Human Activity Recognition Using Smartphone Sensors

Introduction

The Samsung Galaxy S II [1] is a touchscreen-enabled, slate-format Android smartphone designed, developed, and marketed by Samsung Electronics. The standard version has dimensions 125.3 x 66.1 x 8.49 mm and weights 116g.

The database is built from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a waist-mounted Samsung Galaxy S II smartphone with embedded inertial sensors from the UCI Machine Learning repository[2][3].

The purpose of this analysis is to build a function that predicts what activity a subject is performing based on the quantitative measurements from the smartphone.

Methods

Data collection

The data are the Samsung activity data available from the Coursera Data Analysis course website [4]. These data were slightly processed to make them easier to load into R. The raw data was obtained in the UCI ML Repository [2]. The experiments were 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. Using its embedded accelerometer and gyroscope, 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz were capture. The experiments were video-recorded to label the data manually. The obtained dataset was randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data[3].

For each record in the dataset it is 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.

Attributes | Value ——————— | ——————— Dataset characteristics | Multivariate, Time-Series Number of instances | 7352 Number of attributes | 561 Date donated | 2012-12-10 Associated tasks | Classification, Clustering Source | See references[5]

Exploratory analysis and preprocessing

All of the columns of the data set (735x563 matrix; except the last two) represents one measurement from the Samsung phone. The variable subject indicates which subject was performing the tasks when the measurements were taken. The variable activity tells what activity they were performing.
For the task of building a classifier, the training set was designed to be the data from subjects 1, 3, 5, and 6 (1315x561 matrix). The test set was the data from subjects 27, 28, 29, and 30 (1485x561 matrix). The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56s and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain.

Exploratory analysis was performed inspecting the tables and creating bivariate conditinal plots, histograms, scatterplot matrices, boxplots ad interaction plots. No missing values were found. However, subjects 9 and 10 values were missing. The activity attribute was coerced to be a ordered factor (with 6 levels). No clear candidate variables for performing statistical analyses were found. Therefore, and given the large number of variables, it was decided that dimension reduction was necessary.

Thus a singular value decomposition was performed on the training set (1315x561). The singular vectors and singualr values were put in decreasing order in left singular vectors, right singular vectors, and a diagonal singular values matrices. Further exploratory analsys were performed on this new transformed data set, focusing naturally on the first singular vectors. This transformation allowed to see clearly different patterns identifying the activities levels, as can be observed in Figure for the first five left singular vectors.

Statistical modeling

Based on the insights earned from the exploratory analysis and the dimension reduction, a predictive tree model was fit to a formula with the first 28 left singular vectors as predictive variables and the activities vector as dependent variable. The tree was grown using recursive partitioning using the activities and choosing splits from the left singular vectors. The splits which maximized the reduction in impurity by the deviance criterion were chosen.
Moreover, a 10-fold cross validation was performed to find out the deviance as a function of a cost-complexity parameter by pruning the tree. Both deviance and the cost-complexity parameter increased with every pruning, thus no pruning was actually performed.

Reproducibility

The code for preprocessing data and for generating tables and plots of the exploratory analysis and statistical modeling analysis was embedded in a R Markdown document which contains this same text. However, only this text and a figure are being actually submitted, as required in the assignment description.

Results

The classification tree fit resulted in a tree with 12 terminals resulting from the splitting of only 8 left singular vector values. These singular vectors were, in order of size of their corresponding singular values: 2, 3, 4, 5, 6, 7, 13, 14. Interestingly, in order of predictive power, the order was: 2, 4, 13, 3, 6, 5, 7, 14. No other left singular vectors were found to be significantly predictive.

In the following table, the classification rates for the trees fit for the first 2, 5, 9, 28 and 99 singular values are shown:

Num. of left SVs | Training set CR | Test set CR ——————|——————|——— 2 | 0.5535 | 0.6183 5 | 0.7165 | 0.7848 9 | 0.7279 | 0.8586 28 | 0.7562 | 0.9049 99 | 0.7562 | 0.9049

Conclusions

The fact that pruning was unnecessary as dictated by the 10-fold cross-validation as a function of cost-complexity parameter indicates that the found left singular vectors were high quality predictors not needing for further refinement. The classification tree made with splits from the 8 left singular vectors 2, 4, 13, 3, 6, 5, 7, 14 thus showed a classification rate of 90.49% in the training set and of 75.62% in the test set. Taking into account that the benchmark for this 6-level classification is 16.67%, it can be readily concluded that this classification tree based on smartphone measurement variables was highly predictive of the activities the human subjects were doing while carrying the smartphone.

References

1. Samsung Webpage
2. University of California Irvine Machine Learning Repository Webage. URL: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. Accessed 2013
3. Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International

Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012

4. Data Analysis by Jeffrey Leek Coursera class Website. URL: https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda. Accessed 2013

5. Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto. Smartlab - Non Linear Complex Systems Laboratory. DITEN - UniversitÃ degli Studi di Genova, Genoa I-16145, Italy. activityrecognition '@' smartlab.ws, www.smartlab.ws