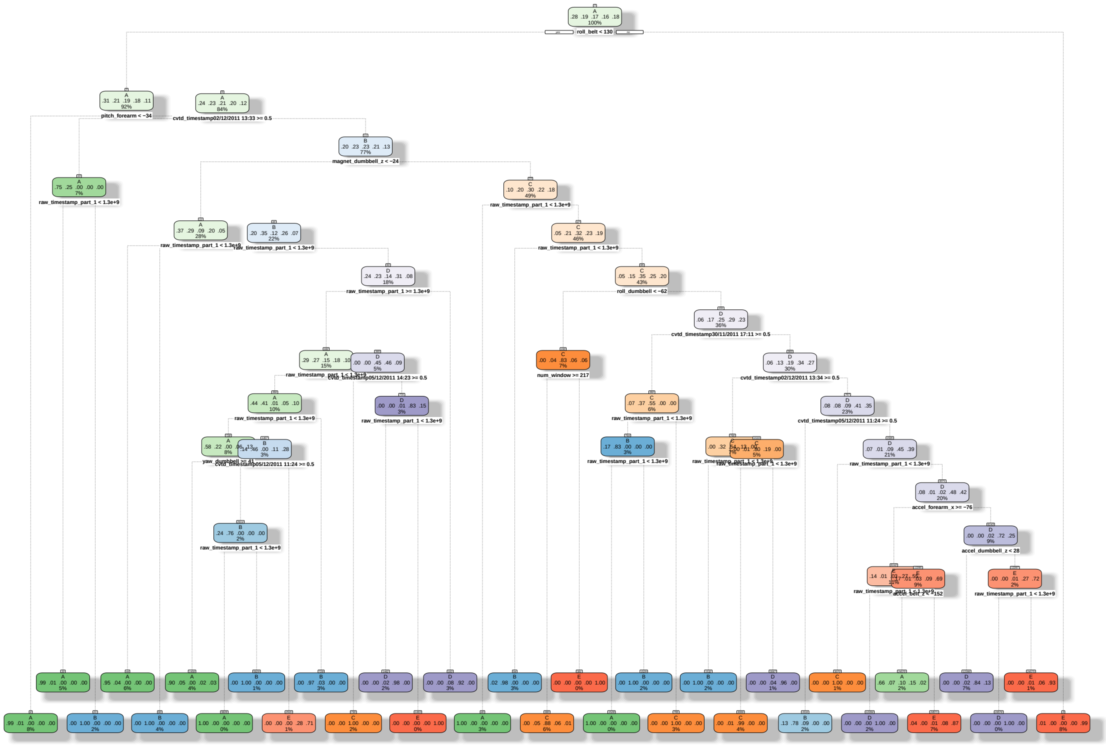
```
## Loading required package: rpart
```

```
model
```

```
## CART
##
## 19622 samples
##    57 predictor
##     5 classes: 'A', 'B', 'C', 'D', 'E'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 17661, 17660, 17661, 17659, 17660, 17659, ...
## Resampling results across tuning parameters:
##
##   cp           Accuracy   Kappa
##   0.004557755  0.9424161  0.92711848
##   0.005056260  0.9386752  0.92237975
##   0.005982054  0.9290633  0.91023552
##   0.006053269  0.9276976  0.90851228
##   0.006124484  0.9266783  0.90722794
##   0.006409343  0.9224534  0.90191268
##   0.011608033  0.8899583  0.86092313
##   0.012462612  0.8795973  0.84784582
##   0.012818687  0.8746231  0.84158494
##   0.014171770  0.8607407  0.82388608
##   0.016308218  0.8507310  0.81138728
##   0.023358496  0.8130732  0.76366447
##   0.027939990  0.7439605  0.67664950
##   0.028023074  0.7350459  0.66552848
##   0.030693633  0.6387560  0.54554019
##   0.030978493  0.6248118  0.52801397
##   0.031441390  0.6110927  0.51071494
##   0.040236434  0.4921298  0.34432996
##   0.045672506  0.4154802  0.21609781
##   0.115154536  0.3213255  0.05630782
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was cp = 0.004557755.
```

```r
rattle::fancyRpartPlot(model$finalModel)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



yes
no
roll_belt < 130
pitch_forearm < −34
cvtd_timestamp02/12/2011 13:33 >= 0.5
magnet_dumbbell_z < −24
raw_timestamp_part_1 < 1.3e+9
raw_timestamp_part_1 >= 1.3e+9
cvtd_timestamp05/12/2011 14:23 >= 0.5
yaw_dumbbell >= 41
cvtd_timestamp05/12/2011 11:24 >= 0.5
roll_dumbbell < −62
num_window >= 217
cvtd_timestamp30/11/2011 17:11 >= 0.5
cvtd_timestamp02/12/2011 13:34 >= 0.5
accel_forearm_x >= −76
accel_dumbbell_z < 28
accel_belt_z < −152

Rattle 2017−Apr−16 17:56:26 HP

Preprocess testing set

```r
test <- read.csv("pml-testing.csv")
test <- test[,colSums(is.na(train)) == 0]
test <- test[,-nzv]
```

Predict on test values

```r
pred <- predict( model, newdata = test)
pred
```

```
##  [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
```