Human Activity Recognition with Smartphones

PREDICTIVE ANALYTICS COM SCI X 450.7
(SPRING 2020)

PROFESSOR: ALFONSO BERUMEN

RUOHAN DANG

1. Problem statement

Human Activity Recognition (HAR) is the problem of identifying a physical activity carried out by an individual dependent on a trace of movement within a certain environment. Smartphones results in behemoth amount of data being collected including motion, location, physiological signals and environmental information.
As stated above, this paper addresses the HAR problem. With the availability of data, analyses can be done to better recognize, classify, cluster and predict what human activities are carried out (e.g. stand, sit, lie, walk) for further decision making.

2. Data overview

The dataset can be freely downloaded from the UCI. The raw data is not available, but the preprocessed version of the dataset is made publicly available to carry out experiments and it was converted in a CSV file. It was built from the recordings of 30 subjects who were within the age bracket of 19 - 48 years performing basic activities and postural transitions while carrying a waist-mounted smartphone with embedded inertial sensors. The dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

The training dataset has a total of 7352 observations and the test dataset has 2947 observations.

The dataset used in this paper had already been pre-processed. The features selected for this database come from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ. These time domain signals (prefix 't' to denote time) were captured at a constant rate of 50 Hz. In particular, the SBHAR data generated around 5-hours of experimental data [3], and was also pre-processed with noise filters sampled by fixed-width sliding windows (2.56 sec with 50% overlap, i.e.128 readings/ window). Other data transformations were also applied including the calculation of Jerk signals from time, body linear acceleration and angular velocity information; magnitude using Euclidean norm and, the frequency domain signals using Fast Fourier Transform (FFT). The preprocessing carried out appears reasonable and justified. A survey of similar studies within the HAR field also indicates the preprocessing was acceptable.

These signals were used to estimate variables of the feature vector for each pattern: '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.

tBodyAcc-XYZ

tGravityAcc-XYZ

tBodyAccJerk-XYZ

tBodyGyro-XYZ

tBodyGyroJerk-XYZ

tBodyAccMag

tGravityAccMag

tBodyAccJerkMag

tBodyGyroMag

tBodyGyroJerkMag

fBodyAcc-XYZ

fBodyAccJerk-XYZ

fBodyGyro-XYZ

fBodyAccMag

fBodyAccJerkMag

fBodyGyroMag

fBodyGyroJerkMag

The set of variables that were estimated from these signals are:

mean(): Mean value

std(): Standard deviation

mad(): Median absolute deviation

max(): Largest value in array

min(): Smallest value in array

sma(): Signal magnitude area

energy(): Energy measure. Sum of the squares divided by the number of values.

iqr(): Interquartile range

entropy(): Signal entropy

arCoeff(): Autorregresion coefficients with Burg order equal to 4

correlation(): correlation coefficient between two signals

maxInds(): index of the frequency component with largest magnitude

meanFreq(): Weighted average of the frequency components to obtain a mean frequency

skewness(): skewness of the frequency domain signal

kurtosis(): kurtosis of the frequency domain signal

bandsEnergy(): Energy of a frequency interval within the 64 bins of the FFT of each window.

angle(): Angle between vectors.

Additional vectors obtained by averaging the signals in a signal window sample. These are used on the angle() variable:

gravityMean

tBodyAccMean

tBodyAccJerkMean

tBodyGyroMean

tBodyGyroJerkMean

---

3. Data preprocessing

The dataset used in this paper had already been pre-processed.

3.a Checking for duplicates : zero.

3.b Checking for missing values : zero.

3.c Checking for class imbalance: looks well-balance.

