Samsung Data Analysis and Activity Prediction

Introduction

The purpose of this analysis is to build a predictive model for the activity that a subject is performing after analyzing the data collected from Samsung smart phone. Android Smartphones like Samsung Galaxy S II come up with embedded accelerometers and Gyroscopes that take quantitative measurement of various activities.

Accelerometers and Gyroscopes

An accelerometer measures movement along a particular axis (up and down, left and right, forward and backward). Some Android smartphone like the Samsung Galaxy S II accelerometers are used to gather this data, it can measure movement along all three of the X, Y and Z axes. The measurements can have positive and negative values to indicate the direction of movement along the axis.

A gyroscope gives the orientation of the phone and how quickly it is changing orientation. For example, if your phone is laying face up on your desk, the gyroscope will tell you it is oriented that way. If you flip it over, the gyroscope will tell you not only that it was flipped over, but how quickly and in which direction it was rotated.

Using the data log collected from the smartphone in intelligent way is a challenge. Using the data collected for intelligent prediction of the user activity could be very useful.

Processing Methods

Data Collection

The dataset provided to us can be downloaded from https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda . The dataset is a part of UCI Machine Learning Repository dataset for activity and the full unprocessed dataset can be downloaded from http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

For a detailed explanation of the information collected one can refer to the explanations provided in above link.

Data Variable Explanation

The dataset altogether has 7352 observations of 563 variables. The Dataset consists of datasets of 30 volunteers within an age of 19-48 years. Each person performed six activities: walking, WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING wearing a smartphone (Samsung Galaxy SII) on the waist. The dataset consists of 3 axial linear acceleration and 3-axial angular velocity captured using embedded accelerometer and gyroscope.

Each record in the dataset following information are provided:

- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.
- Triaxial Angular velocity from the gyroscope.

- A 561-feature vector with time and frequency domain variables.
- Its activity label.
- An identifier of the subject who carried out the experiment.

Missing Data

No missing data are reported in the dataset.

Data Processing

The data contained altogether 84 duplicated variables. Since each of the specific variable contributed to the result, so removing them from the dataset couldn't be justified and we made every variable in the dataset to a unique value using following R command.

Data <- data.frame(samsungData)

Training Set and Test Set data construction

Based on the subject whose records were taken, the training set and test set data were created using the following R-Code.

#Training Set construction based on subject 1, 3, 5, 6

Trainset <- Data [Data$subject %in% c(1,3,5,6),]

#Test Set Construction based on subject 27, 28, 29, 30

Testset <- Data[Data$subject %in% c(27,28,29,30),]

Based on the values that are most affecting the classification we took following variables to construct a prediction model using Tree. For the construction of tree trainset was used. The values are:-

```
[1] "tBodyAcc.max...X"                  "tGravityAcc.mean...X"
 [3] "tGravityAcc.max...Y"               "tGravityAcc.mean...Y"
 [5] "tGravityAcc.mean...Z"              "tGravityAcc.min...Y"
 [7] "fBodyAccJerk.bandsEnergy...17.32" "fBodyAccMag.std.."
 [9] "fBodyAcc.bandsEnergy...17.24.1"   "tBodyAccJerk.max...X"
```

Results

A classification tree was constructed based on the above values mentioned. The above tree had altogether following values:-

```
Number of terminal nodes:  12
Residual mean deviance:  0.1573 = 204.9 / 1303
Misclassification error rate: 0.02357 = 31 / 1315
```

The tree was pruned for best 6 result. The pruned tree had altogether 95 misclassification errors. Various other values for the pruned tree are listed below:-

```
Number of terminal nodes:  6
Residual mean deviance:  0.4939 = 646.5 / 1309
Misclassification error rate: 0.07224 = 95 / 1315
```

