Predicting User Activity from Phone Telemetry

gengelbeck@gmail.com

June 30, 2014

Abstract

The goal of this project was to build a model that could accurately predict the activity a subject is performing using smart-phone inertial sensor data. Using a random forest model we were able to correctly classify about 95 percent of activities in our test dataset.

Contents

1 Introduction **3**

1.1 Background . . . 3

2 Methods **3**

2.1 Data Collection . . . 3

2.2 Exploratory Analysis . . . 3

2.3 Splitting the data . . . 3

2.4 Training . . . 4

2.4.1 Benchmark . . . 4

2.4.2 Picking Predictors . . . 4

2.4.3 Creating a Prediction Model . . . 4

2.5 Testing . . . 5

2.6 Reproducibility . . . 5

3 Results **5**

3.1 Predictive Accuracy . . . 5

3.2 Important Variables . . . 5

4 Conclusions **7**

4.1 Predicting Activities . . . 7

4.2 Limitations . . . 7

1 Introduction

1.1 Background

Using smart-phone inertial sensor data we set out to predict the activities of humans. While desirable, our goal *was not* to create a predictive model that was easy to interpret: For us, predictive accuracy was paramount, understanding was secondary.

2 Methods

2.1 Data Collection

For our analysis we used a dataset of 7352 activity recordings from 30 participants between the ages of 19 and 48 years of age. Each participant in the dataset performed six activities: walking, walking up stairs, walking down stairs, sitting, standing, and lying down.

Activity recordings were collected from a waist-mounted Samsung smart-phone's inertial sensors. Each participant was video taped and their activities were coded from the videos and added to the dataset.

This dataset contains our outcome variable (activity), a participant identifier, and sensor covariates. Briefly, the dataset contains:

- The activity the participant was engaged in. (Our outcome variable.)
- A participant identifier,
- A 561-feature vector with time and frequency domain variables. These values capture:
  - Triaxial acceleration from the accelerometer and the estimated body acceleration,
  - Triaxial Angular velocity from the gyroscope.

More details about our dataset are available at the UCI Machine Learning Repository where the dataset is available for download [8].

The dataset we used was downloaded from the *ds1* course Amazon AWS data store [3] on March 3, 2013 using the R programming language [7].

2.2 Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of our data. Exploratory analysis was used to:

1. Identify missing values – none were found, and
2. Verify the quality of the data

We did not detect any data data quality issues with the dataset.

2.3 Splitting the data

We split our original dataset into three separate datasets: a training, test, and validation dataset.

The training dataset contained the data for participants 1,3,5,6,7,11,14,15,19,22,23,25, and 26 and was used to create our models.

Our validation dataset contained the data for participants 8, 16, 17, and 21 and was used to *validate* our models before applying they to the test dataset.

Finally, the test dataset contained the data for participants 27, 28, 29, and 30. This dataset was used once and it was used to test the prediction accuracy of our model.

2.4 Training

2.4.1 Benchmark

For our model building, we used the estimated probable perfect classification ($\frac{1}{2} \times \textit{test set sample size}$) as a benchmark (See [2]). Our test set sample size was 4444 samples, so our classification benchmark was estimated to be 2222 correct classifications for our test dataset.

2.4.2 Picking Predictors

Given the large number of highly correlated predictors, we set about eliminating unneeded covariates. We used the procedure suggested my Kuhn [6] p. 5 for reducing the effects of multicollinearity. Using this procedure, 344 of 561 covariates were removed from the dataset, leaving us with 217 covariates.

2.4.3 Creating a Prediction Model

Random Forest Model We created a random forest model to predict activities in the training dataset. We fit a random forest models of 50 trees to our test dataset using R's *randomForest* package. (See [5] for a discussion of the *randomForest* package.) The resulting random forest model classified 4331 of 4444 correctly for a correct classification rate of about 0.98.

Validation We used our random forest model to predict activities in our validation dataset. We were able to correctly classify 1408 of 1485 activities for a correct classification rate of about 0.95. This was a drop of about 0.03 in the correction classification rate from the training dataset. Inspecting the confusion matrix, 75 of 77 (97 percent) of the misclassifications of our validation activities were confusions between standing and sitting activities.

Combining Random Forest We decided to see if a model that combined several Random Forest Models would improve our predictions. We fit three random forest of 50 trees to our validation dataset using R's *randomForest* package. We then combined these three models together into one random forest model.

