Title: Samsung Phone Prediction Model

Introduction:

Human activity recognition is a large field of study. In this study, researchers develop models to predict the activity based [1] on data generated from inertial measurement systems. The researches quantify the position and activity of a person using smart phones. Smart phones contain an inertial measurement system to measure the change in position of the phone. Therefore, they can infer the human activity based on these measurements.

In this analysis, I will take a data set from UCI Machine Learning Repository that measure human activity. From this data, I will develop prediction models based on the results of the study [2].

Methods:

Data Collection

The data set was collected by Smartlab - Non Linear Complex Systems Laboratory. The sample is a set of data from 30 volunteers ages 19-48 years of age. Each person performed six activities: WALKING, WALKING_UP, WALKING_DOWN, SITTING, STANDING, LAYING wearing a smartphone (Samsung Galaxy S II) on their waist. Using the phones embedded inertial measurement systems 3-axial linear acceleration and 3-axial angular velocities were collected for each subject. The experiments were video-recorded and labeled manually [2].

Exploratory Analysis

The data set contains 7352 data rows of data with 563 columns with no missing values. The last two columns are the Subject and Activity. The subject is the number of the person that was tested. The outcome is measured by a factor called activity (table 1) and the rest of the data is measurement data from the Samsung phone. The data required pre-processing before it could analyze the header contained characters that required cleaning. The data was subset into a training set which used subjects 1,3,4, and 6 which contains 328 rows for the analysis. The test set which used subjects 27, 28, 29, and 30 and contains 371 rows of data for the verification of the model.

Activity	Laying	Sitting	Standing	Walking	Walking down	Walking up
Count	1407	1286	1374	1226	986	1073

Statistical Modeling

For modeling several packages where loaded into R [3]. A package called randomForest was downloaded to perform random forest analysis of the data [4]. This method was chosen because the outcome was a factor with continuous covariates. Random Forest is an excellent unsupervised method for developing models with low risk of over fitting.

Reproducibility

All analysis performed in this study is reproducible by using the R file samsungData.R. To reproduce the information in this file use the above file and the data record downloaded from the Cousrea class web page.

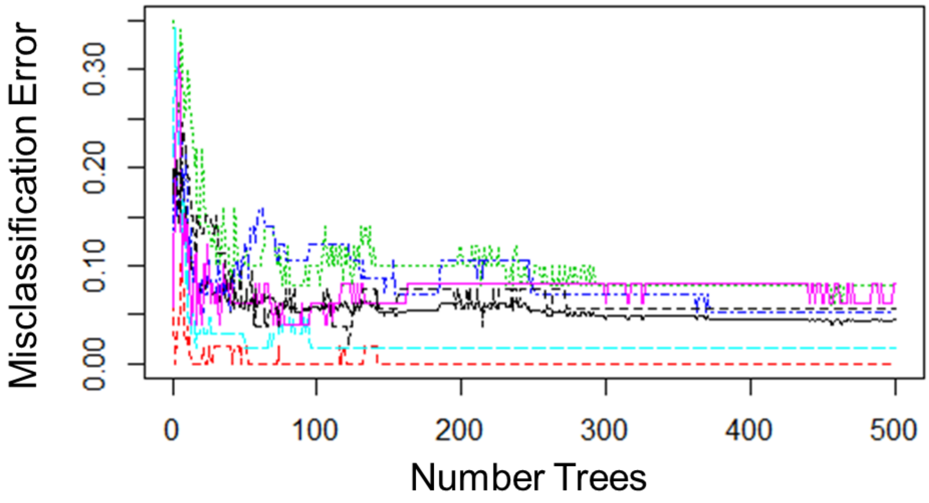
Results

Using the Random Forest function to analyze the data the following results were found. The initial model was run using with the tree parameter set to 100. This returned an error for the model of 4.53% using 23 predictors to minimize the error of the model. The confusion matrix below shows the misclassification of the training data versus the model. Figure (a) shows the misclassification error versus the number of trees in the model. The misclassification decreases as the number of trees increases.

Confusion Matrix	Laying	Sitting	Standing	Walking	Walking Down	Walking Up	Class Error
Laying	55	0	0	0	0	0	0.000
Sitting	0	46	4	0	0	0	0.080
Standing	0	4	53	0	0	0	0.070
Walking	0	0	0	63	1	0	0.016
Walking Down	0	0	0	0	47	2	0.041
Walking Up	0	0	0	0	2	50	0.057

Figure (a)

Misclassification Error vs. Number of Trees



The error reaches a minimum when the number of trees is equal to 500.

