COMPUTATIONAL METHODS IN STATISTICS
Final Project-PARTII
REGRESSION ANALYSIS

February 2, 2020

Author: Bita Nezamdoust
Student ID: 002332793
Instructor: Dr. Xin Qi

Georgia State University
Department of Mathematics and Statistics

INTRODUCTION

The present project deals with classification of human activity based on a human recognition device. The goal is to create a high-accuracy classification model that has the ability to predict six identified human physical activities based on the information recorded by smartphones worn by individuals doing the targeted activities. Classification methods will be employed and using the proper testing methods, misclassification errors will be used as a measure to compare the performances of different models in order to obtain the best classification model.

Data Description

The data set has been built from the recordings of 30 subjects performing daily activities while carrying a waist-mounted smartphone with embedded inertial sensors. The six activities and hence, the six classes included WALKING, WALKING UPSTAIRS, WALKING DOWNSTAIRS, SITTING, STANDING, LAYING. Using its embedded accelerometer and gyroscope, they captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. [2]

There are 561 predictor variables and six classes of human activity as named above in the response variable. While collecting the data, the dataset was randomly split into two sets, where 70% of the volunteers were selected for generating the training data and 30% the test data. In the present report, a subsection of the original data has been used, with 300 observations in the training set and 300 observations in the test set.

While there is multicollinearity among some of the predictor variables as can be seen in 1, the majority of the predictors do not seem to be very correlated according to 2, in which the correlation between a larger section of the data has been depicted.

Prediction Models

The goal is to build a classification model with highest accuracy level to predict one of the six pre-determined human activities based on the collected information. As before, the models' performances are tested on the test set from the data and their misclassification rates are computed and presented in following tables for comparison.

The first column shows the misclassification error resulted from *Usual Logistic Regression* that is equal to 0.09. That means the model is able to offer classification with %91 accuracy. The error becomes larger when *Ridge* and *Lasso Penalty* are applied, which could be due to

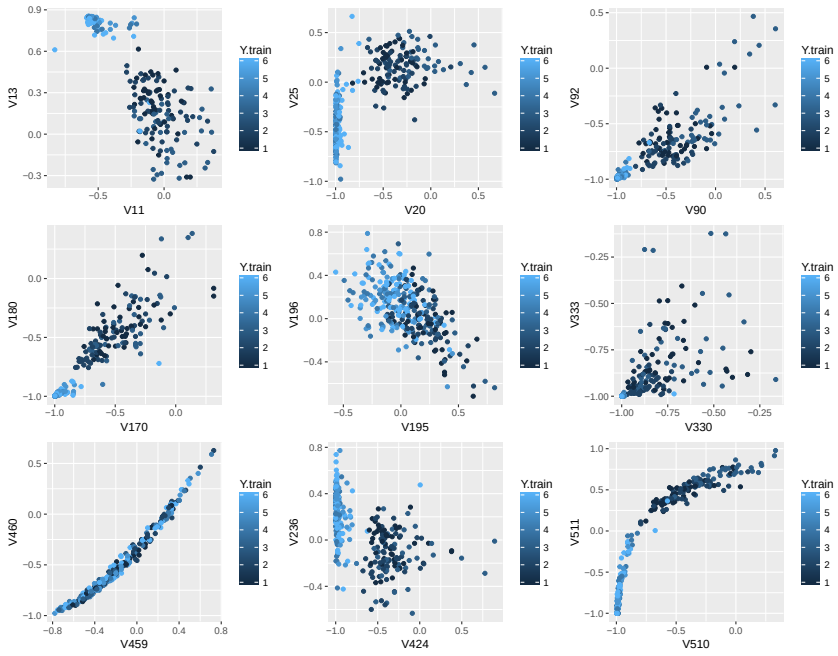

Figure 1: The amount of variability in the data explained by the PC's

the weak multicollinearity of the predictors.

Misclassification Error Comparison						
Logistic Regression	Ridge Regression	Lasso Regression	LDA	QDA	RDA	Tree
0.09	0.15	0.15	0.21	0.21	0.06	0.21

A number of further classification techniques have been employed in this study to achieve the best accuracy that will be briefly introduced here.

Three well-known discriminant analysis techniques include *Linear Discriminant Analysis (LDA, Quadratic Discriminant Analysis (QDA)* and *Regularized Discriminant Analysis (RDA.* The resulting models are built on the basis of *the discriminant functions* and *the Bayes Classification Rule,* that is also known as the "optimal classification rule". The three methods have been employed in this study and misclassification errors obtained are 0.21, 0.21, and 0.06, for LDA, QDA and RDA, respectively. It is apparent that the RDA error is significantly smaller and seems to offer the best classification accuracy so far obtained.

