DATA MINING AND ANALYSIS
Kaggle Project Write-Up

Stanford University
Autumn 2021 – STATS 202

Prepared for:

Professor – Dan Daniel Erdmann-Pham
Department of Statistics
Stanford University

Prepared by:

Team name: **Lieberman** Score: 0.**94303**
Amalya Cox Johnson
Shu Xian Nian

December 1st, 2021

1 Project Background

Human activity recognition (HAR) is a field of study that utilizes sensor data to classify daily human activities. The application of HAR has various applications in disciplines such as health research, medical security, personal safety, living assistance, and so on. This project involves using smartphone sensor data to identify 12 daily activities, namely: Standing (stand); Sitting (sit); Laying (lay); Transitions between the aforementioned (X_to_Y); Walking (walk); Walking downstairs (down); and Walking upstairs (up).

The features selected for this database come from the smartphone's accelerometer and gyroscope 3-axial raw signals. These time-domain signals were captured at a constant rate of 50 Hz and subsequently filtered to remove the noise in the data. Subsequently, the features of the time domain signals were extracted as features for the database. There are a total of 10929 observations, along with 561 different features; the observations are split into 70% training data and 30% test data. The final model is trained using the training data available, and the objective of the model is to classify the associated activity, given the predictor inputs. The remaining test data is used to assess the quality of the model.

This project was conducted entirely by the Liberman team and has not received any previous academic credit at this or any other academic institution. The team has not collaborated with any other teams.

2 Modeling Approach

During the modeling stage, multiple classification techniques were considered, including support vector machine (SVM), K-nearest neighbors, boosting, and random forest. The final modeling method employed by the team is SVM, as it yielded the best performance on the test data.

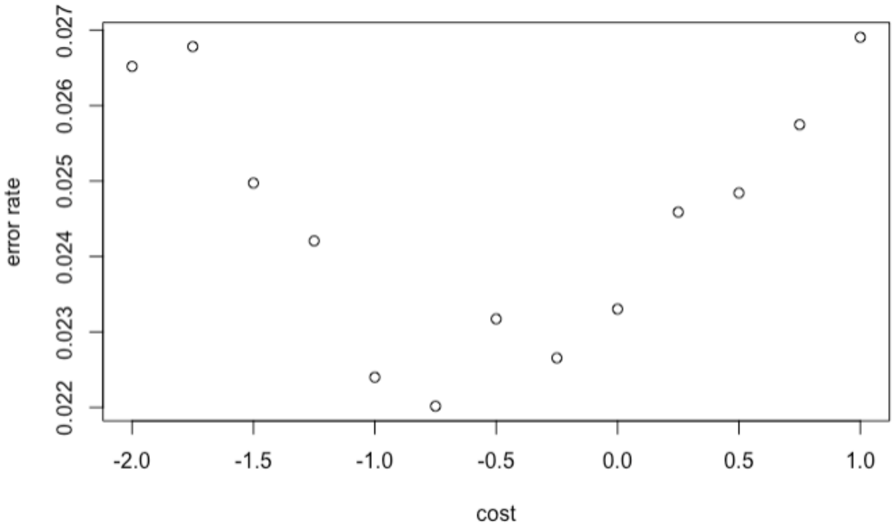
The modeling process for the SVM approach can be aggregated into four main sections: (1) loading and pre-processing the data; (2) tuning the cost parameter for SVM; (3) building an SVM model using the tuned cost parameter; (4) predict the activity for the test data set.

2.1 Loading and Pre-processing

Both the training and test data sets were loaded into R so that the model could be trained based on the training data set, and subsequently produce activity classification results for the test data. The data sets both contain an "Id" field that indexes the observation and is trivial as a feature; therefore, it was removed for the modeling process. Additionally, the "Activities" field contains the labels for this classification problem; therefore, the "Activities" field was encoded as a factor to maintain the categorization of the different activities.

2.2 SVM Hyperparameter Tuning

The accuracy of SVM is dependent on the cost of a violation to the margin; when the cost argument is small, then the margins will be wide and many support vectors will be on the margin or will violate the margin. When the cost argument is large, then the margins will be narrow and there will be few support vectors on the margin or violating the margin. Therefore, a 10-fold cross-validation process was adopted to determine a suitable cost parameter between 0 to 1. Linear kernels were used for the SVM models built during the cross-validation process. The choice of linear kernels is to reduce the complexity of the resulting model, as there are already 561 associated features, using a more complex kernel could potentially lead to overfitting. The resulting cost parameter that yielded the lowest cross-validated mean squared error (MSE) was $10^{-0.75}$.



2.3 SVM Model

The final SVM model was built using the entire set of training data. The cost parameter that was employed in the model was the cost that produced the lowest MSE during the tuning stage, which was $10^{-0.75}$. Additionally, like in the tuning stage, a linear kernel was used to curtail the potential of overfitting. The training error rate in the final model was 0.006179992.

2.4 Prediction

The final prediction of the activities associated with the test data observations was conducted using the final SVM model. The test data set, which contains all the input feature data, was fed into the model to arrive at the final predictions. The predictions were tabulated alongside the associated “Id” in a .csv file, which was then submitted to Kaggle for model performance evaluation. When the predictions were submitted to kaggle, the prediction accuracy was 0.94303.