Human activity recognition project

Heinz Lugo.

21 February 2015

Synopsis

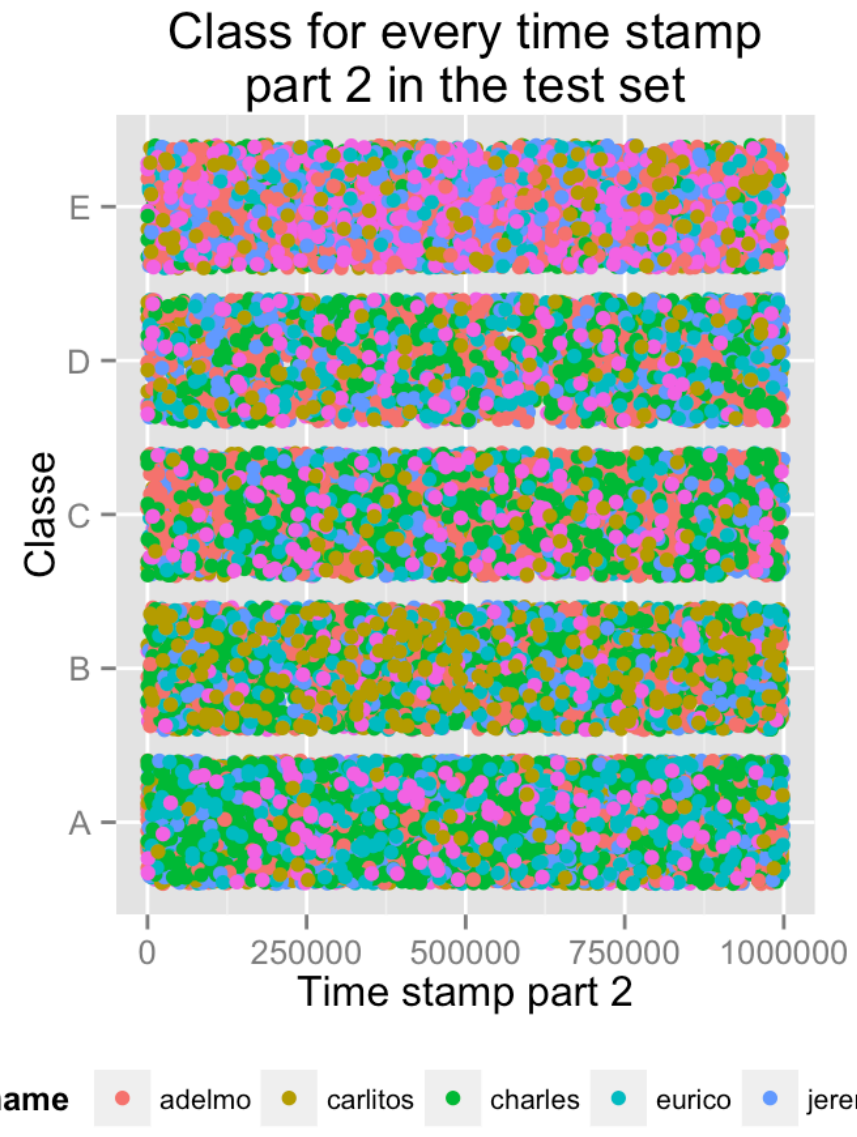
This document describes the methodology followed to develop a machine learning algorithm based on the weightlifting dataset made available by Velloso et al. (2013).

