Stats 503 Project: Predicting the Gestures

Chen Xie, Xun Wang, Xinye, Jiang

April 27, 2019

1. Introduction

The dataset is about 11678 recordings of human hand muscle activity corresponding to four different hand gestures from a prosthetic control system. It has four classes of motions which are “rock”, “scissors”, “paper” and “ok” as the response variable. And the rest 64 variables showing the 8 consecutive readings of all 8 sensors are the explanatory variables. The specific data description can be seen in Table 1.

Table 1: Data Description

Variables names	Type	Description
V65	categorical	Response (rock:0, scissors:1, paper:2, ok:3)
V1-V8	continuous	Reading 1 Sensor 1-8
V9-V16	continuous	Reading 2 Sensor 1-8
V17-V24	continuous	Reading 3 Sensor 1-8
V25-V32	continuous	Reading 4 Sensor 1-8
V33-V40	continuous	Reading 5 Sensor 1-8
V41-V48	continuous	Reading 6 Sensor 1-8
V49-V56	continuous	Reading 7 Sensor 1-8
V57-V64	continuous	Reading 8 Sensor 1-8

Is there any connection between the sensor and the hand gesture? Does the reading account for the hand gesture?

The goal of this project is to predict hand gestures based on the readings and sensors.

In the rest part of the project, we explored the dataset numerically and graphically, done variable selection based on the importance of the variables, and utilized various classification methods including logistic regression, LDA, QDA, KNN, SVM, classification tree and random forest to fit the data and calculated the prediction accuracy respectively.

2. Data Exploration

To have an intuitive sense of the dataset, first look at the barchart of the response variable “gesture” in Figure 1. We saw that the dataset was balanced, as each category of the response variable had around 2900 observations.

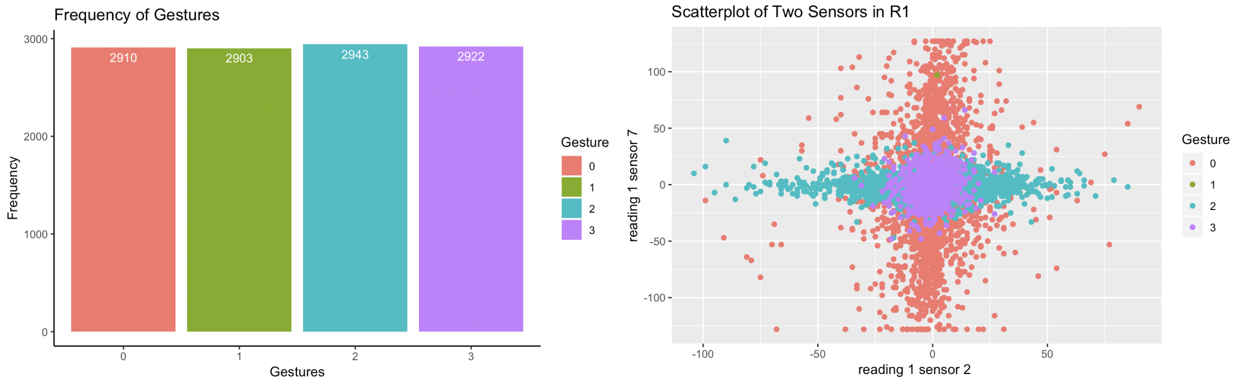

Figure 1: Barchart of Gestures and Scatterplot of Sensor 2 and Sensor 7 from Reading 1

We also drew the scatterplot of two sensors in reading 1 in Figure 1. We had 64 predictor variables which are quite high dimensional, so we tried pairwise scatterplots of this kind to get some sense of the dataset. We can see that there is no obvious structure in the dataset. Combined with the numeric summaries of the predictor variables in Table 2 (note that we only show part of it), all the predictors were focused around 0 and had long symmetric tails. As no specific pattern was detected, PCA won't be a good way to reduce dimension. We tried PCA and found that the first 35 principal components merely covered 80% of the variance which meant that the effect of this dimension reduction method to this dataset was terrible.

Table 2: Numerical Summaries (only part is shown)

Summary	V1	V2	V3	V4	V5	V6	V7	V8
Min.	-116.00	-104.00	-33.00	-75.00	-121.00	-122.00	-128.00	-128.00
1st Qu.	-9.00	-4.00	-3.00	-4.00	-10.00	-15.00	-6.00	-8.00
Median	-1.00	-1.00	-1.00	-1.00	0.00	-1.00	-1.00	-1.00
Mean	-0.52	-0.73	-0.74	-0.73	-0.16	-0.55	-1.27	-0.66
3rd Qu.	7.00	3.00	2.00	3.00	10.00	13.00	4.00	6.00
Max.	111.00	90.00	34.00	55.00	92.00	127.00	127.00	126.00