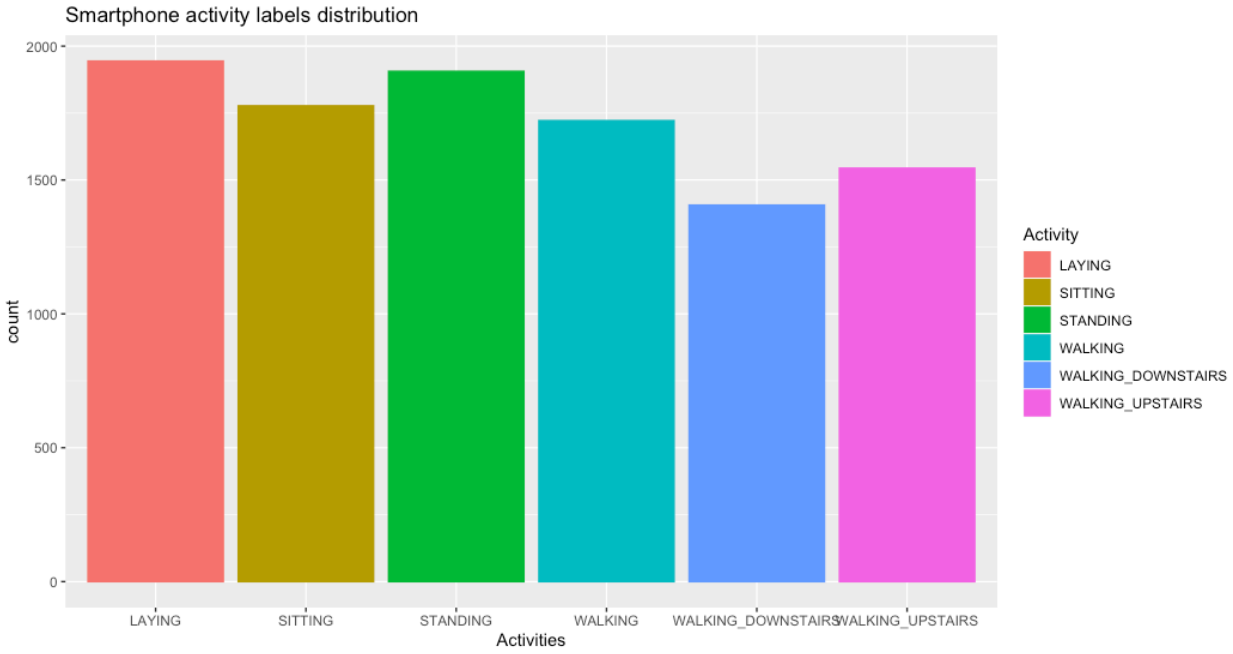
```
> table(overall_data$Activity)

            LAYING            SITTING           STANDING            WALKING WALKING_DOWNSTAIRS   WALKING_UPSTAIRS
              1944               1777               1906               1722               1406               1544
```

