Peer Assessment for Practical Machine Learning - Data Science Specialization

Michael Karp

1/22/2015

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. More information is available from the website here: http://groupware.les.inf.puc-rio.br/har (http://groupware.les.inf.puc-rio.br/har) (see the section on the Weight Lifting Exercise Dataset). The data for this project come from this source: http://groupware.les.inf.puc-rio.br/har (http://groupware.les.inf.puc-rio.br/har).

```
# The goal of your project is to predict the manner in which they did the exercise. T
his is the "classe" variable in the
# training set. You may use any of the other variables to predict with. You should cr
eate a report describing how you
# built your model, how you used cross validation, what you think the expected out of
sample error is, and why you made
# the choices you did. You will also use your prediction model to predict 20 differen
t test cases.

# I built the model for predition of exercise class using a 10-fold cross validation
with a decision tree being the
# underlying model. I initially split the trained the model on a decision tree to det
ermine the which variables would be
# useful for prediction. It was only splitting on the X variable so I had to disregar
d this feature in particular
# because it is irrelevant to the problem. Ultimatly in evaluating the cross validate
d initial tree the first and most
# important split was on roll_belt's standard deviation and since roll belt was not N
A in either dataset. Continued
# down this same line of thought and used total_accel_belt and pitch_belt features as
well

# always set seed to make research reproducible
set.seed(1000)

# Load training set
train <- read.csv("pml-training.csv")
# Load testing set
test <- read.csv("pml-testing.csv")

# inspect dataset
# str(train)
# str(test)

# summary(train)
# use simple decision tree to pick out important variables
library(rpart)
# Cross-validation
# Load libraries for cross-validation
library(caret)
```

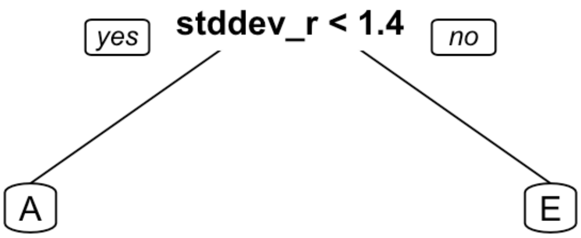
```
## Loading required package: lattice
## Loading required package: ggplot2
```

```
library(MASS)
library(e1071)

# Number of folds
tr.control = trainControl(method = "cv", number = 10)

# cp values
cp.grid = expand.grid( .cp = 0.2)

# cross validation
exercise_cv_cart = train(classe ~ .-X, data = train, method = "rpart", metric = "Accu
racy", trControl = tr.control, tuneGrid = cp.grid)
# $ Accuracy  : num 0.993 probably over fit
# Extract tree
best.tree = exercise_cv_cart$finalModel
library(rpart.plot)
# extract important features using the tree
prp(best.tree)
```

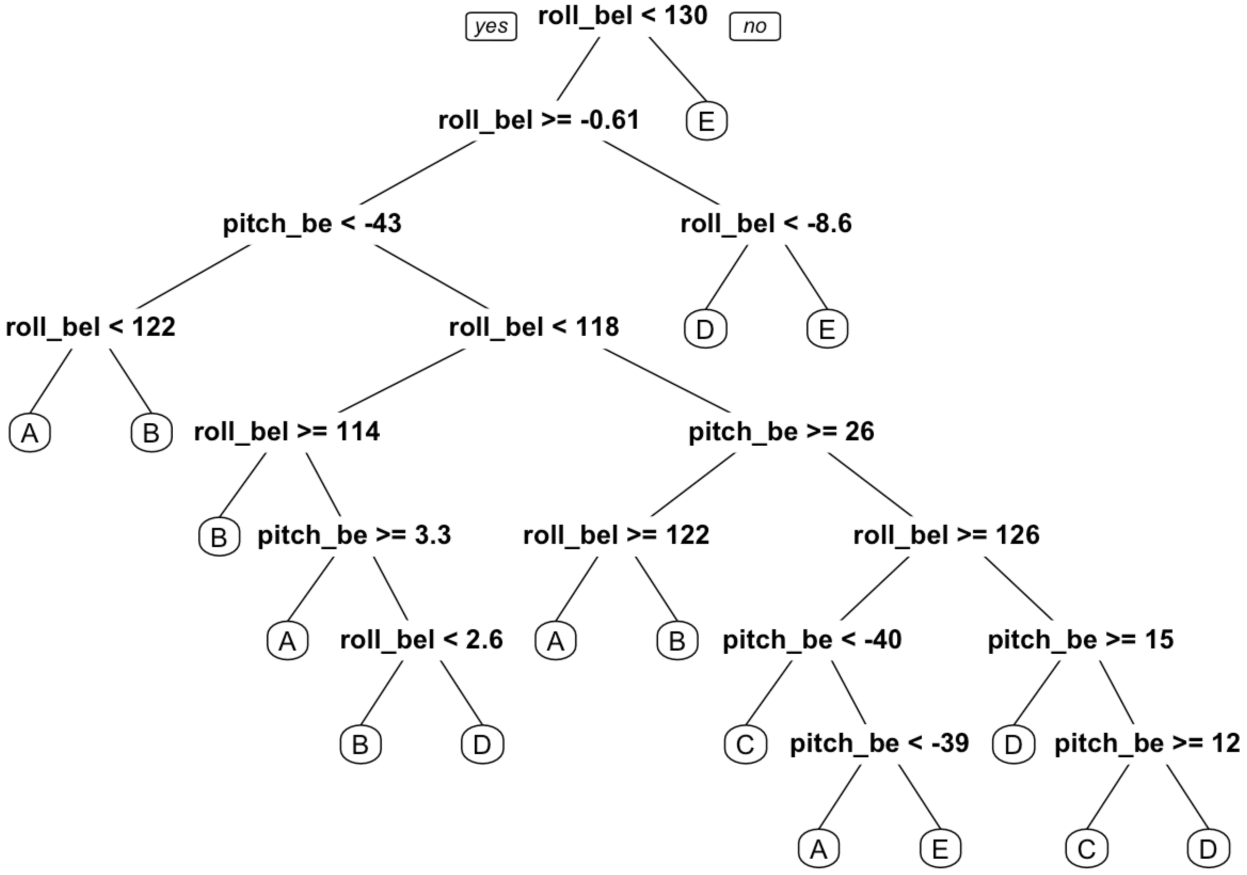


```
best.tree$variable.importance[1:10]
```

```
stddev_roll_belt        var_roll_belt var_total_accel_belt
        47.46829             47.46829             41.05366
```

amplitude_pitch_belt avg_roll_belt amplitude_roll_belt 39.12927 26.94146 25.65854 NA NA NA NA

```
# cross validation - rpart
exercise_cart = rpart(classe ~roll_belt+total_accel_belt+pitch_belt, data = train, me
thod = "class", xval = 10)
prp(exercise_cart)
```



```
predictTest <- predict(exercise_cart, type = "class", newdata = test)
answers <- as.character(predictTest)
answers
```

[1] “A” “A” “B” “D” “A” “E” “A” “A” “A” “A” “A” “A” “B” “A” “D” “A” “A” [18] “A” “A” “B”

```
# 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)
```