Title: Comparison of Two Classifier Models Predicting With Smartphone Data

Introduction:

The task of this analysis is "to build a function that predicts what activity a subject is performing based on the quantitative measurements from the Samsung phone." [1] The Samsung phone data were provided by the Coursera Data Analysis class instructor [1] and originally collected by scientists at Smartlab. [2] The outcome variable for this predictor consists of six classifications and the predictor variables are quantitative. This analysis compares two methods for classification prediction: Multinomial Logistic Regression [3] and Random Forest [4].

Methods:

Data Collection

This dataset consists of sensor measurements collected by Samsung Galaxy S II smartphones worn on the waists of volunteer participants while performing one of six predetermined everyday activities (laying down, sitting, standing, walking, walking up stairs, walking down stairs). The sessions were videotaped and the activity performed was hand labeled by a technician. [5]

This experiment was conducted at Smartlab - Non Linear Complex Systems Laboratory in Genoa, Italy. The sensor measurements of the 3-axial linear acceleration and 3-axial angular velocity were recorded and filtered. The data also includes Jerk signals, signal magnitudes by Euclidean norm, and Fast Fourier Transform. [6]

The dataset for this analysis was downloaded on February 25, 2013 from the Amazon AWS repository for the Coursera Data Analysis class. [7] It was reported on the Coursera website that the data was slightly processed to make them easier to load into R. [1]

The dataset consists of 7352 samples gathered from 21 participants who are represented numerically. Of these samples those from participants 27, 28, 29 and 30 make up the test set. This test set consists of 1485 of the 7352 samples. The remaining 5867 samples make up the training set.

For the purposes of this analysis the training set was divided into two parts for model validation when appropriate. This division was made one time by taking a random sample of 30% of the training set and setting it aside as a validation set, making 4109 samples available for training the models and 1758 samples for validation and determining uncertainty. The sample was taken using the createDataPartition function in the caret R package. [8]

Exploratory Analysis

The first step in the exploratory analysis was data observation of the entire dataset. During this step it was noted that there were no missing values in any sample. Of the 563 variables 561 were observed to be quantitative and had been normalized and bounded in range of -1 to 1 [6].

There were two unusual features observed in the dataset. First, the outcome variable which represented the activity performed by the participant, was a character vector. A character vector for the outcome variable would not be allowed by the classifiers used in this analysis. To resolve this issue the outcome variable was transformed to a factor.

The second unusual feature observed was duplicate column names. There were 42 sets of 3 variables with the same column name. Plotting these variables revealed that there was indeed different data in each column. However, duplicate column names would not only not be allowed by the classifiers they would cause confusion when analyzing results. To resolve this issue the duplicate names were changed.

Statistical Modeling

Two models were used in this analysis: Multinomial Logistic Regression and Random Forest. For each, the training set consisted of 4109 samples (random 70% of the training data) and the validation set consisted of 1758 samples (remaining 30% of the training data). Throughout, the seed was set to 335 to aid reproducibility. The error rate is defined as the percentage of samples misclassified.

The predictors for the classifiers were the set or a subset of the quantitative sensor data collected from the smartphones plus a numeric identifier representing the volunteer participants.

1. Multinomial Logistic Regression

The first statistical model used in this analysis was multinomial logistic regression [3], a classifier used when the outcome variable is categorical, unordered, and has more than two possible values [3]. It uses maximum likelihood estimation to determine the probability for each classification and does not require the parametric assumptions of normality, linearity, or homoscedasticity [9].

For this analysis the multinom function in the R package nnet [10] was used for the estimation of the multinomial logistic regression model. This function was first used to fit the training set using the categorical variable representing the activity performed by the participant as the outcome variable and all remaining 562 variables were used as predictors. When predicting on the validation set the error and uncertainty [11] for this model was found to be 6.20% of the validation samples misclassified.

This model was then tuned with feature selection. Features to include were selected with the step function in the stats R package [12]. This stepwise algorithm was run in the forward direction on the training set with a minimum predictor variable of the measurement for the maximum value recorded for the time domain signal of the body acceleration on the x-axis. The result of feature selection was the 28 variables that produced an optimum Akaike information criterion (AIC) [13] value.

The multinomial logistic regression was run again with these 28 selected variables as the predictors in the multinom formula. The result on the validation set was an error and uncertainty of 3.13% of validation samples misclassified.

The next step in tuning the model was to investigate potential confounders, those variables with a correlation both with the outcome variable of activity performed and with other independent variables. When the purpose of the analysis is prediction confounders are an issue when they introduce information that decreases accuracy when predicting on unseen data. 14 of the 28 selected variables were found to be potential problematic confounders. To address these, variables that appeared to be providing very similar information were removed, with 11 variables remaining. This new formula was run on the training set with a result of 6.09% of validation samples misclassified. These confounding variables were concluded to be non-problematic as removing them decreased accuracy of the classifier.