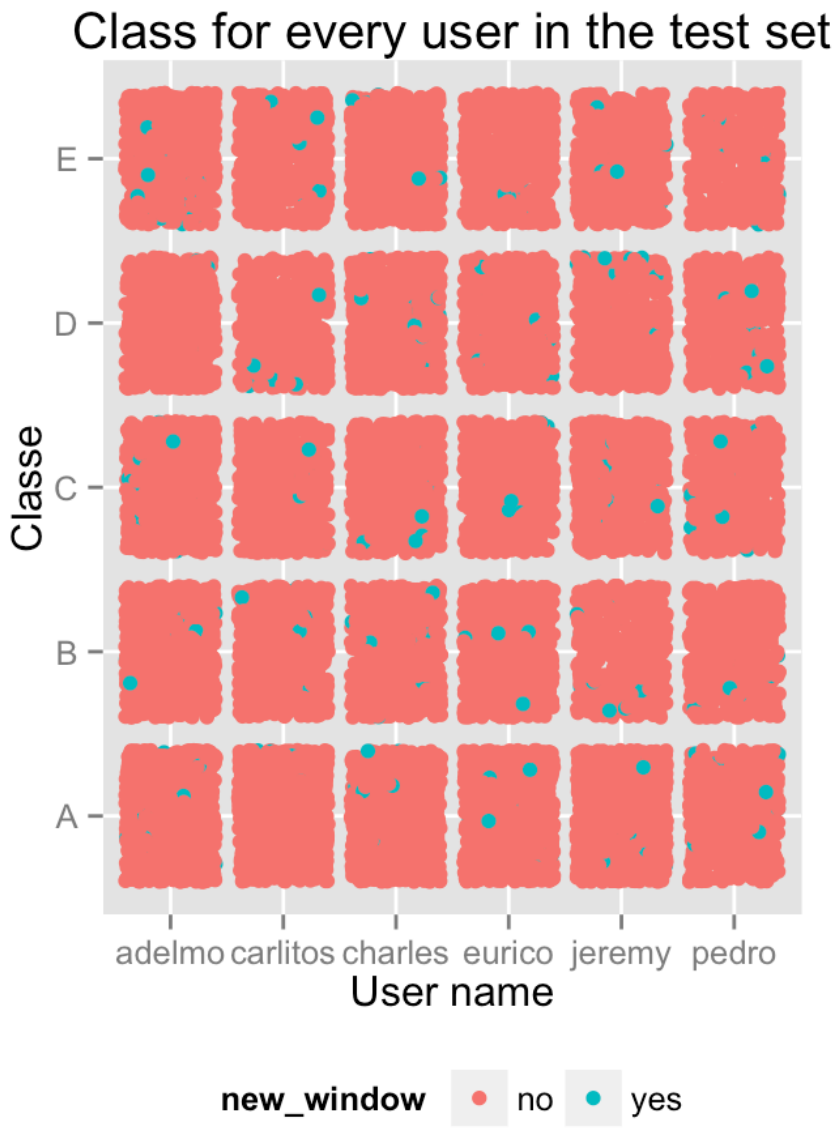
Data preprocessing

After loading the training and test datasets, unsuitable predictors are removed. Three concepts are used to determine if a predictor is unsuitable or not:

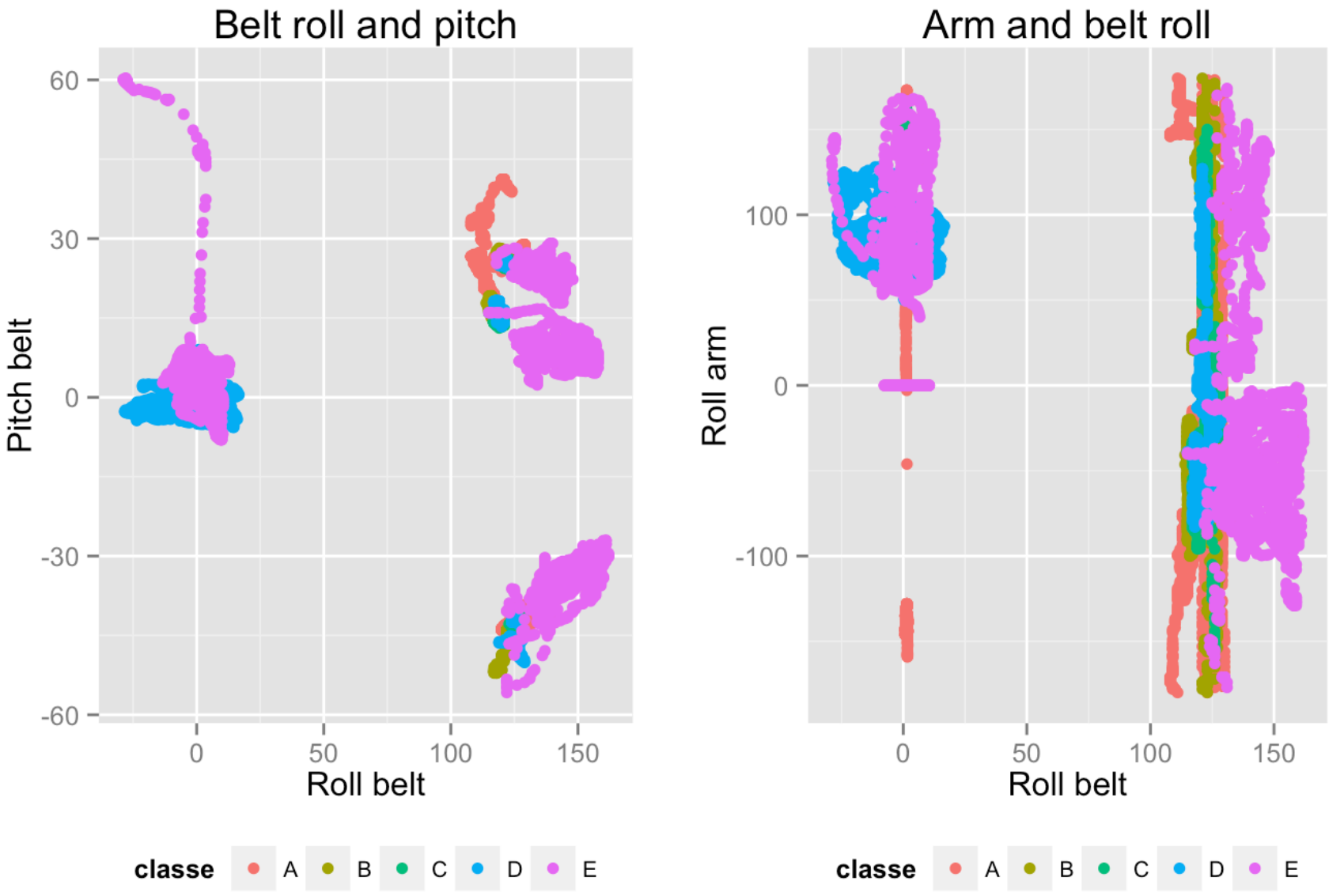
1. Columns with mostly NA or empty values are removed. If there is no data available, their for use prediction is limited.
2. Near zero variability suggests that the variable is not a good predictor.
3. Highly correlated variables can be removed or manipulated (e.g. PCA) to reduce the number of predictors.

