Title: Statistical models for predicting activities with a smartphone.

Introduction:

Smartphones is turning into one of the biggest consumer products industry. Fairly soon there will be more smartphones than people. Since it's creation, this phones has been incorporating multiple kind of devices in order to improve the user experience with the camera, games and much more: GPS, WiFi, touchscreen and gyroscope, accelerometer.

This study centers on both gyroscope and accelerometer sensors for activity recognition. As more and more people uses and carries smartphones on a daily basis, an accurated activity recognition model could be the starting point for building apps for eHeath, assisted living or even sports tracking.

Methods:

Data Collection

For our analysis we used a sample of 21 persons performing a set of six activities: laying, standing, sitting, walking, walking upstairs and walking downstairs. A total of 7352 observations of 561 features with time and frequency domain variables describing triaxial acceleration from the accelerometer and triaxial angular velocity from the gyroscope.

This dataset was downloaded from UCI Machine Learning Repository[1] downloaded on Thursday 07 Mar 2013 using the R programming language[2].

Exploratory Analysis

A subject vs. activity exploratory analysis shows a good amount of all activity for each subject. It has sense to separate the data into training and test set by subject.

After some cleaning, different prediction models were applied to the training set. We've chosen our model by comparing how accurated each model predicts the activity on the test set.

Statistical Modeling

Logistic regression[3], tree modeling[4] and random forest technique[5] were applied, using activity as the outcome dependent variable and the rest of variables as independent. Accuracy was calculated as true/false positive classification rate on the test set.

Reproducibility

Performed with R programming language, the analysis were structured using ProjectTemplate package[6] and it has been published on a Github repository[7]. The *src/reproducible_code.html* file reproduces the whole exploratory analysis and results.

Results:

Running a binomial logistic regression on the training set of activity vs. the rest of variables a poor prediction shows poor accuracy with a classification rate of ~13.7%. No independent variable shows statistically significant. As the outcome variable only takes 6 values (the tracked activities), maybe a decision tree model could fit better, with different activities as leafs.

The classification tree, implemented by r-project tree library, is applied on the training set with a result of 12 leaves, 31/1315 misclassification errors and 205/1300 residual mean deviance (0.82% accuracy). Ten independent variables predicts the outcome as you can see at Fig 1. Time domain variables are more significant for predicting no movement activities (laying, standing, sitting), and both frequency and time domain variables are useful for movement activities (walk, walk upstairs, walk downstairs). tBodyAcc-max()-X determines whether the activity a movement activity or not.
In order to check if the model is overfitting a cross-validation test was performed showing there wasn't a significant deviance for a nine leafs tree. Pruning the tree to nine leafs gave us a 0.78% of accuracy predicting activities on the test set.

Finally, a random forest of 500 tree, implemented by r-project randomForest library, was performed. The prediction results on training test reached a perfect classification rate, and an accuracy of 0.93% predicting activities on the test set. Again time domain variables are given more importance than frequency time variables as you can see on Fig.1 along with the rest of the predictors revealed.

Conclusions:

Both tree and random forest generates a good model for predicting activities based on data collected from a gyroscope and an accelerometer. Time domain variables for Gravity and body data defines whether the subject is moving or not. The table below shows how good the prediction is for the test set with both models, broken down by activities:

Tree prediction model classification rates (test set)

Prediction	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	11	47	158

Random forest model classification rates (test set)

Prediction	laying	sitting	standing	walk	walkdown	walkup
laying	293	0	0	0	0	0
sitting	0	224	40	0	0	0
standing	0	29	254	0	0	0
walk	0	0	0	223	6	0
walkdown	0	0	0	1	194	5
walkup	0	0	8	0	14	194

In spite of the results some other data analysis needs to be performed. The set of activities appears to be simple enough to be detected with less variables than those taken as predictors by tree and random forest models

References

1. http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
2. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: http://www.R-project.org
3. Seber, George AF, and Alan J. Lee. *Linear regression analysis*. Vol. 936. Wiley, 2012.
4. http://cran.r-project.org/web/packages/tree/index.html
5. http://cran.r-project.org/web/packages/randomForest/index.html
6. ProjectTemplate Page. URL: http://projecttemplate.net.
7. Analysis project page at Github. URL: https://github.com/fontanon/samsung-activities-analysis