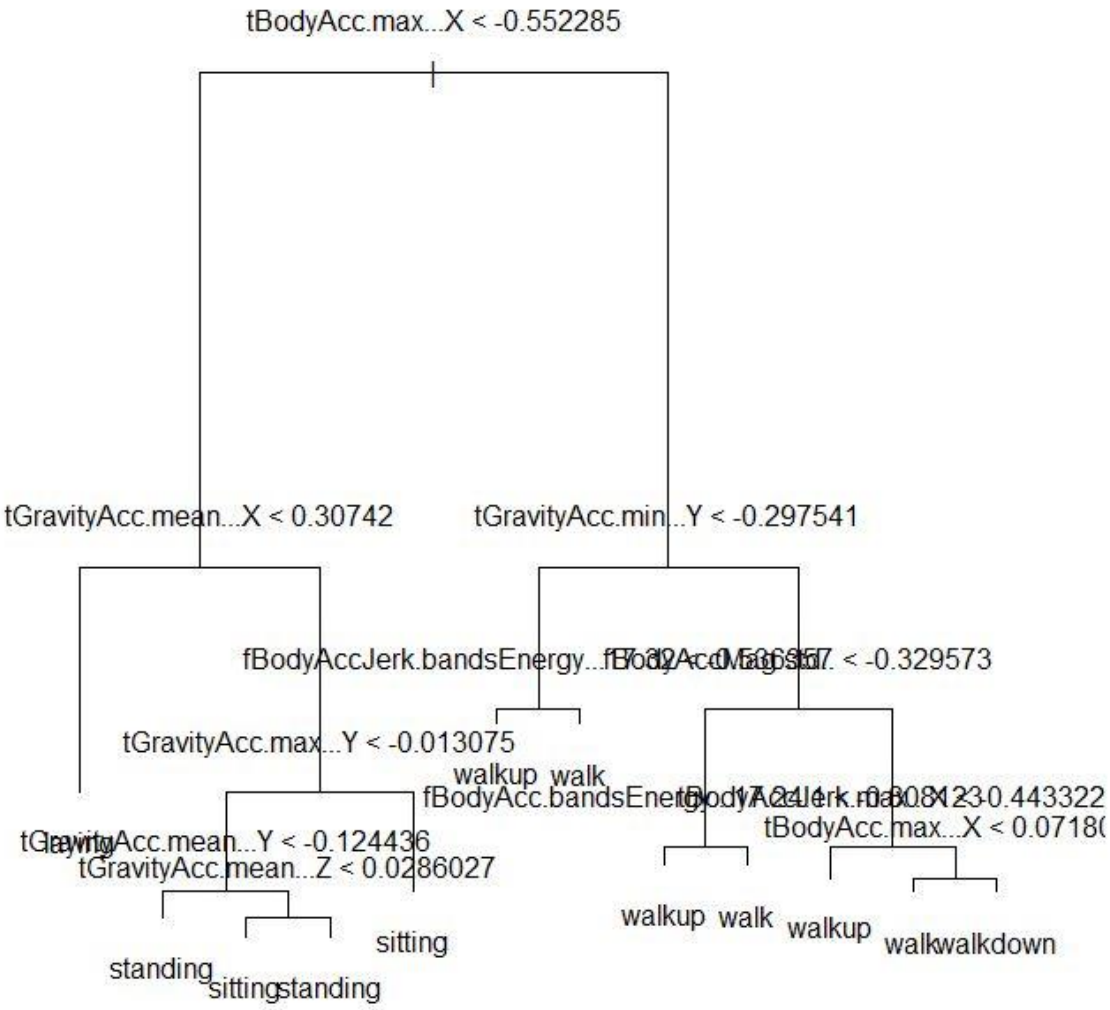
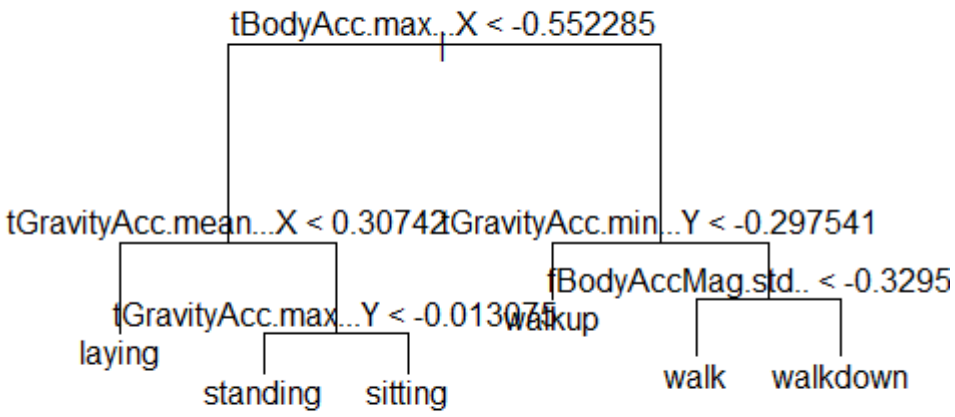


Figure Classification tree after prunning



Finally, after the construction of the prediction tree we fired the test data set to the classification tree.

On firing the test data following data summary were predicted:

```
laying  sitting standing     walk walkdown   walkup
   293      243      304      203      237      205
```

The data had following properties:

```
          laying sitting standing walk walkdown walkup
laying       293       0        0    0        0      0
sitting        0     176       88    0        0      0
standing       0      67      216    0        0      0
walk           0       0        0  190        4     35
walkdown       0       0        0    4      184     12
walkup         0       0        0    9       49    158
```

The above figure being of the prediction study design error test. In the first case of laying all 293 cases were predicted correctly for the test set by the prediction model developed by us using the training set. In the second set of sitting, 176 out of 264 were correctly predicted while the 88 samples were incorrectly predicted to be were made false prediction.

Similar is the case for walk, walkdown, walkup, standing, sitting. The deviation from the actual result was due to a fact that we took few variables to construct a decision tree. On taking

proper number of variables for the construction of the decision tree we could see a better result.

Conclusion

Hence, our analysis constructs a decision tree that could be applied to any data that is collected from any person's activity log and predict what the subject is doing. Decision tree of such nature is very useful for the prediction and analysis of the data.

References

[1] "Wikipedia Decision Tree," [Online]. Available: http://en.wikipedia.org/wiki/Decision_tree.

[2] "Classification trees in R," [Online]. Available: http://www.r-bloggers.com/classification-trees/.

[3] U. M. L. S. S. Data. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.