Coursera Practical Machine Learning Project

July 2014

By: Mario Segal

Objective

The project objective is to try to predict whether an exercise is done correctly by using measurements from sensors placed on test subjects that were asked to do certain exercises in different ways (some correct and some incorrect).

The data was collected by Velloso, Bulling, Gellersen, Ugulino, and Fuks and their results presented in "Qualitative Activity Recognition of Weight Lifting Exercises."; Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) ; Stuttgart, Germany: ACM SIGCHI, 2013. The data was downloaded from http://groupware.les.inf.puc-rio.br/har

The researchers placed sensors on the arm, forearm, belt of the subjects as well as on the dumbbell used. In total five different ways of doing the exercise were performed (labeled as A,B,C,D and E)

Methodology

For the exercise, the data was split into a training and a test set by Professor Jeff Leek and his colleagues.

Consistent with proper methods, I used the training data to explore and develop a predictive model and used the final model on the test data to evaluate my results.

The first step was to properly read the data. Specifically many fields were imported as text and needed conversion to numeric measures. In addition, I identified that many fields contained summary information for a session (such as maximum and minimum readings). These fields were very sparse and were hence removed fro the training set

I performed some visual exploratory analysis of the data by looking at the distributions by exercise class for each variable. From such analysis I was able to identify that while some variables showed very good separation at least among some classes, others did not.

In order to reduce the number of features and reduce noise, I performed Principal Components Analysis using the caret package in R. Specifically I performed a different principal component reduction on the set of measures for each sensor and then combined such measures in a new training set

Initially, I attempted to model using a single CART tree using the caret package with method 'rpart', due to its ease of interpretation. Unfortunately this was not successful as the resulting model did not predict 2 of the classes and had very low accuracy.

I then proceeded to use a Random Forest using the caret packages using method 'rf'. The Random Forest Model performed very well with reported accuracy of 98.2% and Kappa of 97.7%. These measures were estimated for the out of sample using 25 bootstrap samples. In reality when I applied the model to the training data set I was able to predict with 100% accuracy as can be seen in Figure 1.