---

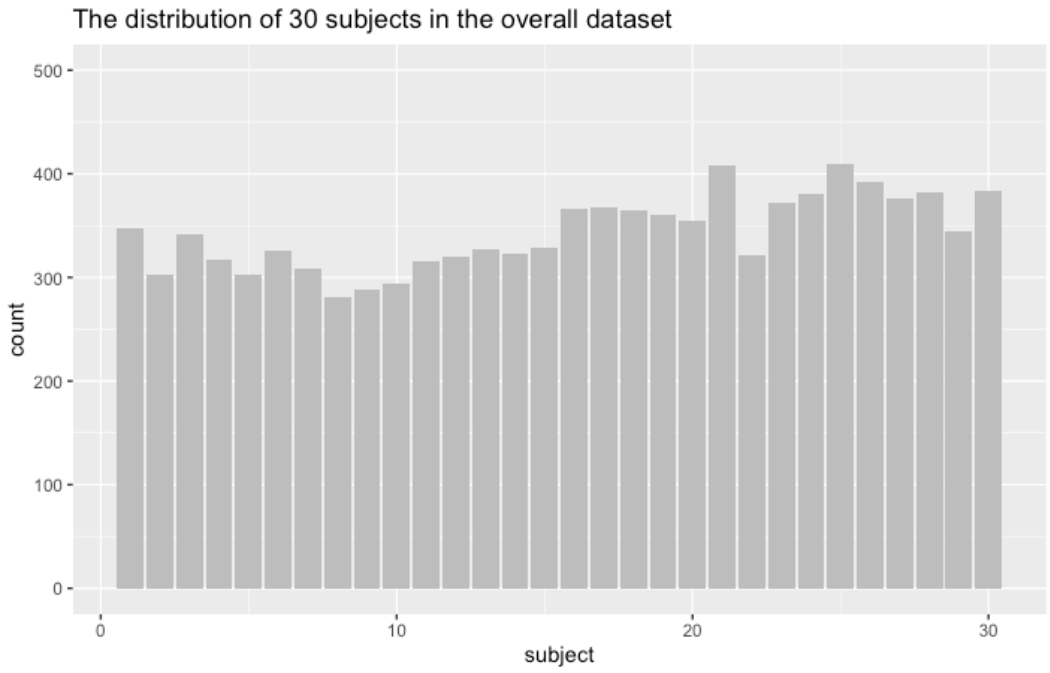
4. Exploratory Data Analysis

4.a Data distribution by activity



Though the observations for each activity are not exactly equal, the data set overall provides a well-balanced distribution of the activity observation. Even after split the data, the balance in observation holds true.

4.b Data distribution by subject



4.c Analysing tBodyAccMag-mean feature

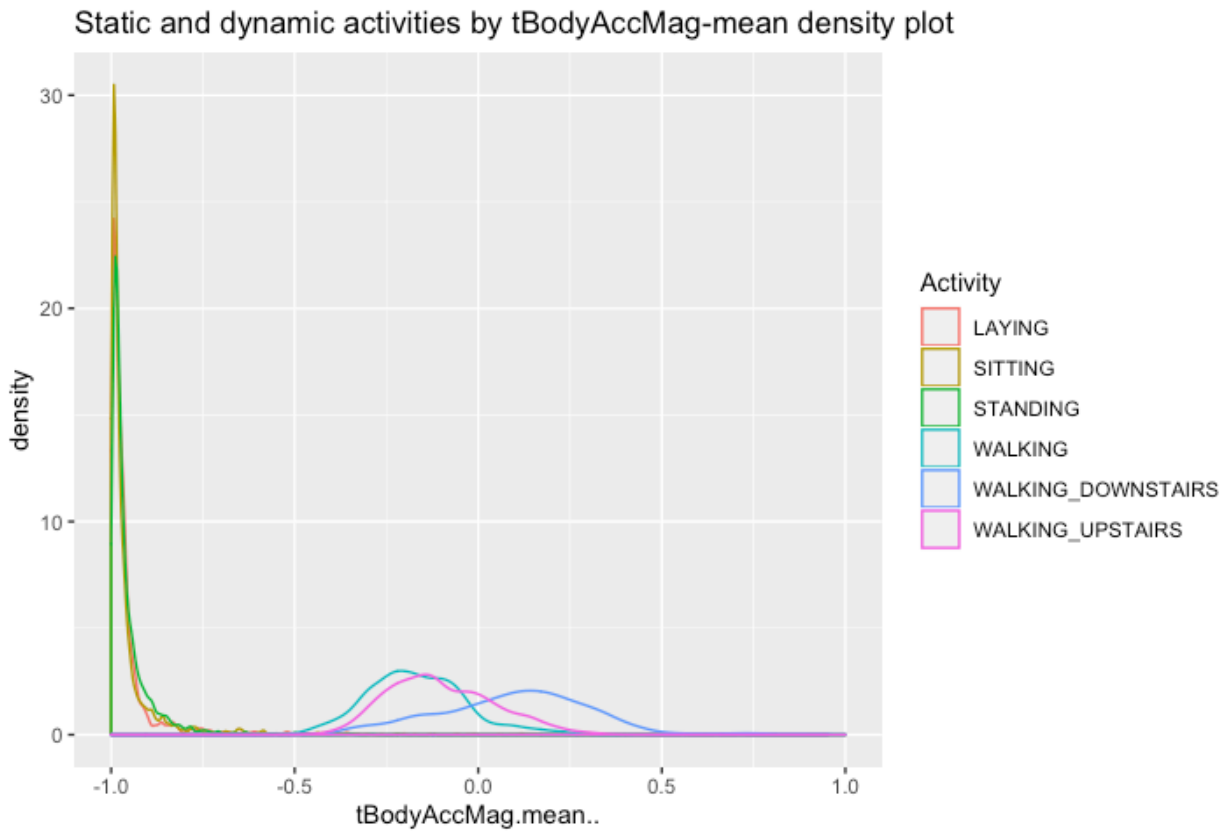
Based on the common nature of activities we can broadly put them in two categories.