I wanted to see is the model could be optimized so Caret package [5] was downloaded. This package builds a wrapper around the Random Forest method to run 10 fold cross validation of the data set to optimize the model. The final model selected 33 predictors to minimize the error of the model. The error of the model was 4.57%. The confusion matrix was also outputted for the model. The confusion matrix still shows error versus the training data after the additional optimization step.

Mtry	Accuracy	Kappa	Accuracy SD	Kappa SD
2	0.908	0.89	0.0381	0.0459
33	0.96	0.952	0.033	0.0396
561	0.951	0.942	0.0379	0.0457

Confusion Matrix	Laying	Sitting	Standing	Walking	Walking Down	Walking Up	Class Error
Laying	55	0	0	0	0	0	0
Sitting	0	45	5	0	0	0	0.100
Standing	0	4	53	0	0	0	0.070
Walking	0	0	0	63	1	0	0.015
Walking Down	0	0	0	0	47	2	0.041
Walking Up	0	0	0	0	3	50	0.057

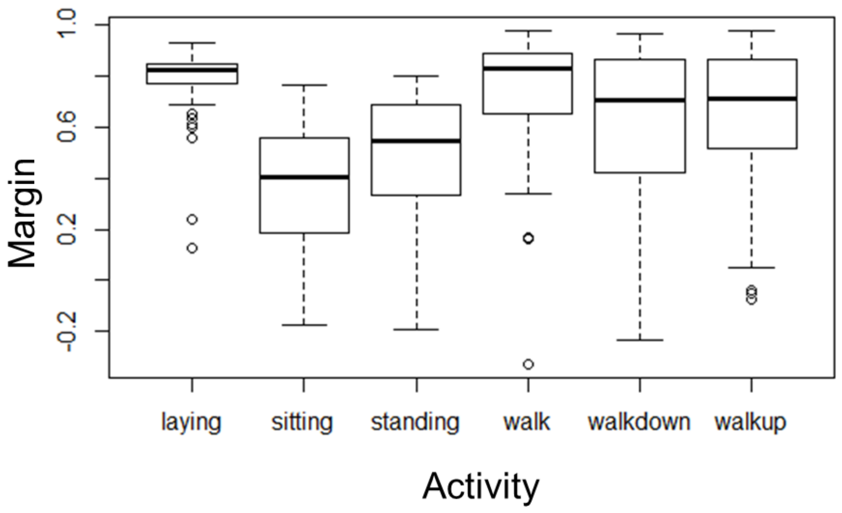
Conclusions:

Next the model was checked against the test set to verify the model against a new set of data. The chart below uses the test data set subjects 27, 28, 29, and 30 to validate the model. The chart is the portions chart for observed and predicted values. It shows the error of predicting each activity.

Predicted	Laying	Sitting	Standing	Walking	Walking Down	Walking Up
Laying	1	00.000	00.000	0.000	0.000	0.000
Sitting	0.000	0.871	0.129	0.000	0.000	0.000
Standing	0.000	0.137	0.863	0.000	0.000	0.000
Walking	0.000	0.000	0.000	1.000	0.000	0.000
Walking Down	0.000	0.000	0.000	0.019	0.904	0.077
Walking Up	0.000	0.000	0.000	0.0517	0.086	0.862

The model is very good at predicting LAYING and WALKING with 100% of the predicated equal to the observed values. The model predicts SITTING 87% of the time but 12.9% of the time predicts STANDING. The results are also very similar results for STANDING. When the subject is WALKING DOWN the model has a 1.9% error rate for WALKING and a 7.6% error rate for WALKING UP. Similar results are predicted when WALKING is observed the model predicts WALKING and WALKING DOWN. Figure (b) shows that negative margin negatively impacts model fit. For example laying and walking have few negative points and a very good fit to the model. The remaining activities contain negative margin therefore resulting in fitting errors.

Figure (b)
Margin per Activity



I tried to improve the model by adding additional data to improve the fit. The results were the same error rates of 4.5%. I also removed the confounding terms from the model it made both the initial fit and the test data fit worse giving an error rate of 8.2%, so these terms were left in the model as to improve the fit. To improve the model data collection methods need to be improved. Additional sampling resolution or noise reduction could be added to the Samsung phone to improve the data set.

References:

1. Rasekh, Amin, Chien-An Chen, and Yan Lu. "Human Activity Recognition using Smartphone."
2. Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto. “Smartlab - Non Linear Complex Systems Laboratory.” URL: http:www.smartlab.ws
3. R Core Team (2012). “R: A language and environment for statistical computing.” URL: http:www.R-projet.org
4. Liaw, A., & Wiener, M. (2002). “Classification and Regression.” by randomForest.R news, 2(3), 18-22.

5. Kuhn, M. (2008). “Building predictive models in R using the caret package.” Journal of Statistical Software, 28(5), 1-26.