The resulting random forest model classified 4444 of 4444 correctly for a correct classification rate of 1. We, of course, suspected that we had over-fit our data.

Validation To see if we had over-fit our data, we validated our combined model against our validation dataset. The combined model was able to correctly classify 1398 of 1485 activities for a correct classification rate of about 0.94. This was a drop of about 0.01 in the correction classification rate from our original random forest model. An inspection of the confusion matrix again showed that most misclassifed activities were confusions between standing and sitting activities: 85 of 86 (98 percent) of the misclassifications. When comparing both model, the combined model also produces more misclassifications: 85 misclassifications for the combined model compared to 77 for our first model. We concluded that the combined model had over-fit our data and we decided to use our original random forest model.

2.5 Testing

Finally, we tested our first random forest model against our test dataset. We were able to correctly classify 1405 of 1485 activities for a correct classification rate of about 0.95. This was a drop of about 0.002 in the correction classification rate from our validation dataset.

2.6 Reproducibility

Our analyses are captured and reproduced in the R script files [4]. These script files are available on request.

To reproduce the exact results presented in this manuscript the cached version of the analysis must be performed, as the data available from course data source may change.

3 Results

3.1 Predictive Accuracy

We created a random forest model that was able to correctly classify about 95 percent of the activities in our test dataset. Table 1 below presents the confusion matrix for our random forest model.

The confusion matrix shows that most prediction confusions occurred when predicting sitting and standing activities. With our test dataset, 76 of 79 (96 percent) of misclassifications were confusions between standing and sitting activities. See Table 1 below for the confusion matrix for our test dataset.

Table 1: Confusion matrix the resulted from using the random forest model to predict activities in our test dataset. Observed misclassification counts have been highlighted.

	laying	sitting	standing	walk	walkdown	walkup
laying	293	0	0	0	0	0
sitting	0	220	44	0	0	0
standing	0	32	251	0	0	0
walk	0	0	0	229	0	0
walkdown	0	0	0	1	196	3
walkup	0	0	0	0	0	216

3.2 Important Variables

Our random forest model uses all 344 covariates available to it. As Figure 1 shows, not all our were covariates were equally important as predictors.

The ten most important predictors as measured by Mean Decrease in Gini coefficients are shown in Table 2. These covariates may indicates the relative importance of distinguishing energy and acceleration caused by humans and that caused by gravitational forces.

