Title: Model for to determine user activity from mobile phone sensors data.

Introduction:

Recent advancement in miniaturization and electronics technology made it possible to produce mobile phones equipped with GHz speed application and graphics processors. Together with higher speed processors, touch screens and 3G and 4G radio technologies these advanced phones known as smartphones are being fitted with different type of acceleration, magnetic and light sensors capable of detecting phone position, speed of movement in x,y,z direction and to collect statistical data from these sensors.

What become apparent was that the data from the movement and magnetic sensors might be used not only to determine phone position with reference to earth but also for prediction of current user activity. If such prediction could be done it could be used for variety of applications for example if we could detect that the phone has been slid to the pocket and the user started walking or running we could switch off all unused components to save battery power (e.g. user is unlikely to use the phone when walking or running). Another application would be a warning system where the user would receive a warning not to use smartphone whilst driving.

Smartphones are also very good tools to study activity detection based on sensors data as they are easily available of the shelf and one can write own software to retrieve all available sensors information.

The data used in this study was gathered just for this purpose as an input for more general studies in machine learning field.

The purpose of this analysis is to create predictive model that will be able to predict with reasonable accuracy what is the subject activity based on input from smartphone sensors.

Methods:

Data Collection

For our analysis we used data consisting of sample of 7532 observations of measurements of 561 different variables generated from raw data obtained from Samsung Galaxy S3 smartphone accelerometer and gyroscope and described in [1]. The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist [1].
The data were downloaded from Amazon Web Services hosting service [2] on 8th February 2013 using the R programming language [3].

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, and (3) determine the variables that can be used for our final predictive model.

During exploratory analysis it was found that there were no missing values. The only data transformation necessary was to coerce “activity” variable to a factor class. During modeling we have however identified problem with some repeated variable names and non-alphanumeric characters in most of the names. This was cured by coercing raw data Frame to modified data Frame using following line of code in R:

```
samsungData <- data.frame(samsungData)
```

Statistical Modeling

For statistical modeling we have split data samples in two non-overlaping sets: training set and test set. It was imposed that the training set will be subset of full set containing observations for subjects 27, 28, 29, and 30. We have decided to use all remaining data as a training set because Random Forest package [4], [5] was used for final model construction and random forest algorithm is cross-validating the data itself using bootstrap approach [6] hence there is no need for additional validation set.

Misclassification error defined as a ratio of misclassified samples to all samples in the set was used as an error metric for measurement of model fitness. After different trials of modeling the tests set with single trees using “tree” [7] and “rpart” [7] packages we have decided to use random Forest predictor as giving best smallest misclassification error on the training set.

Results:

Due to complexity of random forest models we will not present whole list of decision nodes of all trees of our end model here due to limited space allowed. Following line of code was used for random forest generation and can be used to reproduce model in R language:

```
trainData <- samsungData[!is.element(samsungData$subject, c(27,28,29,30)),]
testData <- samsungData[is.element(samsungData$subject, c(27,28,29,30)),]
library(randomForest)
set.seed(33833)
forest1 <- randomForest(as.factor(activity) ~ .,data=trainData, prox=TRUE)
```

First two lines of above code split full data set into two subsets: “trainData” – containing observations for all subjects apart of 27, 28, 29, and 30. The remaining observations “testData” were used as a testing set and contained only observations for subjects 27, 28, 29, and 30.

We have used default “randomForest” package parameters. This has resulted in generating random forest model `forest1` with 500 trees and 23 variables used at each tree split.

The resulting model trained on the “trainSet” gave us misclassification error of OOB= 1.65% (OOB-> Out Of Bag error in randomForest package). In Table 1 below confusion matrix of the generated model with all 500 trees used is shown together with classification error for each of the activity classes:

Table 1. Confusion matrix and classification error for the predicted values obtained by applying “forest1” random forest model to the training data set. Values represent number of classifications in the predicted data. Row refers to classification by the model and column the actual classification in the data.

```
Confusion matrix:
          laying sitting standing walk walkdown walkup class.error
laying     1114       0        0    0        0      0 0.000000000
sitting       0     989       32    0        0      1 0.032289628
standing      0      35     1056    0        0      0 0.032080660
walk          0       0        0  983        7      7 0.014042126
walkdown      0       0        0    5      777      4 0.011450382
walkup        0       0        0    1        5    851 0.00700116
```

In attached Figure 1 on the left hand graph plots of classification errors for all classes of activities is shown together with total misclassification error (OOB) as a function of number of trees used in the model. It can be seen that different activities have different sensitivity of classification error to number of trees used. For example ‘laying’ activity is easiest to predict and around 50 trees would be sufficient to get 0% classification errors in training data set. In contrary “standing” activity classification continues to improve up to 500 trees and it might benefit even more with higher number of trees. The overall classification error OOB reaches minimum for around 200 trees and doesn’t seem to improve with higher number of trees used. Contrary, “sitting” and “walkdown” activities seem to have higher classification errors for number of trees higher than around 200.