2. Random Forest

The second statistical model used in this analysis was Random Forest, a classifier that works by generating many classification trees [14] and making predictions based on the mode vote of all the trees in the ensemble or “forest” [4].

This analysis used the randomForest function of the R package randomForest [15]. This model was run several times, with the formula consisting of the categorical variable for the activity performed as the outcome variable and the remaining 562 variables as predictors, to find the number of trees producing the smallest error and amount of uncertainty on the validation set. The optimum number of trees was found to be 750 and the error and uncertainty on the validation set was 1.8% of samples misclassified.

Feature selection was then used to try and improve upon this result. The function rfcv from the randomForest package was used and the optimum number of variables was found to be 140. Using the randomForest importance function the 140 variables with the highest gini importance [16] were selected. The variable with the highest gini importance was the minimum value recorded for the time domain signal of the gravity acceleration on the x-axis.

The model was run several times with this new formula to find the optimum number of trees (1100) and the error and uncertainty on the validation set was found to be 2.05% samples misclassified.

As with Multinomial Logistic Regression the potential problematic confounders were identified and removed and the model was run again with a result of 2.1% of validation samples misclassified. These confounding variables were found to be non-problematic as removing them appeared to have very little effect on overall accuracy of the classifier.

Results:

The best model from each of Multinomial Logistic Regression and Random Forest was used to predict the activity performed from the test set.

For Multinomial Logistic Regression the model with feature selection performed the best on the validation set. The resulting error and uncertainty on the test set was 3.57% of test samples misclassified.

For Random Forest the best performing model on the validation set was that which used all variables as predictors. The resulting error and uncertainty on the test set was 4.85% of test samples misclassified.

Conclusions:

In conclusion, both Multinomial Logistic Regression and Random Forest performed well on the Samsung dataset. Each was able to correctly predict which activity the participant was performing based on the quantitative sensor data more than 95% of the time.

One potential problem with this conclusion is the lack of inclusion of other statistical models that may perform better than either model analyzed here, such as support vector machines [Support vector machine http://en.wikipedia.org/wiki/Support_vector_machine]. If another model produced errors much lower than the two models analyzed here the conclusion that they “performed well” would require re-evaluation. A solution to this problem would be to analyze this model with the same methods applied here and compare the results.

It is interesting to note that the accuracy by activity performed differed between the two models. Multinomial Logistic Regression had a higher error rate for the activities walking and walking up. Random Forest had a higher error rate for sitting, standing, and walking down. Both classifiers performed equally well on laying down, with 100% accuracy for each. (Figure 1.)

A follow up to this analysis could include analyzing which predictors best predicted for the activities with the highest error rates in each model and giving those predictors more weight. Further, blending these models could increase the overall accuracy of prediction.

References:

[1] Coursera Data Analysis Class, Data Analysis Project 2

https://class.coursera.org/dataanalysis-001/human_grading/view/courses/294/assessments/5/submissions
[2] UCI Machine Learning Data Repository, Human Activity Recognition Using Smartphones Data Set http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
[3] Multinomial Logistic Regression
http://en.wikipedia.org/wiki/Multinomial_logistic_regression
[4] Random Forest
http://en.wikipedia.org/wiki/Random_forest
[5] features_info.txt included in UCIHARDataset.zip
http://archive.ics.uci.edu/ml/machine-learning-databases/00240/UCI%20HAR%20Dataset.zip
[6] README.txt included in UCIHARDataset.zip
http://archive.ics.uci.edu/ml/machine-learning-databases/00240/UCI%20HAR%20Dataset.zip
[7] Coursera Data Analysis class Amazon AWS repository for Samsung Data
https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda
[8] caret R Package Manual http://cran.r-project.org/web/packages/caret/caret.pdf
[9] Multinomial Logistic Regression
http://www.unt.edu/rss/class/Jon/Benchmarks/MLR_JDS_Aug2011.pdf
[10] R Package nnet http://cran.r-project.org/web/packages/nnet/nnet.pdf
[11] Structure of a Data Analysis, “Get a Measure of Uncertainty”
https://d19vezwu8eufl6.cloudfront.net/dataanalysis/structureOfADataAnalysis2.pdf
[12] step function in the R stats package
http://stat.ethz.ch/R-manual/R-devel/library/stats/html/step.html
[13] Akaike information criterion http://en.wikipedia.org/wiki/Akaike_information_criterion
[14] Decision tree learning http://en.wikipedia.org/wiki/Decision_tree_learning
[15] random Forest R package
http://cran.r-project.org/web/packages/randomForest/randomForest.pdf
[16] Random Forests, Gini Importance
http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#giniimp