Joseph Elliott

Data Analysis Assignment 2

Prediction of Physical Activity from Cell Phone Data

1 Introduction

Researchers carried out an experiment to determine how motion could be detected by a smartphone. A group of 30 volunteers between the ages of 19-48 was selected for this experiment, where different physical activities were performed while wearing a smartphone (Samsung Galaxy S ll) strapped to the participant's waist. Using the phone's accelerometer and gyroscope, data was gathered about the phone including three-axis linear acceleration and three-axis angular velocity (Anguita, Ghio, Oneto, Parra, & Reyes-Ortiz, 2012; Frank & Asuncion, 2010).

In this analysis, the data from the experiments was utilized to form a prediction of

This analysis is most beneficial in the field of medical research and physical therapy. In cases where physical therapy is needed, physicians could use a smartphone to monitor a patient's motion, preventing the need for high-cost sensors or in-patient testing and monitoring. This research also may have implications for sports/fitness professionals as well, allowing them to track physical activities easily with little need for extra tracking or timing devices. Future analysis may be able to detect differences in acceleration patterns between different groups of people, enabling the study of walking gaits, postures, and possibly the detection of orthopedic or podiatric problems in patients even before symptoms arise.

To predict the type of physical activity based on the cell phone data, a random forest was fitted to a training set of data, and used to predict a test set of data. This random forest correctly predicted the activity for 94.88% of the test set observation, a misclassification rate of just 5.12%.

2 Methods

2.1 Data Collection

A pre-processed data set was provided through the Coursera Course website.

https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda

And was downloaded on March 3, 2013 using the R programming language (The R Project for Statistical Computing, 2013). The data set consisted of 7352 observations for the 30 subjects involved, on 563 variables. Of these variables, 561 consists of a vector that records triaxial acceleration, triaxial angular velocity, and calculated variables such as jerk (3rd derivative), entropy, and energy. This vector consists of time and frequency domain variables and information such as min, max, mean, standard deviation, skewness, etc. for many variables.

The observations also record a numeric identifier for the subject, and the physical activity being performed, as one of the following:

- walking
- walking upstairs
- walking downstairs
- sitting
- standing
- laying

2.2 Data Processing

Although the data was provided in a somewhat processed form, additional processing was needed to convert the subject identifier and physical activity labels into factor variables. There were no "NA" values present in the dataset.

2.3 Statistical modeling

The statistical modeling methods performed to develop a prediction function were the random forest method. This method was chosen because the predicting problem involves classification, where a factor variable is to be predicted, and it is likely that there is not a linear relationship between the outcomes and the variables. Classification trees and random forest are methods that are best suited to non-linear data.

Random forest begins with prediction trees, which evaluate the variables provided for a given data set and split on the variables which best describe the data (Leek, 2013). A random forest approach fits multiple trees, and for a particular observation, the forest 'votes' (each tree has an outcome) on how to classify it, with the most prevalent classification winning out (Breiman & Cutler).

To perform the random forest approach with this data set, the data set was divided into a training and a test set, with the following characteristics:

- Training set – all data except the subjects included in the test set. 5867 total observations, 17 subjects.
- Test set – data from subjects (27,28,29,30), 371 total observations, 4 subjects.

3 Results

3.1 Resulting prediction rate

As an exploratory analysis, a single classification tree was fitted to the training data, with the outcome as activity and the predictors being all variables in the data set except for the subject id number. This tree was then pruned to 10 leaves to provide the best prediction, based on the deviance and misclassification errors being minimized at that point. This classification tree correctly predicted 86.79% of the test set, with a re-substitution error rate on the training set of 89.80%. However, it was supposed that the random forest could give better prediction with less errors.

Joseph Elliott

Data Analysis Assignment 2

A random forest was fitted to the data, using activity as outcome and all other variables (except subject id number) as predictors. The random forest had 500 trees, with 23 variables tried at each split. This random forest was able to predict the training set with 100% accuracy upon re-subsitution. In addition, the random forest achieved a prediction accuracy of 94.88% on the test data set (misclassification rate of 5.12%). The out-of-bag (OOB) error rate for the random forest was 1.77%. The confusion matrix for the random forest can be seen below.