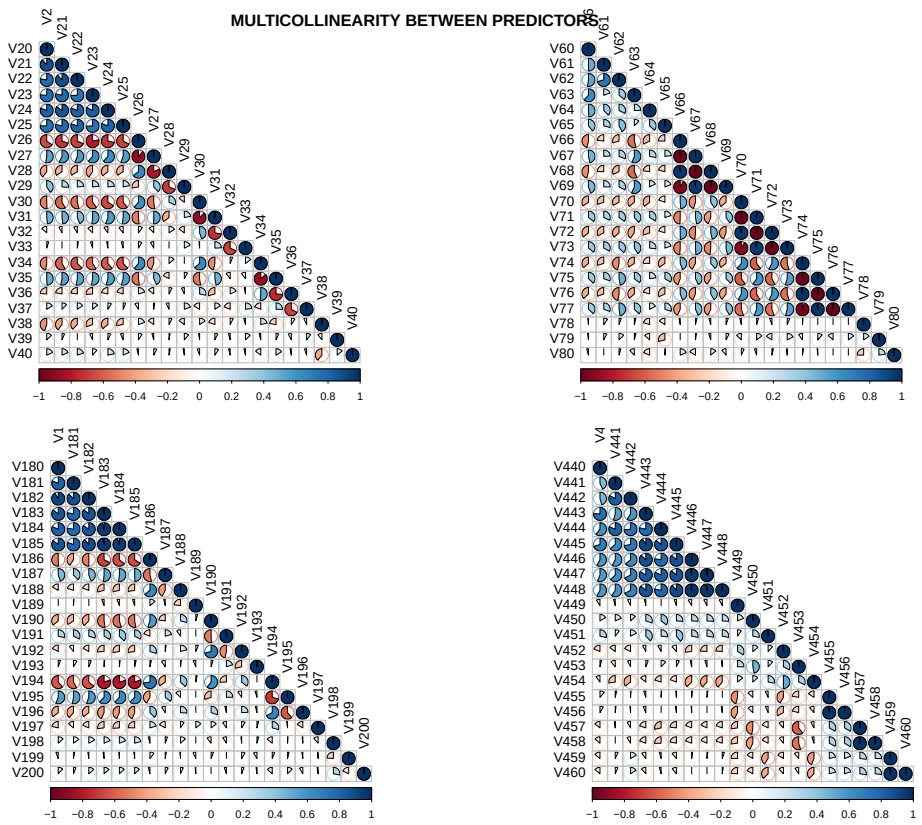

Figure 2: The amount of variability in the data explained by the PC's

Misclassification Error Comparison			
SVM (Linear)	SVM (Polynomial)	SVM (Radial)	SVM (Sigmoid)
0.09	0.08	0.08	0.1
cost = 0.05	C = 1, d =2, $\gamma = 0.0005$, $\gamma_0 = 10$	C = 5, $\gamma = 0.0005$	C = 5, $\gamma = 0.0005$, $\gamma_0 = 0$

Support Vector Machine (SVM is another linear classification method that is designed to make optimal decision boundaries. The method uses the concept of *margin*, defined as the shortest distance from all the observation points to the linear decision boundary, and attempts to identify the classification rule that maximizes the margin, so that the probability of faulty classification due to random fluctuation of the data is minimized. SVM is computed for four types of *kernel functions* as named in the table above and the misclassification errors for the four models include 0.09 for Linear SVM, 0.08 for Polynomial SVM, 0.08 for Radial SVM and 0.1 for Sigmoid SVM. The values to the tuning parameters for each model could as well be found in the table.

Lastly, a non-linear classification method have also been employed here that is the *Classification Tree*. The tree uses a criteria such as the misclassification rate to select one variable

and then start splitting the values and assigning class labels. The misclassification rate for the tree is 0.21 which in comparison cannot compete with the other rates computed before it.

Therefore, the best model for the prediction of the human activity using the smatrphone data is obtained by **Regularized Discriminant Analysis (RDA)** with the prediction error of 0.06 and %96 rate of accuracy. The tuning parameters for the model have been selected by 10-fold cross-validation that include the optimum *alpha* equal to 0.99 and the optimal *delta* equal to 0.

CONCLUSION

In this study, the best classification model was obtained to classify six categories of human activity based on the data from smartphones as recognition tools. The best final model was found to be the *Regularized Discriminant Analysis Model* with %96, that is up from %91 accuracy obtained by the usual logistic regression model initially built in the study. Therefore, with the selection of appropriate tuning parameters in the classification methods applied, the accuracy of the models were significantly improved.

REFERENCES

[2] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.[online] available at:
http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones