Predicting Exercise Type From Wearable Device Data

Alex Chertok

June 12, 2018

Exersice Type Prediction From Data Recorded By Wearable Devices

Introduction

In the goal of this project is to predict the type of exercise being performed by a volunteer by using data collected by an accelerometer within a device worn by the volunteer while performing the exercise.

Data Cleaning and Analysis

The training data contains 19,622 records of 160 columns. Trying to visualize a dataset with such a large number of variables doesn't seem feasible, so other methods of exploratory analysis must be employed.

One of the most fundamental considerations that needs to be made is whether any data is missing. Summing up the number of NA values by column, an interesting pattern emerges: of the 160 variables, 6 have only NA values, while another 94 of them have between 19,216 and 19,301 NAs (approximately 98% of all records). This means that 6 of the 160 variables are essentially non-existent, while another 94 are apparently very poor candidates for usable predictors.

One possible way that variables which are mostly, but not entirely, NAs could be potentially used is as a factor where all NA records are one and others are zero. If it is the case that different types of exercise generate valid values for some of these 94 measurements at different rates, then that fact could be useful for classification. To investigate this possibility, I made a table of the exercise class and the number of NAs by record in the training dataset:

Distribution of Number of NA Values Among 94 Columns of Interest, by Value of the classe Variable

	0	2	4	5	6	7	9	11	12	14	16	22	94
A	0.0097	0.0014	0.0027	0e+00	0.0011	0.0032	5e-04	2e-04	2e-04	2e-04	2e-04	2e-04	0.9805
B	0.0126	0.0005	0.0037	0e+00	0.0000	0.0037	0e+00	0e+00	3e-04	0e+00	0e+00	0e+00	0.9792
C	0.0108	0.0009	0.0044	0e+00	0.0003	0.0035	3e-04	0e+00	3e-04	0e+00	0e+00	0e+00	0.9795
D	0.0109	0.0022	0.0050	6e-04	0.0000	0.0019	0e+00	0e+00	9e-04	0e+00	0e+00	0e+00	0.9785
E	0.0119	0.0003	0.0028	3e-04	0.0006	0.0061	0e+00	0e+00	0e+00	0e+00	0e+00	0e+00	0.9781

This table shows that the distribution of number of NAs is quite similar across different exercise types, with around 98% of records having only NAs and around 1% of records having no NAs for each type. Records with some, but not all of the 94 variables containing NAs are a very small minority (substantially less than 1%) for each exercise type and overall. Since there is no evidence that NA distribution between exercise types is different, it seems reasonable to exclude all 94 of the variable considered in this analysis from contention as possible useful predictors.

This leaves a dataset with 60 variables, including the response, and no NA values remaining. Of the remaining variables, there are a few that do not logically belong in any prediction model. The unique record number (V1) clearly has no relationship to the response variable and the three timestamp variables (raw_timestamp_part_1, raw_timestamp_part_1, cvtd_timestamp) also seem to have no viable relationship to exercise type. This leaves 56 variables in the training set, and therefore 55 possible predictors.

Modeling

Since this is high-dimensional dataset, selecting the optimal set of variables for linear or generalized linear models has the potential to be a difficult an time-consuming process. With that in mind, I decided to fit a linear discriminant analysis (LDA) model, which includes dimension reduction and is also computationally efficient. Fitting LDA models on all 55 predictors with explained variance thresholds of 80%, 90%, 95%, and 99% and evaluating by 5-fold cross-validation gave an accuracy of approximately 75% in every case. This is clearly not an adequate rate of accuracy given that the goal is to correctly classify all 20 of the test set records, even before considering that the test accuracy may be lower than the cross-validated accuracy rate. Applying the LDA predictions to the test set gives 13 correct predictions out of 20 for each threshold used, giving a test error of 65%. A representative model output (using an 80% threshold is shown below).

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.7445213	0.6764272	0.0017617	0.0020765

In order to improve the prediction accuracy, I elected to use a random forest, which is a substantially more powerful, but computationally intensive classification method. Fitting a random forest (method = “rf” in the caret package train() function) with all 55 predictors and using 5-fold cross-validation to select the best tree in the ensemble gave an accuracy rate of 99.8%. This a very substantial improvement over the LDA accuracy. In theory, assuming that all records are independent, this accuracy rate should result in a more than 96% chance of predicting all 20 of the test set records. Applying the random forest predictions to the test set shows that in fact all 20 of the predictions were correct. Below is the random forest model output.

mtry	Accuracy	Kappa	AccuracySD	KappaSD
2	0.9963818	0.9954232	0.0015947	0.0020172
28	0.9986240	0.9982596	0.0006883	0.0008705
55	0.9968914	0.9960678	0.0010564	0.0013364

Conclusion

These results make it clear that the random forest approach works very well for predicting the exercise type in our dataset. However, it should be emphasized that there is a high computational cost, as the random forest training took more than 100 times longer to run than the LDA training (several minutes against a few seconds).