Practical Machine Learninng - Project

Gopala Krishna

Sunday, May 24, 2015

Background

Dependent Variables

- Class A: exactly according to specification
- Class B: throwing the elbows to the front
- Class C: lifting the dumbbell only halfway
- Class D: lowering the dumbbell only halfway,
- Class E: throwing the hips to the front.

Predictor Variables

For data recording, the authors used 4 inertial measurement units (IMU), which provide three-axes: acceleration, gyroscope and magnetometer data. The sensors are mounted in the users' glove, armbarnd, lumbar belt, and dumdbell. The author also used "a sliding window approach with diifferent lengths from 0.5 second to 2.5 seconds, with 0.5 second overlap. In each step of the sliding window approach they calculated features on the features on the Euler angles (roll, pitch and yaw), as well as the raw accelerometer, gyroscope and magnetometer readings. For the Euler angles of each of the 4 sensors they calculated 8 features: mean,variance, standard deviation, max, min, amplitude, kurtosis and skewness. (Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) . Stuttgart, Germany: ACM SIGCHI, 2013).

Data Preprocessing

Removing variables with too many NA's

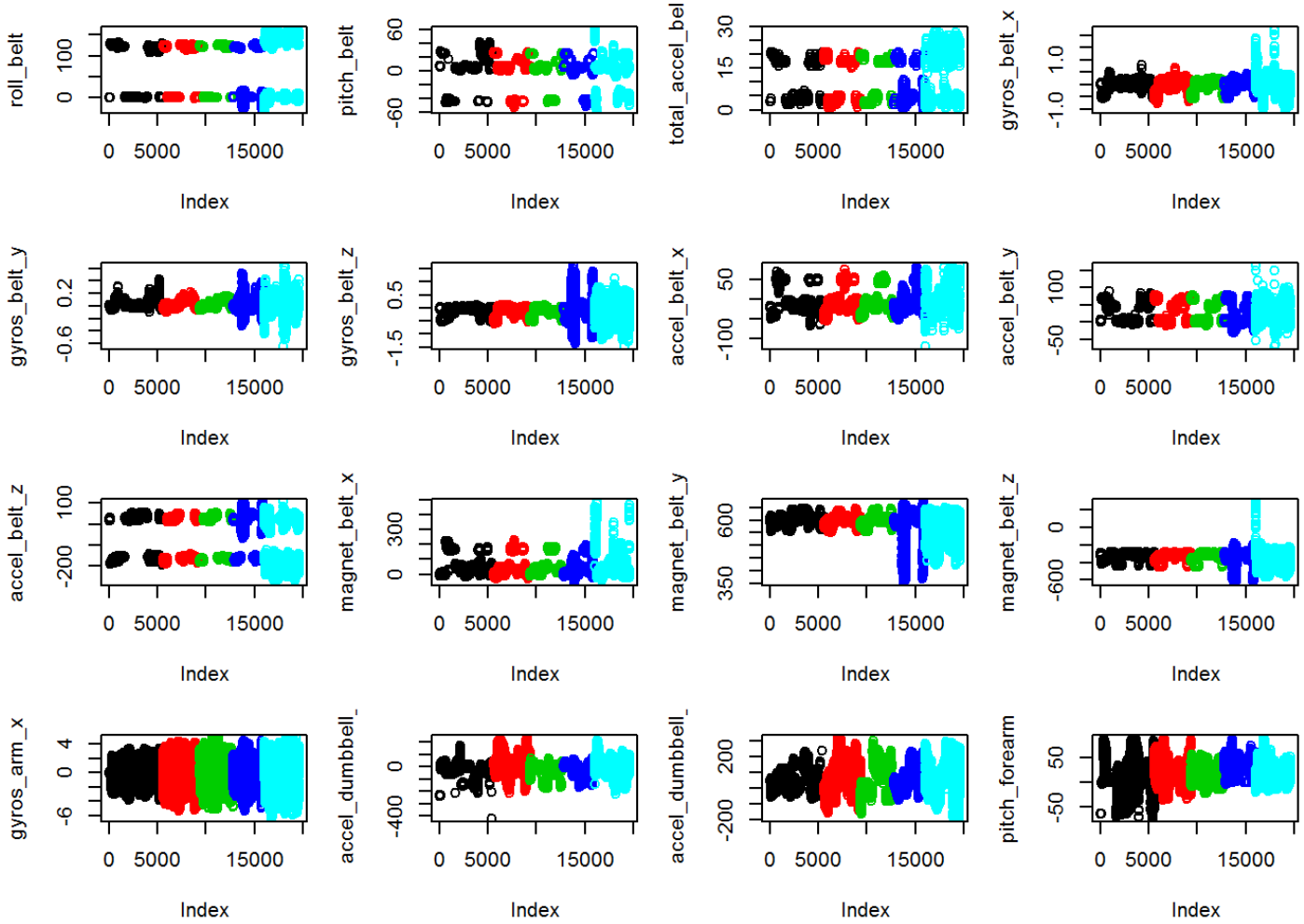
The table below shows that in the training set, out of 160 variables, there are 67 variables that have 97.9% of missing values.

```
##
##     0 19216
##    93    67
```

Therefore, we only keep 93 variables that have enough data points for data exploration and prediction.

Ploting potential predictors

With the remaining 93 variables, we will visualise each variable's possible correlation with the dependent variable (viz., classe) by ploting them against the index, and coloured by the five "classe" type. We then choose the 16 variables that show the most variations accross 5 classes (i.e., A, B, C, D, E).



Modelling

We choose the the Bagged CART method (i.e., method = "treebag") for the classification problem. This is because its computation advantage against the ramdom forest. Importantly, the treebag method implement a cross-validation process.

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

In particular, there were 25 resampling bootstraped perform with the Accuracy of 0.941712, Kappa of 0.926323.

```
## Bagged CART
##
## 19622 samples
##    16 predictor
##     5 classes: 'A', 'B', 'C', 'D', 'E'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
##
## Summary of sample sizes: 19622, 19622, 19622, 19622, 19622, 19622, ...
##
## Resampling results
##
##   Accuracy  Kappa      Accuracy SD  Kappa SD
##   0.942708  0.9275843  0.004066217  0.005113137
##
##
```

Prediction

Finally, the fitted Bagged CART model is used to predict the classificaiton for the 20 cases in the test set. When submitting in the submission part ofthe project, it achieved 95% accuracy (i.e., 19 out of 20). Clearly, the cross-validation process in the Bagged CART model has reduced the overfitting of the CART model; thus, increase the accuracy in the testing set. The predicted classification for each problem_id is as follow:

```
##    problem_id predict
## 1           1       B
## 2           2       C
## 3           3       B
## 4           4       A
## 5           5       A
## 6           6       E
## 7           7       D
## 8           8       B
## 9           9       A
## 10         10       A
## 11         11       B
## 12         12       C
## 13         13       B
## 14         14       A
## 15         15       E
## 16         16       E
## 17         17       A
## 18         18       B
## 19         19       B
## 20         20       B
```