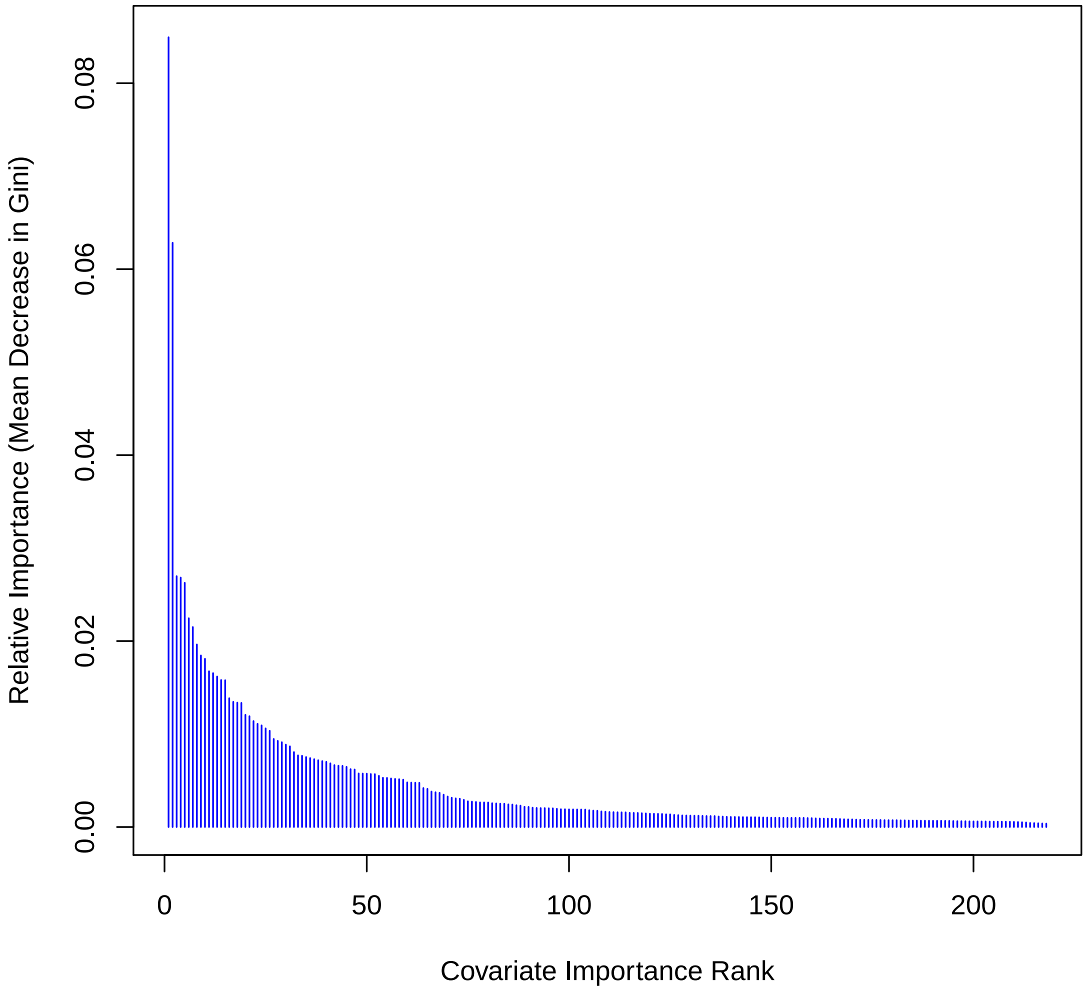

Figure 1: Bar plot of the relative importance of each of the 344 covariates used by forest tree model.

Table 2: List of the ten most important covariates in the random forest model as measured by Mean Decrease in Gini coefficient.

Col No.	Covariate	Mean Decrease Gini
53	tGravityAcc.min...X	313.87
58	tGravityAcc.energy...Y	232.21
64	tGravityAcc.entropy...Y	99.70
397	fBodyAccJerk.bandsEnergy...9.16	99.14
561	angle.Z.gravityMean.	97.05
382	fBodyAccJerk.bandsEnergy...1.8	82.99
66	tGravityAcc.arCoeff...X.1	79.48
506	fBodyAccMag.max..	72.56
59	tGravityAcc.energy...Z	68.18
410	fBodyAccJerk.bandsEnergy...1.8	66.86

To increase the comprehensibility of the model – a secondary goal for us – it may be worthwhile to reduce the number of convariates used by the model. Liaw & Wiener [5] report that variable importance can be used for eliminating noise variables and reducing the complexity of models.

4 Conclusions

4.1 Predicting Activities

We were able to create a random forest model that was able to classify about 95 percent of the activities in our test dataset correctly.

4.2 Limitations

Our understanding of the covariates provided in the dataset was limited. It appeared to us that a many of the covariates in the dataset could be combined into more accurate and sensible activity measures with physical correlates that would be easier to interpret and communicate. Greater knowledge of the domain may provide us with a model that is both understandable and accurate.

Additionally, our model is limited to the activities covered by the experiment. Our predictive model *would not* generalize to activities that are not in the original model (e.g., playing a video game).

Finally, the dataset does not reflect the *real-world* base rates for these activities (e.g., people are more likely to go walk around than to walk upstairs). If applied to real-world activities, our model would need to reflect differences in the base rates of activities.

References

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

[2] Leek, J., *Prediction Study Design* `https://d19vezwu8eufl6.cloudfront.net/dataanalysis/predictionStudyDesign.pdf`. Accessed 3/2/2013.

[3] Leek, J., *Samsung Activity Data for Data Analysis Assignment 2* `https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda`. Accessed 3/2/2013.

[4] Lemon, Jim, *Kickstarting R - Writing R scripts Page,*. `http://cran.r-project.org/doc/contrib/Lemon-kickstart/kr_scrpt.html` Accessed 2/17/2013.

[5] Liaw, A., & Wiener, M., *Classification and Regression by randomForest. R News: The Newsletter of the R Project*, 18-22, `http://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf`(2/3), December, 2002.

[6] Kuhn, M., Building Predictive Models in R Using the caret Package, *Journal of Statistical Software*, *28(5)*, 2008.

[7] R Core Team, *R: A language and environment for statistical computing.* `http://www.R-project.org` Accessed 2/17/2013.

[8] UC Irvine Machine Learning Repository. *Human Activity Recognition Using Smartphones Data Set* `http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones`. Accessed 3/2/2013.