Static and dynamic activities :

SITTING, STANDING, LAYING can be considered as static activities with no motion involved;

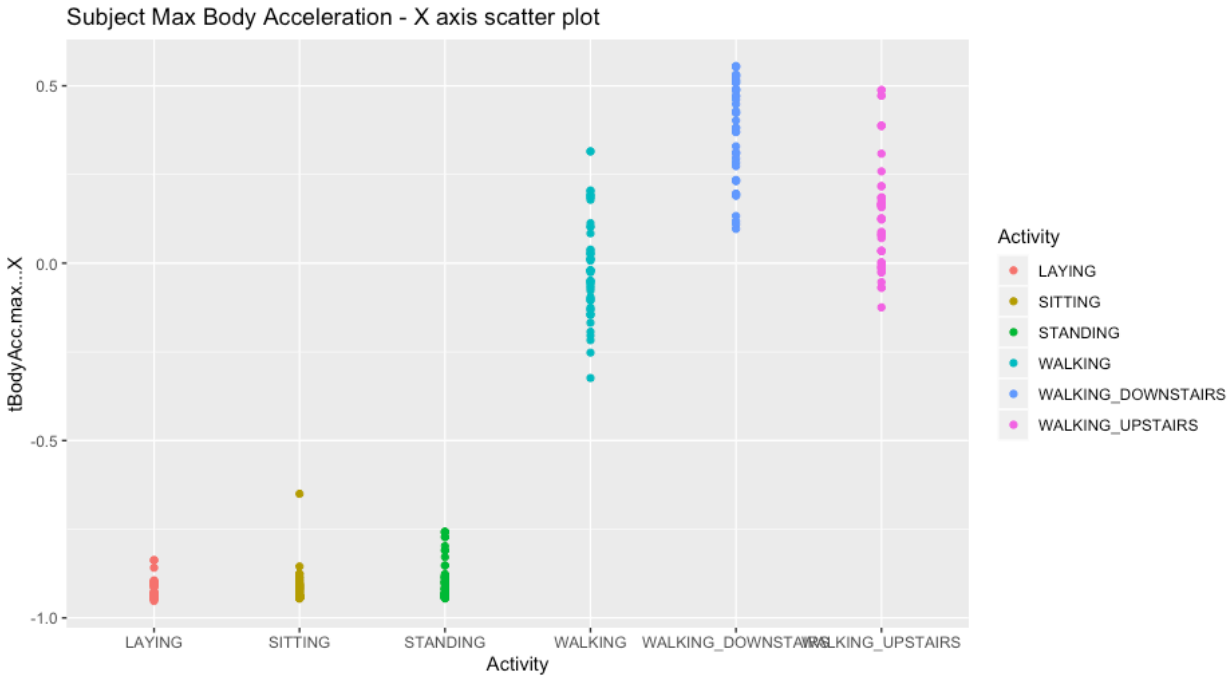
WALKING, WALKING_DOWNSTAIRS, WALKING_UPSTAIRS can be considered as dynamic activities with significant amount of motion involved.

Using the density plot we can easily come with a condition to separate static activities from dynamic activities.



