

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 dumbbell 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

```
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

```
library(caret)
```

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

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

```
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

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

Constructing trees with caret package

```
fitControl <- trainControl(method = "repeatedcv",
                           number = 10,
                           repeats = 10,
                           classProbs = TRUE)

model<-train(classe~.,
             data=trainvar,
             method="rpart",
             tuneLength=20,
             trControl = fitControl)
```

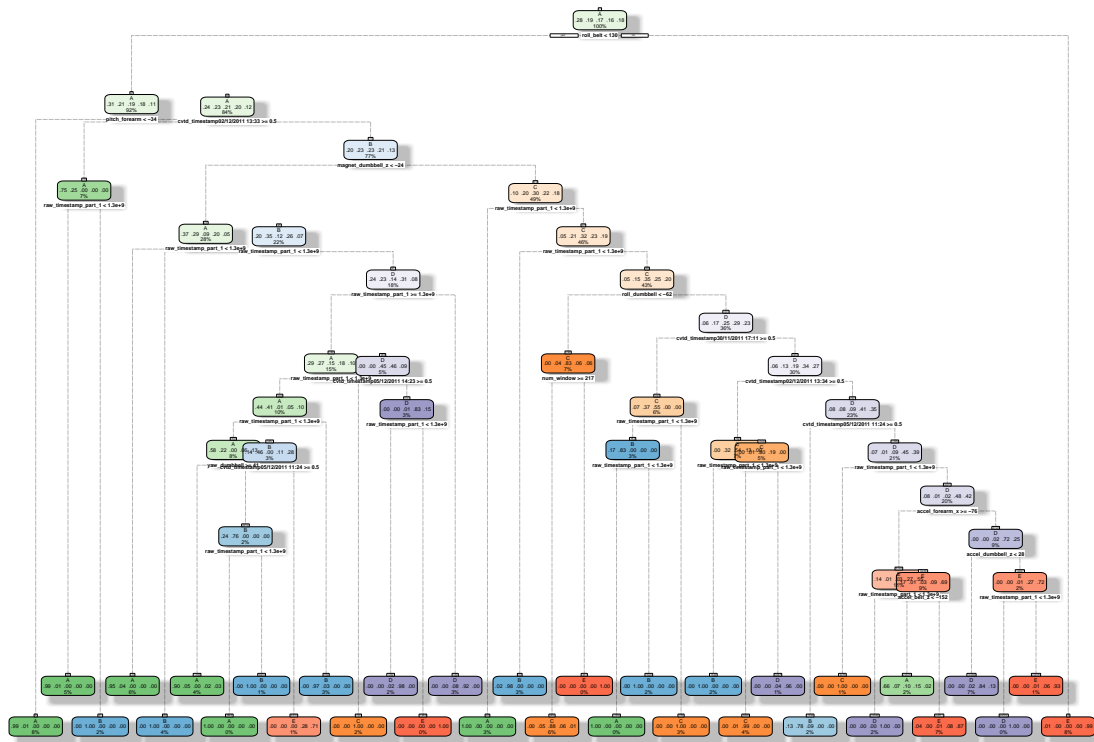
```
## Loading required package: rpart
```

```
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: 17659, 17660, 17660, 17661, 17659, 17660, ...
## Resampling results across tuning parameters:
##
##    cp          Accuracy    Kappa
##  0.004557755  0.9439719  0.92909941
##  0.005056260  0.9405014  0.92470989
##  0.005982054  0.9332802  0.91558818
##  0.006053269  0.9320365  0.91401912
##  0.006124484  0.9301968  0.91169870
##  0.006409343  0.9259055  0.90629029
##  0.011608033  0.8961133  0.86873094
##  0.012462612  0.8869308  0.85716906
##  0.012818687  0.8816351  0.85047263
##  0.014171770  0.8658319  0.83032442
##  0.016308218  0.8555681  0.81739648
##  0.023358496  0.8228063  0.77581435
##  0.027939990  0.7524549  0.68722511
##  0.028023074  0.7365985  0.66748290
##  0.030693633  0.6407289  0.54787586
##  0.030978493  0.6312307  0.53599288
##  0.031441390  0.6076203  0.50605028
##  0.040236434  0.5041084  0.36236624
##  0.045672506  0.4220457  0.22584255
##  0.115154536  0.3252014  0.06224181
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.004557755.
```

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

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



Rattle 2017-Apr-16 17:35:43 HP

Preprocess testing set

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

Predict on test values

```
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
```