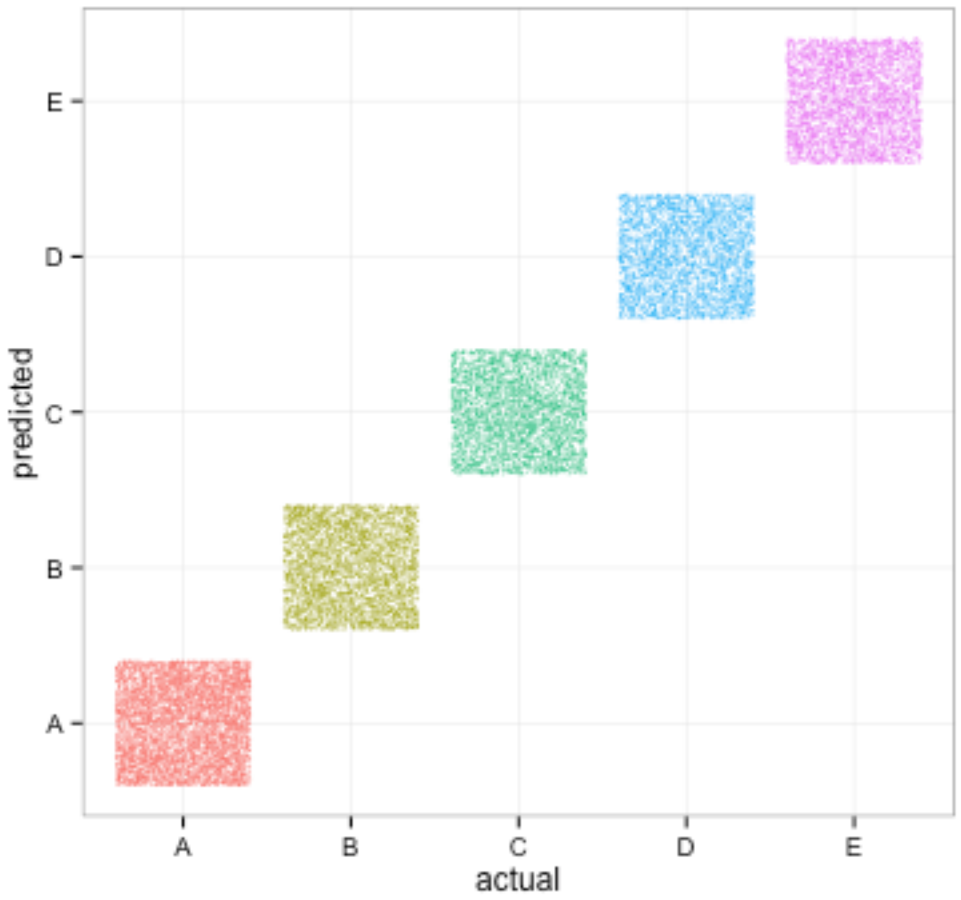

Figure 1

Figure 2 shows a jittered scatter plot of the actual and predicted values for the testing dataset – it shoes that in this case the Random Forest Model predicted with 100% accuracy on the training set

Results

As can be seen in Figure 1 below, The Radom Forest Model has 4 dumbbell and 1 belt principal components as the most important measures, with importance decreasing rapidly afterwards. This result suggests that a sensor in the dumbbell, which should be easy to accomplish, could provide sufficient feedback to the exerciser, however we did not test the accuracy of a dumbbell only sensor for prediction.

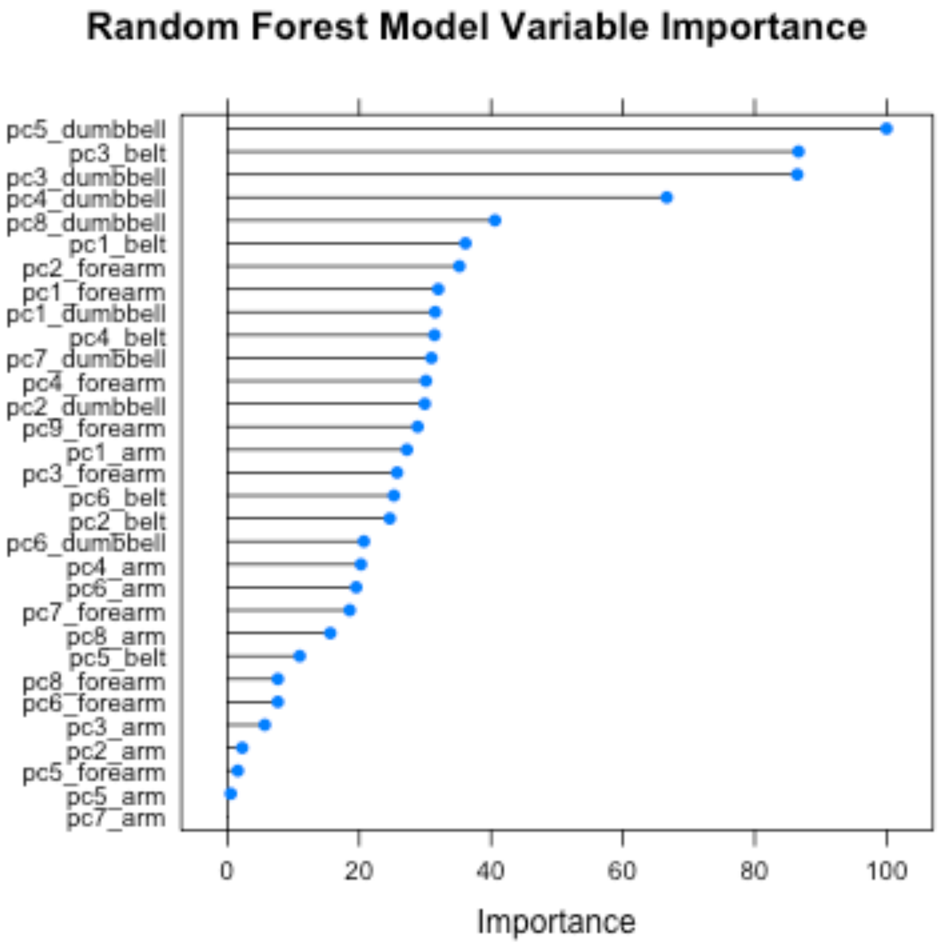

Figure 2

Variable Importance Plot for the Radom Forest Model. 4 of the 5 most important variables are principal components of the dumbbell measures. This result implies that a sensor in the dumbbell could be sufficient for proper prediction.

