Title: Predicting A Person's Movements Using Accelerometer and Gyroscope Data

Introduction:

The Samsung Galaxy SII is a smartphone built with an internal accelerometer and gyroscope which can be used to quantitatively measure the movement of the phone [1]. SmartLab leveraged this technology by attaching it to 30 subjects to measure their movements when performing six different activities using the smartphone [2]. The quantitative data gathered can be used to then predict a person's activity. By building accurate movement prediction technology, we can develop new and improved assistive technology applications, such as software to monitor daily exercise or recognize when a phone is being used. The goal is to parse through the quantitative variables and determine if it is possible to accurately predict new subjects' activities based upon data provided.

Methods:

Data Collection

To build the model, we were given a processed set of data collected from a Samsung Galaxy SII, which was provided by Professor Jeff Leek on Coursera [3]. The original dataset is available on the UCI Machine Learning Repository's website [4].

Exploratory Analysis

This dataset contains 7,352 observations of 561 quantitative variables as well as columns identifying the Subject and the Activity. We immediately subsetted the dataset into a training set and a test set. The training set was further divided into the training set for building a predictive model, and a validation set for testing the model before applying the model to the actual test set.

Subjects 27, 28, 29, and 30 were reserved for the test group and contained 1,485 observations. Subjects 22, 23, 25, and 26 were used as the validation group and contained 1,494 observations. Data from the remaining thirteen subjects were all used in the training set for building the models; a total of 4,373 observations. There were no NAs in the data, but the Activity column had to be converted into a factor variable for statistical analysis. There were also a number of repeated column names, which had to be edited to be unique.

To make sure that there were no skews in the number of measurements from any one subject and that there was a reasonable data on each activity, we used a heatmap (Figure 1) to visualize the data [5]. It can be seen that every subject's activity level was fairly similar, although it is clear that Subject 1 did more "walking" than others, while Subject 21 did more "laying" than others.

There is most likely a number of confounding variables among the quantitative columns too, however, the size of the dataset made this difficult to investigate. They did not appear to cause problems in our initial prediction models, so they were not researched more in-depth.

Statistics

The end goal was to try and predict the activity being performed based on the quantitative measurements. Since the variable we were predicting was a factor, we thought predictive trees would be most suitable. Random forests is an ensemble of many decision trees which is similar to bagging, but frequently produces more accurate predictions [6]. Since the dataset was not extremely large and would not overload our processor, we decided random forests would most likely yield the best results. Although there is the risk of overfitting, we were confident that using the validation set, we would be able to overcome this issue.

Results:

We built two predictive models using random forests—one including the Subject column, and one with the Subject column removed. Both models generated 500 trees with 23 variables tried at each split. We found the model which did not include the subject had a slightly higher estimated error rate—OOB estimate of error rate: 1.85% vs. OOB estimate of error rate: 1.69%—but both consistently yielded a 10% error rate with our validation set.

According to the confusion matrix, the two activities which were most misclassified were "sitting" and "standing". This is not surprising since both activities involve little movement. It is surprising though that "laying" was almost never misclassified despite also involving little movement. This misclassification pattern held true when we applied the prediction model to our validation set. Ultimately, we decided to use the model including the Subject column due to its slightly lower estimated error rate.

Our final predictive model performed well in predicting the test group's activity. The error rate of the predictions was 5.9% which means the predictions were correct 94.1% of the time. A surprise was that most of the errors were misclassifications between the three "walking" activities, which is different from in our previous results.

Conclusions:

Our analysis concludes that accelerometer and gyroscope measurements can be used to predict a subject's activity relatively accurately. It is easy to distinguish the group of non-mobile activities (laying, sitting, standing) from the mobile activities (walking, walking down, walking up), however there were still a number of misclassifications within each group of three activities.

This prediction model could be further improved with more data, but it is already very effective. However, this was also a very limited experiment, wherein subjects only performed six different activities. Although it is a good start, it would be interesting to see if comparable predictive models can be built when there is a much larger number of activities to predict, up to a potentially infinite number of activities. This small sample proves that building such models should be possible, but that it may require significantly more data and computing power.

