Practical Machine Learning Project Analysis

Jiatong Li

Choosing the predictors:

When I looked at summary(data), I found that many variables contained too much blank values, NAs,or #DIV/0!s. For example:

```
min_yaw_forearm     amplitude_roll_forearm amplitude_yaw_forearm
            :19216      Min.   :   0.000                      :19216
#DIV/0!:      84      1st Qu.:   1.125                #DIV/0!:      84
-1.2   :      32      Median :  17.770                 0.00   :    322
-1.3   :      31      Mean   :  24.653
-1.4   :      24      3rd Qu.:  39.875
-1.5   :      24      Max.   : 126.000
(Other):     211      NA's   :19216
```

From the first variable X, simply a sequence vector, we can know there are 19622 rows. We can also get 19622 by adding the numbers in the first or third variable I just listed above. As 19216/19622=0.9793, those blank or NA takes up at least 98% of the rows. Even knnImpute may not properly handle them! So I discarded those columns and used the columns with all effective numbers to train and predict. The summary of those useful columns look like:

```
total_accel_forearm
Min.   :  0.00
1st Qu.: 29.00
Median : 36.00
Mean   : 34.72
3rd Qu.: 41.00
Max.   :108.00
```

I also discarded the first 6 columns. The first column, X is just the sequence 1,2…19622 which cannot contribute to predicting the classe. Everyone was asked to perform barbell lifts correctly and incorrectly in 5 different ways, so user_name also has nothing to do with classe. So did the time stamps. Since new_window has about 98% “no”, it may contribute nothing to our prediction. So the final columns I selected were c(7:11,37:49,60:68,84:86,102,113:124,140,151:160). Of course, I have to include the final column, the classe.

The model:

The classe is a factor variable. So classification trees is a good choice. Besides, random forests is one of the two top performing algorithms. As a result I chose random forests.

Cross validation:

I split the data into training and testing part, with p=0.7. After training with the training part, I predicted on the testing part and got a table:

```
x          A       B       C       D       E
   A 1674       5       0       0       0
   B       0 1133       6       0       0
   C       0       1 1020      11       0
   D       0       0       0     953       3
```

E 0 0 0 0 1079

There are total 5885 samples in the testing part, and 5859 of them are classified wright and 26 of them are wrong. So I expect the out of sample error to be 26/5885=0.44%, or the accuracy is 99.56%. It's a high accuracy rate.

I predicted on the given 20 test cases. It gave:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

B A B A A E D B A A B C B A E E A B B B

I submitted and all of them were right.

In a word, the key to a high accuracy rate is choosing the useful columns. I chose them manually, a little bit silly but effective. I guess I can use functions to choose them next time…