A predictive analysis of Accelerometer data to discern type of activity

Introduction:

An accelerometer is a device that is used to measure the actual acceleration of the body on which it is placed [1], where actual acceleration is defined as the weight experienced by a small test mass present on the device. In modern times, these devices have become an integral part of hand held cellular devices. An accelerometer records the acceleration in the x, y and z directions, which can further be divided into a number of sub-fields such as the body acceleration, acceleration frequency, acceleration jerk etc. [2]. The present dataset that was analyzed consisted of numerous such acceleration measurements and the corresponding activity that was being carried out with each of the measurements namely, **sitting, laying, standing, walking, walking up or walking down**. The aim of the present analysis was to develop a classifier that could be used to predict the type of activity being undertaken by an individual, being given the acceleration data from the Smartphone of the individual. The applications of this technique could be in the fitness sector, in healthcare, in advertising and in other areas too.

In the present analysis a large dataset was given with accelerometer measurements for 21 individuals performing all the six tasks stated above. The total number of rows of data was 7352 and there were 561 types of accelerations data recorded per observation. To conduct an unbiased analysis, the data set was divided into two parts, the test set and training set. Initially, exploratory analysis was carried out on the training set to recognize the fields which could distinguish the data well. Then, based on the exploratory analysis, statistical analysis was carried out on the data using generalized linear models, logistic regression and a random forest model. The result of the analysis showed that in this case, the multiclass logistic approach [3] showed the best results among all three approaches with an average F-value of 0.9783832 on the test data.

Methods:

Data Collection

The data was downloaded from the following URL: https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda on March 6, 2013 using the R programming language [4]. The data was taken from Smartlab's experimental data at the DITEN University, Geneva [2].

Exploratory Analysis

The first step of the exploratory analysis was pre-processing the raw data to make it easier to analyze. Firstly, the data was checked for missing and infinite values. No such data was found. Then the columns were renamed with generic names to make the accessing of individual columns simpler. The activity data was converted to factor form to make it simpler to analyze this data using R modules. Then the data was divided into the training and the test sets. The test set is the one on which the accuracy of the predictive algorithm is finally tested and in this case it

contained a randomly selected 20% of the overall data set. The training set is the one on which the analyses are carried out, i.e. where the predictive algorithm is trained. This consisted of the remaining 80% of the given data. A few exploratory graphs were plotted with some of the columns in the data set versus the data index. These were scatter plots showing the column acceleration data on the y-axis and the index number of the data or another acceleration parameter on the x-axis, colored according to the activity being carried out by the individual. These graphs showed, as also shown in Prof. Leek's videos, that some fields, such as *tBodyAcc-mean()* in x and y directions did not help separate the data, while *tBodyAcc-max()* in the x and y directions separated the sedentary activities were from those involving motion. Fig. 1 (a) attached separately shows one such graph, plotting tGravity Acceleration Max in X and Y directions, colored by the type of data. This graph shows that these parameters can easily separate between the sedentary activities like Laying and Standing and the other activities. I tried *glm()* with the Gaussian linear modeling on some variables against the output data but did not get promising results. This is because the problem was inherently a multi-class classification problem, even though it could be coerced into a regression-based solution.

Metric for Model Selection

Since the data consists of 6 unevenly distributed labels, and the problem is a multi-class classification problem, the best metric would be to look at weighted precision and recall measures for each of the classes. The F-measure, which is the harmonic mean of both these measures, thus is a good indicator of precision and recall. The **weighted average F-measure** was thus used as the metric for model selection.

Statistical Modeling

Initially, the problem was attempted as a regression problem, by assigning numbers to each label and attempting to predict those numbers using generalized linear models. The results appeared promising. Using 10 fold cross-validation, we estimated, that the best model, using model selection was the one, which used all input variables as training data. This implied that the approach was not able to account for all the variation in the data. Since this approach was not really suitable for multi-class classification, the new model that was used was an L2-regularized logistic regression model. Before applying this approach, the data was centered and scaled. The results of this approach were significantly better than the glm approach. The cost parameter for this approach was calculated using a heuristic given in [3]. The third approach attempted was a random forest. As per the lectures, the Random Forest method is essentially a bagged tree approach but includes some additional steps, which optimize the process and also make it more robust, though it might bias the prediction analysis to the training data. This approach picks a fraction of the data and a fraction of the input variables and creates a decision tree. This process is repeated multiple times and the large number of tress that are made (500 in the present analysis) all vote in the final prediction to give accurate predictions.