Table 1 – Confusion matrix for from the random forest prediction.

	laying	sitting	standing	walk	walkdown	walkup	class.error
laying	1113	0	0	0	0	1	0.000897666
sitting	0	993	28	0	0	1	0.028375734
standing	0	43	1048	0	0	0	0.039413382
walk	0	0	0	983	7	7	0.014042126
walkdown	0	0	0	4	775	7	0.013994911
walkup	0	0	0	0	6	851	0.007001167

The variables were then ordered according to their importance and a plot was used to visualize the importance of the variables based on their Gini coefficient.

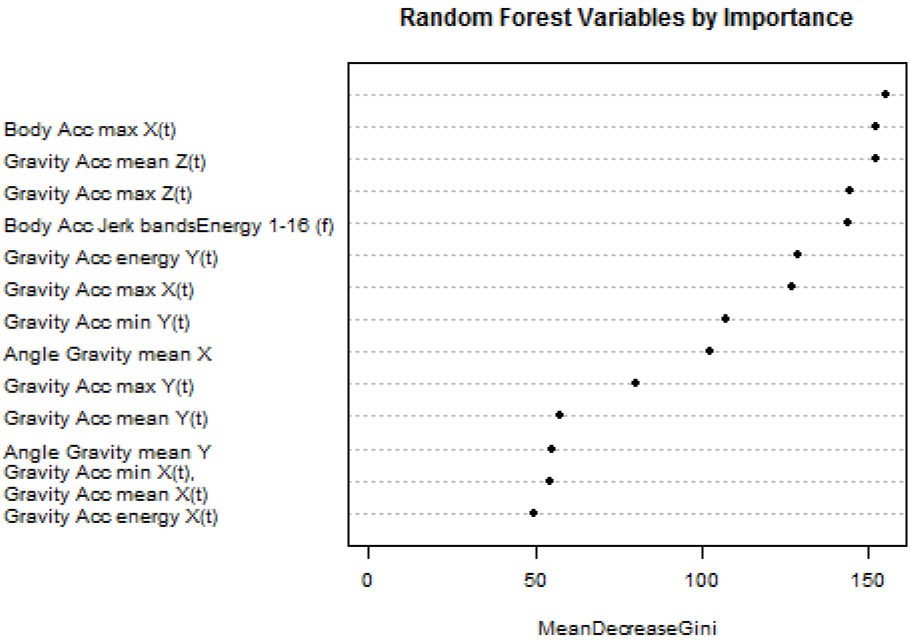

Figure 1 – In this figure, the top 14 variables from the Samsung Data as selected by a random forest prediction are shown plotted against their mean decrease in their Gini index. This shows how important the variables are to the prediction model.

According to Breiman and Cutler, there is no need to perform cross-validation on a random forest prediction to get an unbiased test set error rate, as this performed during the bootstrapping process used during the random forest generation. The OOB error can serve as an unbiased estimate of test set error (Breiman & Cutler).

3.2 Potential problems

In this analysis, as there were over 500+ variables, and the relationships between them and the activity were not clear, no attempt was made to select variables as predictors, instead all measured variables were provided to the random forest and it selected the best options for classification. This approach, while effective, could be improved by more knowledge of the subject. The random forest calculation with many variables is quite computationally intensive (and seems to run only on one core of a multi-core processor in R), so limiting the number of predicting variables may decrease compute time.

With respect to potential confounders, it was not really clear from this non-linear data set which variables can be considered confounders.

4 Conclusion

The random forest method can be a powerful way to predict classification of non-linear outcomes. Utilizing this approach on this set of cell phone data can generate very good predictions on the type of activity being performed based on records from the cell phone's accelerometer and gyroscope. This prediction information may assist health and fitness professional to monitor their patient's remotely without the need for complicated sensory equipment.

5 Works Cited

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2012 December). Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. *International Workshop of Ambient Assisted Living (IWAAL 2012)*.

Breiman, L., & Cutler, A. (n.d.). *Random Forests*. (UC Berkeley) Retrieved March 7, 2013, from http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Frank, A., & Asuncion, A. (2010). *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].* Irvine, CA: University of California, School of Information and Computer Science.

Leek, J. (2013). *Predicting with Trees.* Retrieved March 8, 2013, from Coursera.org: file:///Users/jtleek/Dropbox/Public/courseraPublic/week6/004predictingTrees/index.html

The R Project for Statistical Computing. (2013). Retrieved 2013 12-February from http://www.r-project.org/