

Title: Predicting Subject Behaviors from Accelerometer and Gyroscope Data

Introduction:

The UCI Machine Learning Repository [1] provides experimental data from their machine learning experiments to be used in experiments by other researchers. One such experiment involved recording and analyzing data gathered by accelerometers and gyroscopes embedded in Samsung Galaxy S II smartphones while human subjects performed a variety of physical activities [2]. This large data set is the subject of the data analysis that follows, wherein we attempt to construct a function which will predict the human subject's activity from the accelerometer and gyroscope data collected by the smartphone.

By validating such a model, we can demonstrate the data's predictive power which can be applied toward innovative smartphone applications (or for other devices) that are sensitive to or may have an interest in the motions of the human user.

Methods:

Data Collection

This analysis used a data sample from the UCI Machine Learning Repository [2] which was prescribed by instructor Jeff Leek for the Data Analysis class on Coursera.org [3]. The data contained 7,352 observations, each containing 563 variables (including the activity type and subject identifier). These data were downloaded from the course website on February 28, 2013 using the R programming language [4].

After collection, the data were partitioned along subject lines into training and test sets. The training set contained subjects numbered 1, 3, 5, and 6 (as prescribed by instructor Jeff Leek [3]); the test set contained subjects numbered 27, 28, 29, and 30 (again, as prescribed by instructor Jeff Leek). The remaining subjects were randomly placed into the two sets; the randomization seed was **20130228** and the final sets included:

- **training:** 1, 3, 5, 6, 7, 14, 15, 16, 19, 21, and 26
- **test:** 8, 11, 17, 22, 23, 25, 27, 28, 29, and 30

Exploratory Analysis

Several exploratory analyses were performed on the training data set to assess its quality and to identify potentially meaningful factors. First, the activity observations were coerced to be an R factor [4]. Second, columns with duplicate names (e.g., "fBodyAccJerk-bandsEnergy()-1,8") were modified (using prefixes X . , Y . , and Z .) to be unique. Lastly, a data frame was reconstructed from the original data frame as a means of sanitizing the illegal characters (e.g., (,) , - , and ,) from the subject data.

Since there were so many variables to consider, we performed a singular value decomposition (SVD) [5] in an attempt to narrow down the number of factors for the next phases of analysis. By plotting the values of the SVD, the "elbow" suggested that there were approximately five variables of interest, with **fBodyAccJerk-min()-X** (the minimum measurement of the frequency domain signal of the subject's body's linear acceleration along the x-axis) likely accounting for approximately 51% of the variance.

Finally, an exploratory tree model [6] was generated citing **activity** as the outcome and using all other variables as the covariates. This exploratory classification tree revealed seven variables as strong predictors:

- **fBodyAccJerk-std()-X** - the standard deviation of the frequency domain signal of the subject's body's linear acceleration along the x-axis;
- **tGravityAcc-min()-X** - the minimum measurement of the time domain signals of the gravity acceleration signal along the x-axis;
- **tGravityAcc-max()-Y** - the maximum measurement of the time domain signals of the gravity acceleration signal along the y-axis;
- **tBodyAccMag-std()** - the standard deviation of the time domain signals of the subject's body's acceleration

magnitude;

- **tGravityAcc-arCoeff()-Y,2** - the autorregresion coefficient (with Burg order equal to 4) of the time domain signals of the gravity acceleration signals along the y-axis for the second sample period;
- **fBodyAccJerk-maxInds-X** - the index of the frequency component with largest magnitude of the frequency domain signal of the subject's body's linear acceleration along the x-axis; and
- **tBodyGyro-arCoeff()-Y,1** - the autorregresion coefficient (with Burg order equal to 4) of the time domain signals of the subject's body's angular velocity.

The classification tree had **8** terminal nodes (see Figure 1), revealed a residual mean deviance of **0.6049**, and had a misclassification error rate of **0.1008**. We interpreted this as a fairly successful model and would use it as the basis for our predictive model.

Interestingly, there was only *one* intersecting variable between the list of contributors identified by the SVD and the actual variables used in the exploratory classification tree model. That intersecting variable was **fBodyAccJerk-maxInds-X**. (Curiously: this was not the *maximum* contributor from the SVD which (*vide supra*) was fBodyAccJerk-min()-X.) Of further interest was the observation that the most salient variables from both the SVD and the tree model were prefixed with “fBodyAccJerk” (the frequency domain signal of the subject's body's linear acceleration). Variables from this “fBodyAccJerk” family appeared as 3 of the 7 top contributors (as identified by the SVD) and as 2 of 7 variables used in the exploratory classification tree model. Indeed, obvious clustering of “energetic” (i.e., walk, walkdown, walkup) and “sedantary” (i.e., laying, sitting, standing) activities emerged when plotted (see Figure 2).

Statistical Modeling

Using the results from the exploratory classification tree as the basis for the statistical model, we prepared a refined classification tree, this time explicitly declaring the covariates from the actual tree construction (cited above). Not surprisingly, the statistics produced from this tree were identical to the original exploratory tree.