Results:

The following are the results of the various models that were applied. Initially linear models were applied, but the result from their outputs was not too accurate. Following this two approaches were applied: L2 Regularized Logistic Regression and Random Forest.

1. Generalized Linear Model:

Ten fold cross validation was used to calculate the best model, where the features used to predict on the test data was varied. The choice of the features to build the models with was on the basis of the features, which explained the maximum variance in the model. Based on this study, the result was that the best model was the one considering all the features, with CV F-Measure of 0.9000. On testing this model on test set, an F-measure of 0.9054 was obtained.

2. Logistic Regression

An L2-Regularized Logistic regression model was used to build a multi-class classifier. The data was scaled and centered and the cost parameter was calculated using the heuristicC() function given in [3]. On the training data, the cross validation average F-measure was found to be 0.9767 and on the test data, 0.9784.

3. Random Forest:

The random forest method was applied on the training set, considering all 561 columns of accelerometer data. Predictions were calculated for the training set and they were stored both as the predicted outcome activity and as a matrix containing the probabilities of each of the rows of data being a type of activity. The importance function applied to the random forest model gave the relative importance of each of the columns of data in terms of the magnitude of the mean decrease of the Gini coefficient associated with the model. On the basis of this, the important parameters were angle(X, gravityMean), angle(Y, gravityMean), fBodyAccMag-energy(), fBodyGyro-maxInds-Z, fBodyAccJerk-bandsEnergy()-1,16, tGravityAcc-arCoeff()-Y,2, tGravityAcc-energy()-Y, tGravityAcc-energy()-X", tGravityAcc-min()-Y, tGravityAcc-min()-X, tGravityAcc-max()-Y, tGravityAcc-max()-X, tGravityAcc-mean()-Y, tGravityAcc-mean()-X and tBodyAcc-correlation()-X,Y. One of the trees of the random forest has been plotted in Fig. 1 (b). The tree shows that at the initial splits, Laying, Sitting and Standing get separated out while it is more difficult to separate Walk, Walk up and Walk down.

On the test data, the cross validation average F-measure was found to be 0.9744.

4. Application of the three models on the test set

All three models were applied on the test set finally, to see which algorithm was superior, when applied to data which was independent of the training set. These results are tabulated below.

Table 1: F-measure when predictive algorithms were applied to the test set

Model	F-Measure
Generalized Linear Model	0.9054
Logistic Regression	0.9784
Random Forest	0.9744

Based on these results the most accurate results were given by the **Logistic Regression method**. Fig. 1 (c) shows the test data plotted with the same axes as Fig. 1 (a), colored by the activity. The size of points is enlarged when the point is incorrectly classified. It can be observed that the misclassifications are mainly in separating Walk, Walk up and Walk down, as was expected from data observation during expository analysis.

Conclusion:

The present report carried out a predictive analysis of using accelerometer data from a Smart Phone to predict the type of activity being undertaken by the individual whose data is being recorded. The appropriate method for the prediction study was L2-Regularized Multi-Class Logistic Regression. This approach, when applied to the test set, gave an F-measure of 0.9784.

Other methods of statistical analysis might have made the analysis more robust by finding the predictive variables, which had a greater impact on the output. These methods include SVD decomposition and Support Vector Machine (SVM) methods. SVD could reduce the dimensions of the problem while SVM could have been another classifier in the combined approach. Another possible step that might have assisted the analysis was to look into the meanings of each of the variables and the possible impacts they might have on the output to perform feature engineering.

Reference

[1] Wikipedia "Accelerometer" Webpage. URL: http://en.wikipedia.org/wiki/Accelerometer. Accessed on: 03/09/13.

[2] "Human Activity Recognition Using Smartphones Dataset" Webpage. URL: https://www.dropbox.com/s/rrsm7nv5j7y3rsi/analysis2info.pdf. Accessed on 03/09/13.

[3] "Package 'LiblineaR'" URL: http://cran.r-project.org/web/packages/LiblineaR/LiblineaR.pdf. Accessed on 03/09/13.

[4] R Core Team (2012). "R: A language and environment for statistical computing." URL: http://www.R-project.org.

[5] “Pragmatic Programming Techniques” Webpage. URL: http://horicky.blogspot.com.au/2012/06/predictive-analytics-decision-tree-and.html. Accessed on 03/09/13.