Daniel Lee
18379784
Project 3

Predicting Parkinson’s Disease Progression with Smartphone Data

Summary

This project is an attempt to not only predict whether a person has Parkinson’s disease or not based on the accelerometer activity on the person’s smartphone but also to be able to measure the disease progression. I only focused on predicting whether a person has Parkinson’s disease or not. The data was collected from smartphones carried around by 16 people. Seven of the 16 people had Parkinson’s disease. The subjects were asked to carry a supplied smartphone with them for at least 4 to 6 hours per day. The participants were asked to carry the smartphones for at least 8 weeks. The data consists of audio, accelerometry, compass, ambient light, proximity, battery level, and GPS. My attempt was to focus on the data collected by the accelerometer to predict whether a person has Parkinson’s disease or not. I used the support vector machine classification method with linear and radial basis kernels. I used 10-fold cross validation on my training set. For my testing set, I was able to obtain ROC curves for both methods, with the AUC for the linear kernel essentially equal to 1. The reason for such high accuracy could be that it’s quite easy to distinguish between a patient with Parkinson’s disease and someone who does not have Parkinson’s disease. One of the winning methods followed these procedures on MATLAB. I tried my best to replicate these procedures on R. Even though I only examined predicting whether a person has Parkinson’s disease, I believe a more difficult question would be to predict the progression of the disease, which I did not explore for the moment.

Competition Question

Based on the data collected from the smartphones, predict whether a person has Parkinson’s disease.

Data Processing

Though there was a lot of data available, I focused on the acceleration data. The accelerometer recorded the subject’s movements every two seconds. The csv files that were available were divided into data from one-hour periods for each patient. There were 5663 csv files collected by the accelerometer, each csv file representing data for one subject during a one-hour period. I took the x-mean, y-mean, and z-mean coordinates available every two seconds and took the aggregate root-mean-square of the data for the entire hour. Given that there were 5663 csv files, I ended up with 5663 root-mean-square values for the x-y-z coordinates. I did the similar root-mean-square aggregation for the four power spectral density channels. Therefore, I was left with five features to perform classifications.

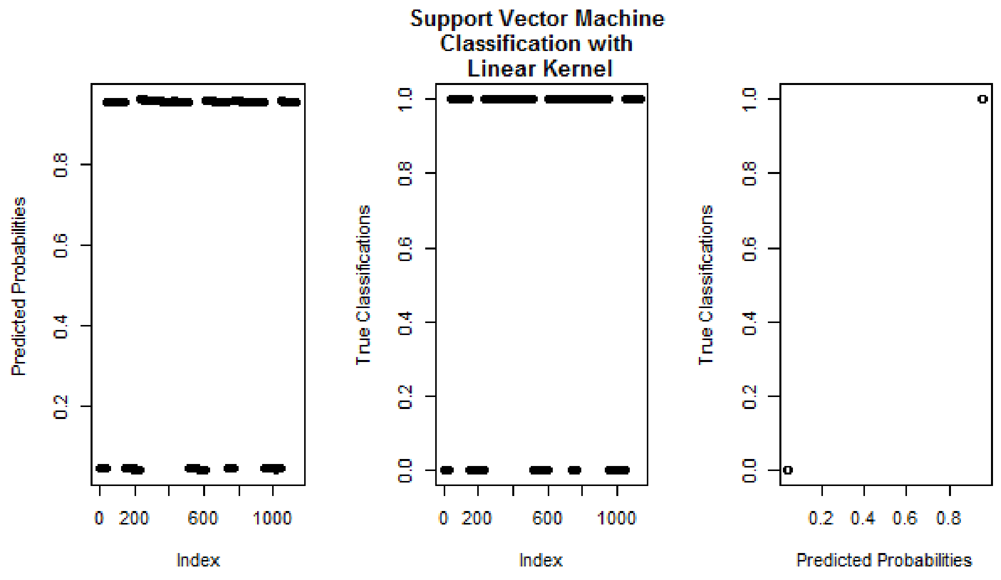
Data Analysis

Summary of Winning Method

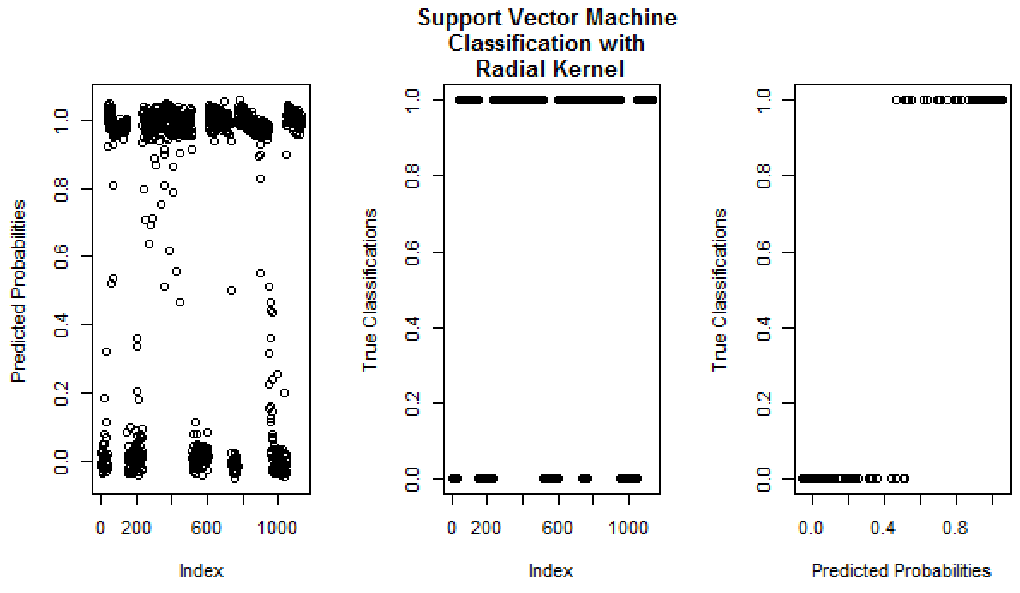
One of the winning methods analyzed the data as I analyzed the data. The person focused on the accelerometer data and calculated the root-mean-square values of the x-y-z coordinates. She also focused on the four power spectral density channels. Then, she performed a k-fold cross validation on the support vector machine algorithm using the LIBSVM library in MATLAB. She was able to yied an accuracy of ~86.6% using the linear separator while the radial basis classifier yielded accuracies of ~99%. Other teams also performed support vector machines, though their feature selections were slightly different. The winning team also attempted to predict the progression of the disease.