Satisfied with the tree model, we performed a standard multivariate linear regression [7]. Factors for the model were selected based on the outcomes of the exploratory and refined trees (*vide supra*). Coefficients were estimated with ordinary least squares [7].

This linear model would be used for predictions.

Confirmation

Predictions from the linear model were validated against the training data set using a misclassification function. The outcome factors (i.e., “activity” values) were coded as numeric values in the data frame. The predictions generated decimal values which were rounded and matched to the actual values; a sum of **0** indicated a match, while all other values were returned as their absolute value. The matches and mismatches were summed and divided by the sample size. The misclassification formula appears as follows:

$$MCr = \frac{\sum |Av - Round(Pv)|}{n}$$

Where MCr is misclassification rate, Av is the actual value, $Round(Pv)$ is the rounded predicted value, and n is the sample size.

Results:

We fit a regression model that looked at the activity as the outcome and accounted for seven covariates. Our final regression model was:

$$Act = b_0 + b_1(BAJmi_x) + f(BAJsd_x) + g(GAmin_x) + h(GAmax_y) + i(BAMsd) + j(GAac_{y,2}) + k(BGac_{y,1}) + e$$

where:

- b_0 was the intercept term;
- b_1 was slope of the activity;
- $BAJmi_x$ was fBodyAccJerk-maxInds-X;
- $f(BAJsd_x)$ was fBodyAccJerk-std()-X;

- $g(GAmin_x)$ was $tGravityAcc-min()-X$;
- $h(GAmax_y)$ was $tGravityAcc-max()-Y$;
- $i(BAMsd)$ was $tBodyAccMag-std()$;
- $j(GAac_{y,2})$ was $tGravityAcc-arCoeff()-Y,2$;
- $k(BGac_{y,1})$ was $tBodyGyro-arCoeff()-Y,1$; and
- e represents the error term (all unmeasured and unmodeled sources of variance in the activity outcome).

(See the “Exploratory Analysis” (above) for detailed descriptions of these covariates.)

We observed highly statistically significant relationships between activity and six of the seven variables ($fBodyAccJerk-maxInds-X$, $fBodyAccJerk-std()-X$, $tGravityAcc-min()-X$, $tGravityAcc-max()-Y$, $tBodyAccMag-std()$, $tBodyGyro-arCoeff()-Y,1$; $P < 0.001$ in all six cases). No statistical significance was found for the seventh variable ($tGravityAcc0-arCoeff()-Y,2$). Confidence intervals for these variables appear as:

Variable	2.5 %	97.5 %
$fBodyAccJerk-maxInds-X$	-0.76298715	-0.5823629
$fBodyAccJerk-std()-X$	0.31112851	0.6792702
$tGravityAcc-min()-X$	0.42053407	0.5797983
$tGravityAcc-max()-Y$	-1.59414443	-1.3636571
$tBodyAccMag-std()$	1.28391600	1.6099426
$tGravityAcc-arCoeff()-Y,2$	-0.05943848	0.1171762
$tBodyGyro-arCoeff()-Y,1$	-0.37110171	-0.1268098

We then collected a prediction from our training data set and applied it against our misclassification function. The misclassification function reported a misclassification rate of **0.03789474** (approximately 3.7895%); compare this with the original tree's misclassification rate of 0.1008 (approximately 10.08%). Finally we collected a prediction using our model, using the test set as our data and applied it against our misclassification function. The misclassification rate for the test data prediction was **0.05320946** – which was still a significant improvement over the original tree's misclassification rate.

Conclusions:

Given the low misclassification scores, this model appears to have excellent predictive power for inferring a subject's activity type from their movements as recorded by the accelerometer and gyroscopes embedded in the Samsung Galaxy S II smartphone. In particular, measures of linear acceleration appear to have the most predictive power (illustrated in Figure 2). With respect to potential confounders, it is worth noting again that this data set is extremely rich, recording literally hundreds of data points, much of them noise, but many of them *appearing* to have statistical significance toward predicting the subject's activity.

Future research may wish to incorporate additional activities (e.g., turning, running, jumping) to further refine the predictive model, or to create a family of related functions to infer these activities from the available data. As is, such a predictive model has applications in pedometers and other motion trackers, as well as geo-social platforms.

References:

1. UCI Machine Learning Repository: About. URL: <http://archive.ics.uci.edu/ml/about.html>. Accessed 3/6/2013.
2. UCI Machine Learning Repository: Human Activity Recognition Using Smartphones Data Set. URL: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones> Accessed 2/28/2013.
3. Coursera.org: Data Analysis Assignment #2. URL: https://class.coursera.org/dataanalysis-001/human_grading/view/courses/294/assessments/5/submissions. Accessed 2/28/2013.
4. R Core Team (2013). “The R Project for Statistical Computing.” URL: <http://www.r-project.org>. Accessed 2/28/2013.
5. Baker, Kirk. “Singular Value Decomposition Tutorial”. URL: http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf. Accessed 3/1/2013.
6. Hastie, Trevor, et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (PDF). URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf Accessed: 3/7/2013.
7. Howell, David C. *Fundamental Statistics for the Behavioral Sciences*. Wadsworth Cengage Learning, 2011.