There are a 100 columns from the original dataset with a total of 19216 NA or empty (i.e.“”) values, these are removed.





The previous graphs do not show a pattern based on the subject, the time window or the timestamps. These variables can also be removed and only the data related to physical measured variables (e.g. acceleration) should be considered as predictors. The near zero variability evaluation showed that only the new_window variable had near zero variability, so no new information was acquired.



The correlation matrix shows that the roll_belt is highly correlated with total_accel_belt, accel_belt_y and accel_belt_z. One could remove the roll_belt but neither the total_accel_belt, the accel_belt_y or accel_belt_z are correlated with each other however, as it is only one variable it is decided to leave it as part of the model. Other variables are correlated but there is no clear relationship between variables so they are left as part of the model.

The previous figures (i.e. Belt roll and pitch, Arm and belt roll) show that within each accelerometer and across accelerometers there does not seem to be a relationship that could be exploited to reduce the number of predictors. Although one could try a Principal Competent Analysis (PCA) the resulting vectors would be difficult to physically interpret.

Machine algorithm

Being the classe variable a factor, a tree approach is suitable. However, with so many potential predictors and with no way of knowing their accuracy and quality they should be considered weak predictors. Based on this a boosting with trees approach is followed. ### Cross validation A k-fold cross validation with 10 folds is followed, this is setup as part of the train function using a trainControl object. According to the results of the model the expected in-sample accuracy is close to 0.97.

```
Stochastic Gradient Boosting

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

No pre-processing
Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 17659, 17660, 17661, 17659, 17660, 17659, ...

Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy  Kappa   Accuracy SD  Kappa SD
  1                   50      0.7502    0.6832  0.011095     0.014112
  1                  100      0.8228    0.7757  0.011564     0.014705
  1                  150      0.8531    0.8141  0.012345     0.015652
  2                   50      0.8559    0.8174  0.011808     0.015034
  2                  100      0.9076    0.8831  0.008620     0.010920
  2                  150      0.9345    0.9171  0.008143     0.010307
  3                   50      0.8980    0.8709  0.009190     0.011606
  3                  100      0.9435    0.9285  0.007073     0.008928
  3                  150      0.9630    0.9531  0.006163     0.007791

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Accuracy was used to select the optimal model using  the largest value.
The final values used for the model were n.trees = 150,
 interaction.depth = 3 and shrinkage = 0.1.
```

Prediction on the test data.

	user_name	prediction
1	pedro	B
2	jeremy	A
3	jeremy	B
4	adelmo	A
5	eurico	A
6	jeremy	E
7	jeremy	D
8	jeremy	B
9	carlitos	A
10	charles	A
11	carlitos	B
12	jeremy	C
13	eurico	B
14	jeremy	A
15	jeremy	E
16	eurico	E
17	pedro	A
18	carlitos	B
19	pedro	B
20	eurico	B