My Solution

After simplifying the data with the five feature selections for 5663 csv files, I sampled a random subset of 1133 datasets (approximately 20% of the entire dataset) and set it aside for it to be used as a testing set. With the remaining 4530 datasets, I performed a10-fold cross validation for the support vector machine classification method using the e1071 library. First, I trained the data on the linear separator. Then I trained the data using the radial based separator. The AUROC were quite high. The AUROC for the linear separator came out to be one. The AUROC for the radial based separator came out to .9999. Below, I included plots that show the predicted probabilities and the actual classifications for the test set. It can be seen that the predicted probabilities and the true classifications match very closely for the fits generated by the linear separator. The fits generated by the radial based separator also does well, but not as accurate as the linear separator.

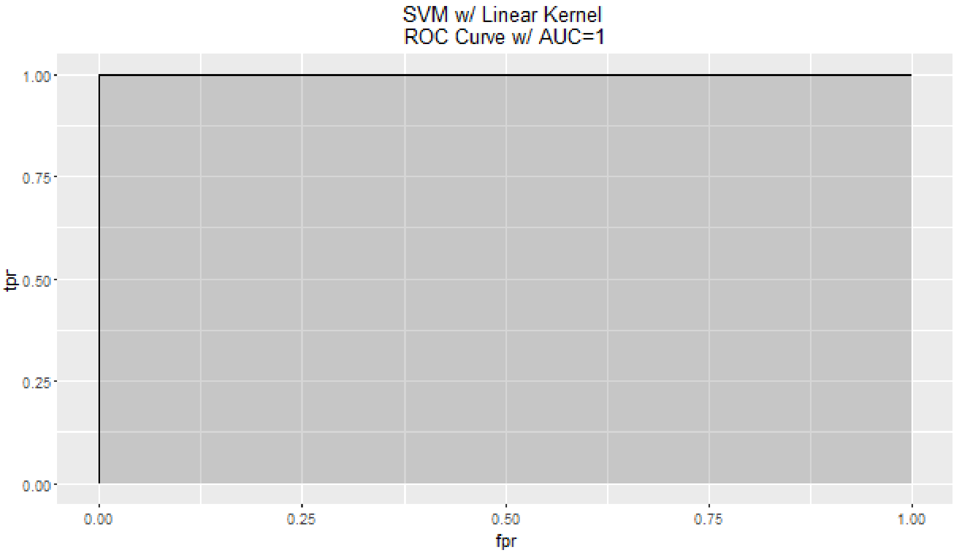


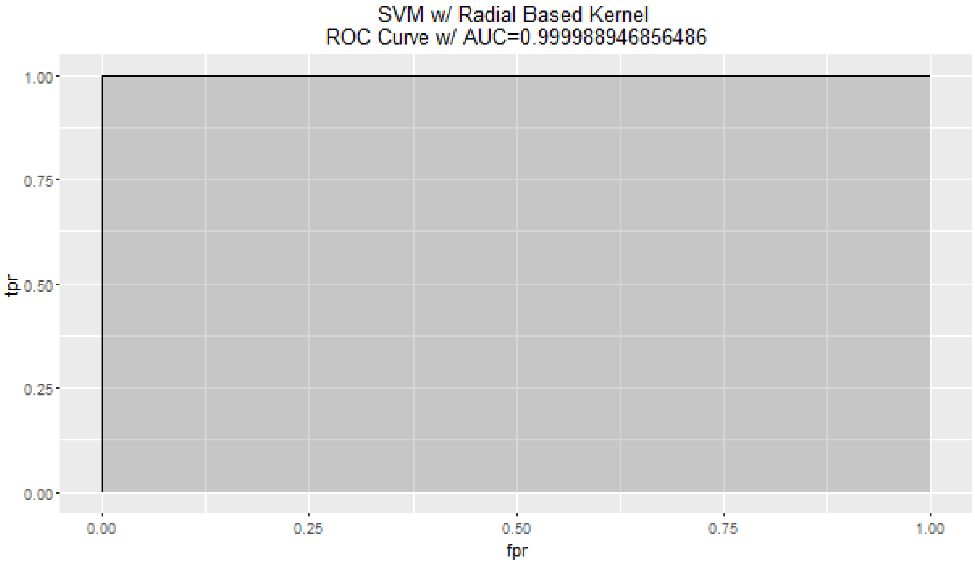
Support Vector Machine Classification with Linear Kernel



Support Vector Machine Classification with Radial Kernel

Below are the ROC curves for both the support vector machine using the linear kernel as well as the radial based kernel. AUC for linear kernel is one, and AUC for radial based kernel is 0.9999.





My Own Question

How do I use the data to predict the development and the stage of Parkinson’s disease a patient is in?

Proposed Solution to Own Question

One idea is to examine the magnitudes of the movements to determine what stage of the disease the patients may be in. Another thought could be to incorporate other data provided, such as GPS location and audio, to predict the stage of disease.