When I applied the model to the test set (after transforming it with the same principal component used for the train set), I was able to correctly predict all 20 cases, this result is a bit surprising as one never expects perfect prediction.
The researchers mentioned that expert trainers monitored the test subjects and that a relatively light dumbbell was used to avoid risk of injury. I suspect that in a real world scenario with less supervision and heavier dumbbells the model will not predict perfectly.

The interpretation of the meaning of principal components is always challenging. In this case it is more so given that the measurements of the sensors deal with concepts not very familiar to most people (yaw, roll, pitch). To show how they work Figure 2 plots the 2 most important PCs for the dumbbell sensor with the observations colored by exercise type

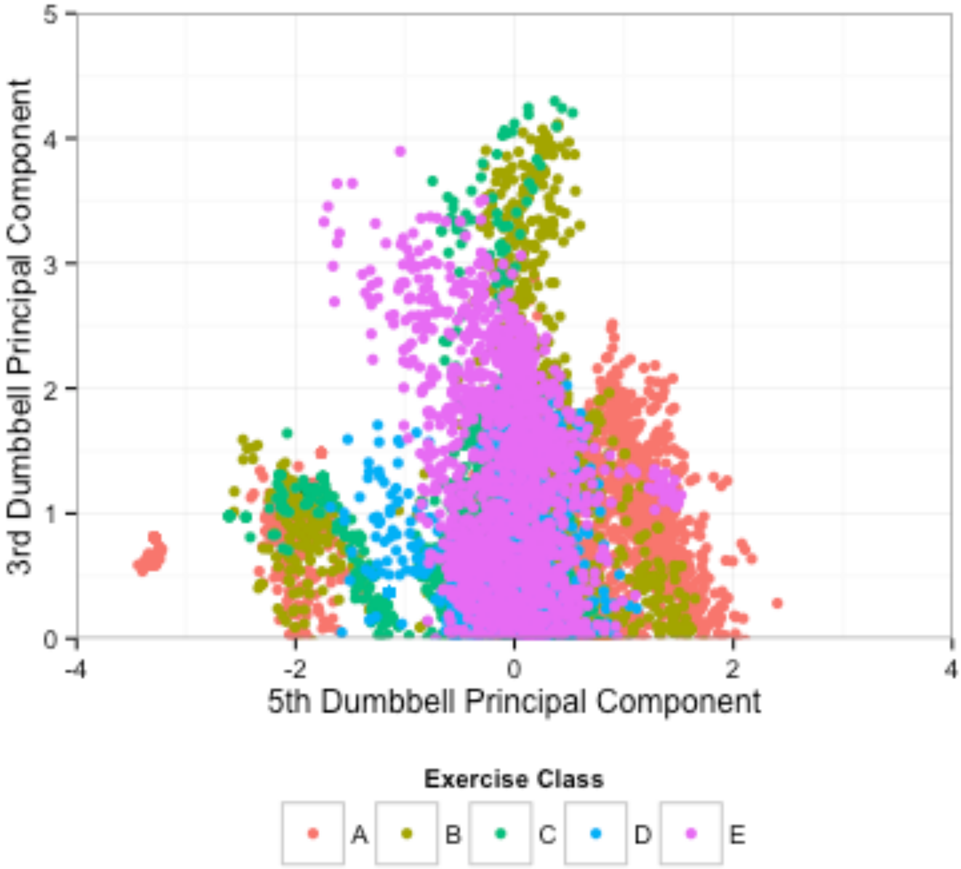

Figure 3

Figure 3 shows a scatter plot of the 2 most important principal components for the dumbbell sensor. One can see that observations are clustered together especially those for exercise E.

Conclusions

I find that the predominance of smart phones and other smart devices will provide more opportunities to collect data and also to use such data to help us (of course data can always be used in bad ways, which makes this a challenging topic).

I normally use my phone to track my training runs, it provides me with details like tempo and speed. I can see how the addition of some sensors could provide additional information such as fluid loss, and even critique my running form (for example a reminder that you are losing form can be all that one needs to focus on it which normally improves performance)

In the example in question, people could use sensors and have a phone tell them if they are doing the exercise properly, which is desirable for people who want to work out on their own or have no access for a trainer (but less desirable for trainers)