4.d Analyzing Max Body Acceleration feature for subject 20

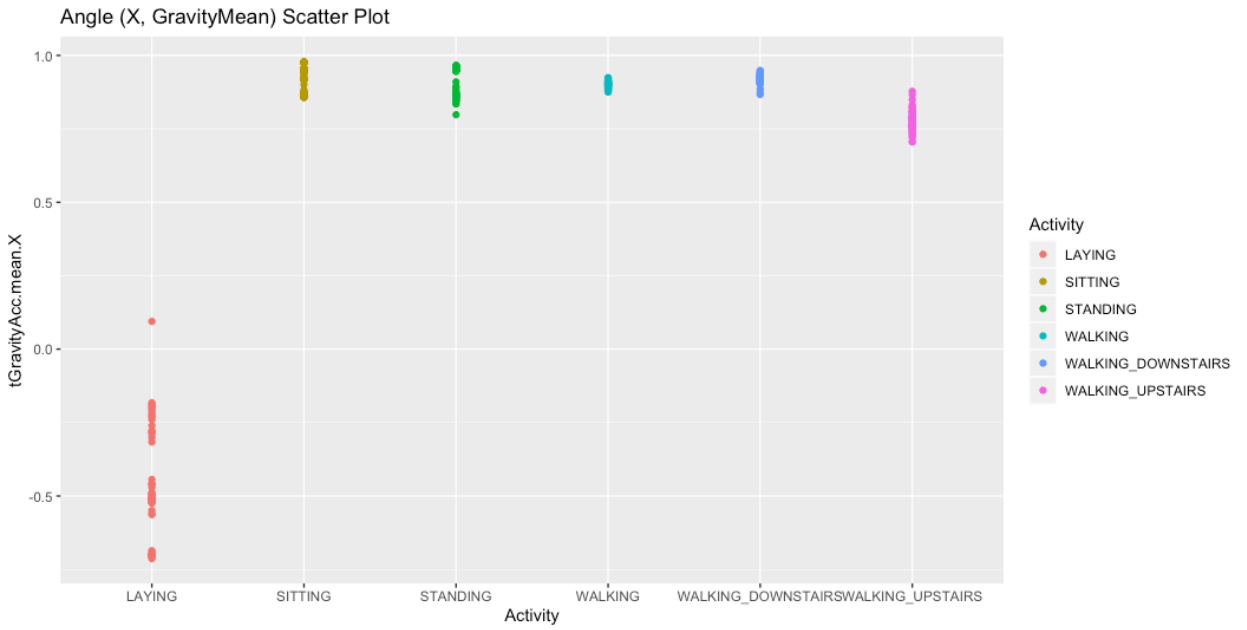
This analysis is carried out to understand the variation in data for each activity. This and next experiment were carried out on the data recorded only for subject 20.



The value along the x-axis can help us tell distinguish between walking, walking upstairs and walking downstairs, but it doesn't provide any insights into the passive activities. This may indicate that the acceleration alone is not sufficient enough for activity recognition.

4.e Analyzing Angle between X-axis and gravityMean feature subject 20

This plot shows a clear distinction for the “laying” activity from the others. In a similar way, the features engineered provide important insights into recognizing human activities.



5. Modeling

5.a GLM model with lasso

Firstly, I chose a generalized linear model with lasso to classify the classes. In this part, I split the original data into two part to train the model and validate the model to get a best lambda. The variables also decreased by using lasso in this processing.

5.b Linear SVM model

Linear SVM is the extremely fast machine learning (data mining) algorithm for solving multi-class classification problems from ultra large data sets.

5.c Naive Bayes model

Naive Bayes models are a group of extremely fast and simple classification algorithms that are often suitable for very high-dimensional datasets. Because they are so fast and

have so few tunable parameters, they end up being very useful as a quick-and-dirty baseline for a classification problem.

5.d Decision tree model

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease to interpretation. They are adaptable at solving any kind of problem at hand.

