Peer graded assignment

Code

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://web.archive.org /web/20161224072740/http:/groupware.les.inf.puc-rio.br/har (http://web.archive.org /web/20161224072740/http:/groupware.les.inf.puc-rio.br/har) (see the section on the Weight Lifting Exercise Dataset).

Requirement

The goal of your project is to predict the manner in which they did the exercise. This is the “classe” variable in the training set. You may use any of the other variables to predict with. You should create 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 different test cases.

Packages and seed

Hide

```
library(doParallel)
```

```
Error in library(doParallel) : there is no package called âdoParallelâ
```

Read in the data

Investigating NAs - finding the right percentage

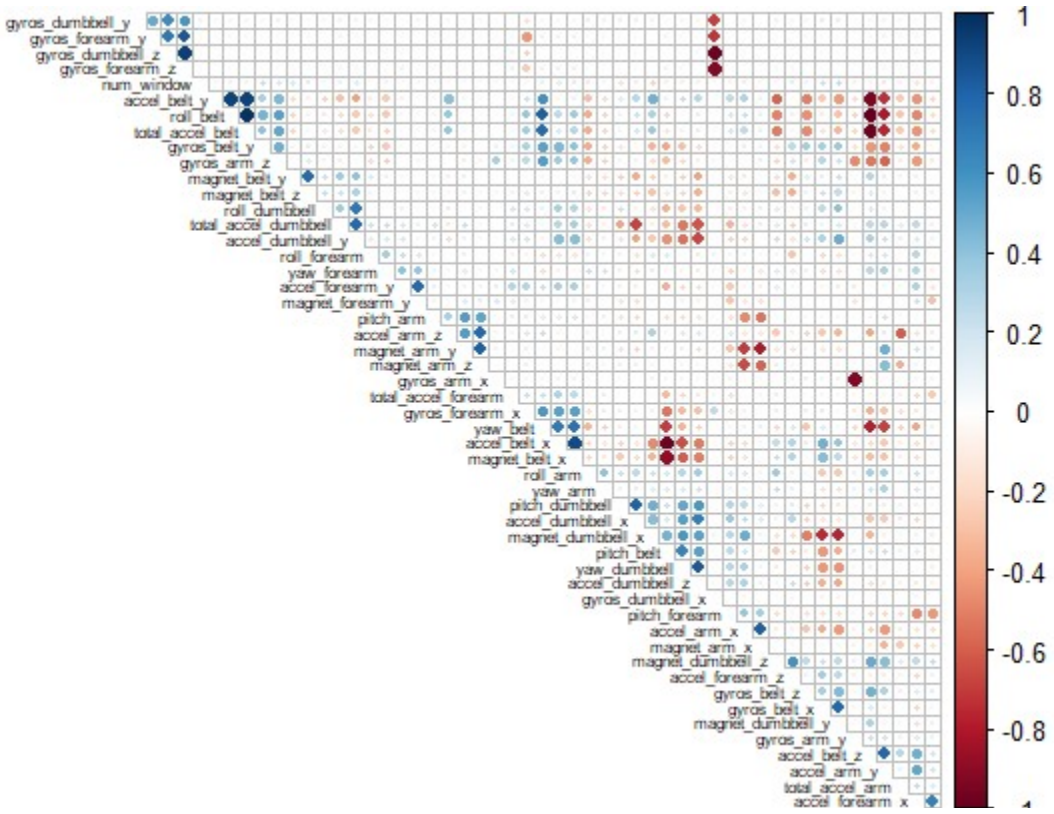
Hide

```
sum(sapply(training, function(x) sum(is.na(x))/length(x))>0.9)
```

```
[1] 100
```

Exclude 100 variables with more than 90% NAs as well as the first few

Test for highly correlated variables



Columns with 75% or more correlation - to be removed

Hide

```
length(highcorrcolms) # number of columns which should be removed
```

```
[1] 20
```

Modelling

Random forest

Hide

```
fit_rf
```

```
Random Forest

19622 samples
   33 predictor
    5 classes: 'A', 'B', 'C', 'D', 'E'

No pre-processing
Resampling: Cross-Validated (2 fold)
Summary of sample sizes: 9810, 9812
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.9933239  0.9915546
  17    0.9967894  0.9959387
  33    0.9941901  0.9926514

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 17.
```

Decision tree

Hide

```
fit_rpart
```

```
CART

19622 samples
   33 predictor
    5 classes: 'A', 'B', 'C', 'D', 'E'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 15697, 15698, 15697, 15699, 15697
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.02770261  0.5759876  0.4630511
  0.03945307  0.5474489  0.4236320
  0.04226606  0.4308715  0.2363863

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.02770261.
```

Using the RF model since the accuracy is better

My final predictions

Hide

```
pred_rf
```

```
 [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
```