Figure

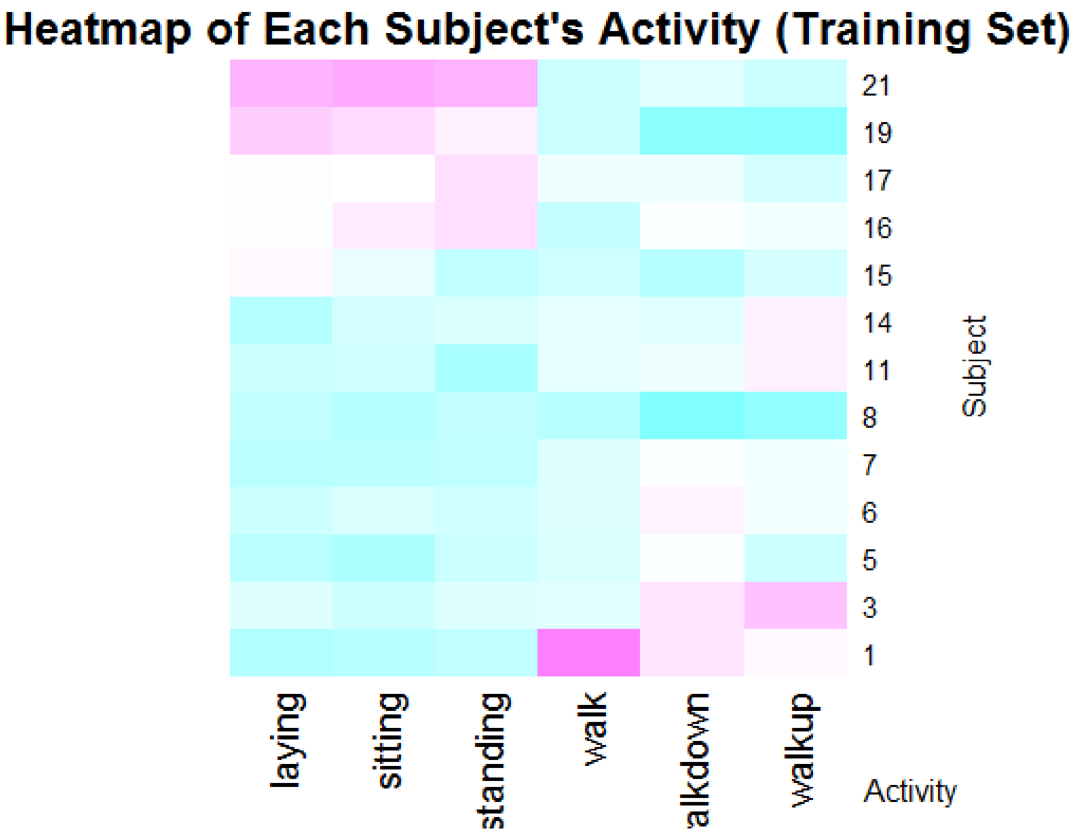

Figure 1: This heatmap shows the number of measurements for each subject's activity, relative to other subjects in the training set. The darker purple represents activities that one subject performed more than the other subjects. The darker blue represents activities which one subject performed less than their peers. For example, Subject 1 did more "walking" than the other subjects; and Subject 8 did less "walking down" than others.

References:

1. Chipworks "Silicon Summary in the Samsung Galaxy S II" Page. URL: http://www.chipworks.com/blog/recentteardowns/2011/07/25/silicon-summary-in-the-samsung-galaxy-s-ii/. Accessed 03/09/13.
2. smartLab "ESANN 2013 Special Session" Page. URL: http://smartlab.ws/component/content/article?id=59. Accessed 03/09/2013.
3. Coursera "Assessment Details | Data Analysis" Page. URL: https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda. Accessed 03/07/2013.
4. UCI Machine Learning Repository "Human Activity Recognition Using Smartphones Data Set" Page. URL: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. Accessed 03/07/2013.
5. Flowing Data "How to Make a Heatmap – a Quick and Easy Solution" Page. URL: http://flowingdata.com/2010/01/21/how-to-make-a-heatmap-a-quick-and-easy-solution/. Accessed 03/08/2013.
6. Wikipedia "Random Forests" Page. URL: http://en.wikipedia.org/wiki/Random_forest. Accessed 03/09/2013.