In attached Figure 1 on the right hand graph the value of a mean decrease of GINI coefficient [8] is shown for the 30 most important variables as classified by “randomForest” algorithm. Gini criterion is the splitting criterion used in random forest (see p. 11 in [5]). Mean decrease of GINI coefficient is one of the metrics used by “randomForest” algorithm to classify importance of each variable in contribution to the final model. It is apparent from the plot that there are 9 covariates listed in Table 2 that have particularly high mean decrease of GINI coefficient value above 120. For all other variables the mean decrease of GINI coefficient values are below 80.

Table 2. 9 Most important “forest1” model covariates having highest Mean Decrease in Gini coefficient above 120.

```
"tGravityAcc.max...X"
"tGravityAcc.mean...X"
"angle.Y.gravityMean."
"tGravityAcc.mean...Y"
"angle.X.gravityMean."
"tGravityAcc.energy...X"
"tGravityAcc.max...Y"
"tGravityAcc.min...X"
 "tGravityAcc.min...Y"
```

The resulting model `forest1` was finally used to predict values of test dataset. The misclassification error err=4.8% was obtained by testing our model on test data set.
The Table 3 below shows confusion matrix for the predicted data on test data set together with misclassification error calculated for each of the activity classes.

Table 3. Confusion matrix and classification error for the predicted values obtained by applying "forest1" random forest model to the test data set. Values represent number of classifications in the predicted data. Row refers to classification by the model and column the actual classification in the data.

	laying	sitting	standing	walk	walkdown	walkup	class.error
laying	293	0	0	0	0	0	0.000000000
sitting	0	226	26	0	0	0	0.143939394
standing	0	38	257	0	0	0	0.091872792
walk	0	0	0	228	2	1	0.004366812
walkdown	0	0	0	0	194	0	0.030000000
walkup	0	0	0	1	4	215	0.004629630

From Table 3 it is apparent that sitting activity has highest classification error of 14.4% and standing being activity with next highest misclassification error of 9.2%.

Conclusions:

We have applied random forest statistical predictive modeling approach to create predictive model for detecting of subject activity based on the 561 covariates obtained from raw data obtained from Samsung Galaxy S3 smartphone accelerometer and gyroscope. By applying our final model to the test data set we have obtained activity classification error of 4.8% which can be considered good result for very complex data. In fact the difference in error rate obtained for test set (4.8%) and train set (1.65%) might suggest that our model was in fact over fitted to the train set.

Our model was based on all 561 variables available in data set and does not explain any physical relationships between measured variables and activity. It is however good starting point for further improvements where we could try to create models based on smaller number of variables starting with 9 variables from Table 2 that have showed to have most important contribution to model classification decisions. Having only 9 variables to choose from linear models could be applied and random forests with smaller number of variables. Such models could hopefully limit suspected over fitting to training data set and possibly give better error rates for new datasets.

References

[1] D. A. A. G. L. O. Jorge L. Reyes-Ortiz, "Human Activity Recognition Using Smartphones Data Set," DITEN - Università degli Studi di Genova, Genoa I-16145, Italy. , Dec 2012. [Online]. Available:

http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones#. [Accessed 3 March 2013].

[2] "Human Activity Recognition Using Smartphones Data Set," [Online]. Available: https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda. [Accessed 3 March 2013].

[3] R Core Team, "R: A language and environment for statistical computing," 2012. [Online]. Available: http://www.R-project.org. [Accessed 8 February 2013].

[4] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32., 2001.

[5] L. Breiman, 2002. [Online]. Available: http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf.

[6] "Bootstrapping (statistics)," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Bootstrapping_(statistics). [Accessed 10 March 2013].

[7] F. J. H. O. R. A. a. S. C. J. Breiman L., Classification and Regression Trees, Wadsworth, 1984.

[8] "Gini coefficient," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Gini_coefficient. [Accessed 10 March 2013].

[9] “Lending Club Page,” [Online]. Available: https://www.lendingclug.com/home.action.

[10] W. community, "Credit score in the United States," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Credit_score_in_the_United_States. [Accessed 17 February 2013].

[11] "Codebook for the available dataset," [Online]. Available: https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf. [Accessed 8 February 2013].