5.e KNN model

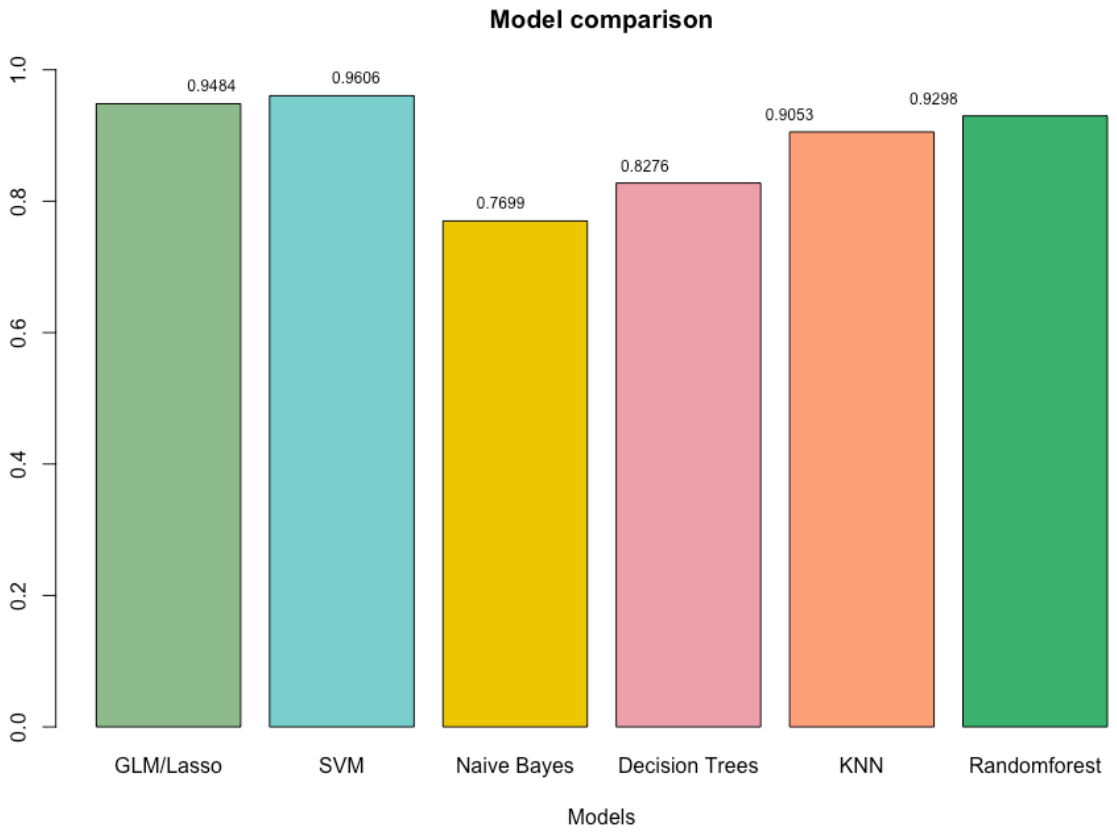
K-nearest-neighbor (kNN) classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. Its principle is that an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

5.f Random forest model

It is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object.

5.g Model Comparison

Compared with the accuracy of the models in the training data set, we are more concerned about their performance in the test data set. The classification accuracy of each model is shown in the figure below.



6. Conclusion

The classification performance of the above models in this case is quite different. GML and SVM outperformed the other models. KNN and Random forest are also good and their accuracy are above 90%. The performance of Naive Bayes and Decision trees are a little underwhelming.

Overall , support vector machine won this competition with a higher accuracy.

7. Limitations

The target of this data has 30 subjects, distributed between 19 and 48 years old, excluding children, adolescents and the elderly. Therefore, the model's generalizability is limited.

Not enough parameter fine-tuning experiments were conducted.

Neural network algorithm is not used in this project due to hardware issues.

Reference:

https://www.kaggle.com/uciml/human-activity-recognition-with-smartphones