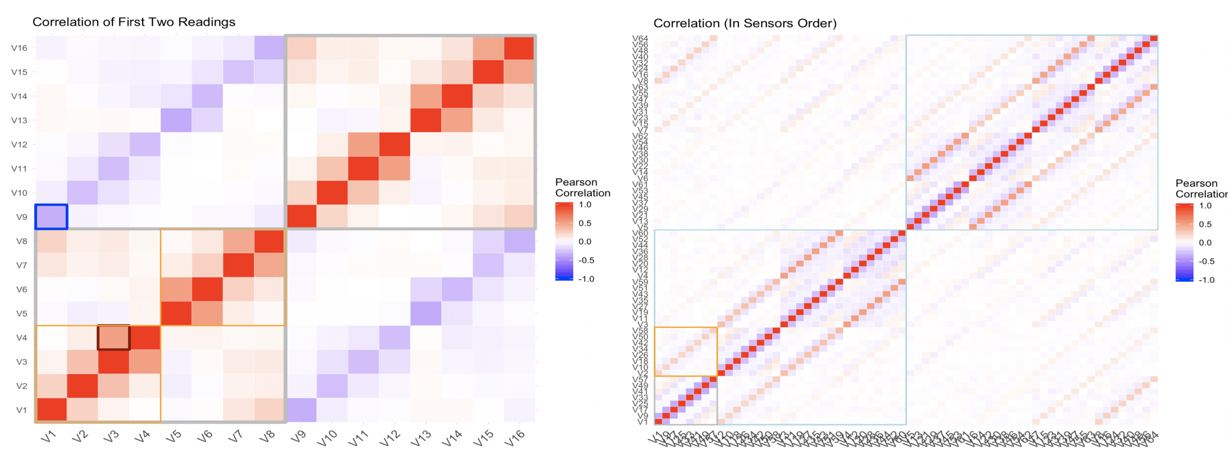

Figure 2: Correlation Plots

To figure out the correlations between the predictors, look at the correlation plots in Figure 2. From the left plot of the correlations of first two readings, we saw that the same sensors in the adjacent readings were negatively correlated, pairwise adjacent sensors in each reading were positively correlated while the first 4 and last 4 sensors in each reading were nearly uncorrelated. The right plot of the correlations in sensors order reconfirmed our conclusions before.

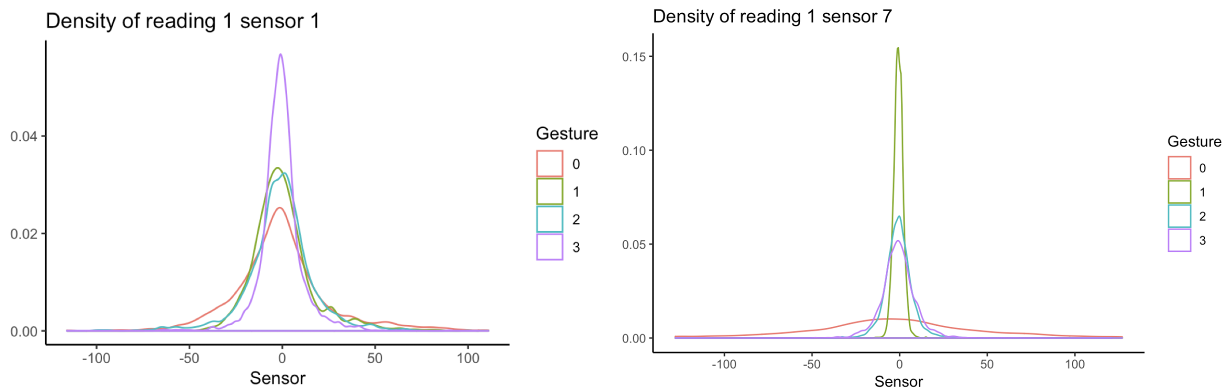

Figure 3: Density Plots

Using the density plots in Figure 3, we addressed the different significance of each predictor variable in prediction for gesture. For example, the densities of sensor 1 in reading 1 for 4 gesture classes were quite similar, while the densities of sensor 7 in reading 1 were quite different. In this case, the sensor 7 should be put on more weights for prediction than the sensor 1. The importance diversity encouraged us to do model selection to reduce dimensions.

To visualize the data more comprehensively, look at the plot of the contour of density of V2 and V7 in Figure 4. The data looked like normal with unequal variances. Based on this plot, we expected that the decision boundary between the classes were non-linear and thus those classifiers generating linear boundaries such as LDA, logistic regression would perform poorly in this setting.

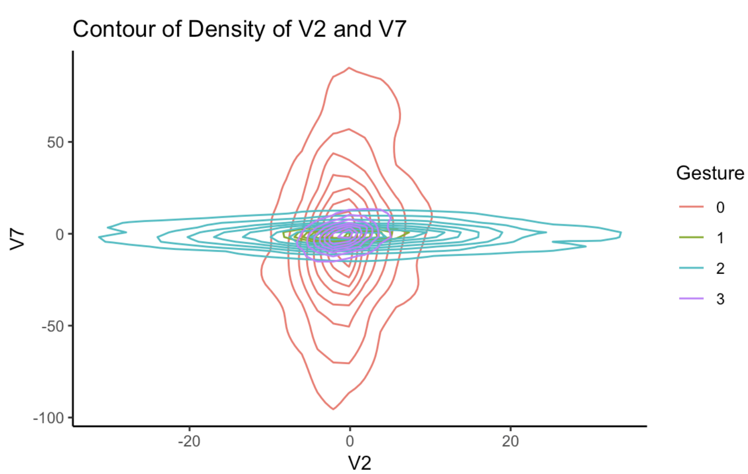

Figure 4: Contour of Density of V2 and V7

Variable Selection

As mentioned before, we had 64 predictor variables and some were more important than others. Therefore, we tried two methods to do variable selection to reduce redundance and better the prediction performance and interpretability.

The first method was to use random forest with mtry=8 to fit the whole dataset and check the importance of the variables. We found that the last but one sensor in each reading was most important and the second sensor in each reading was next most important, which can be seen in Figure 5.

The second method was to utilize stepwise forward selection based on AIC and BIC values to choose the corresponding optimal logistic regression models respectively and find the important variables. The important variables found in this method were basically the same as the ones we found before using random forest.

Finally we separately checked the overall prediction accuracy using 8 predictors, i.e., the 7th sensors in 8 readings, and the accuracy using 16 predictors, i.e., the 7th and 2nd sensors in 8 readings. And we reached the conclusion that the latter one had overall better performance than the former one. Thus, we finally decided to carry on the remaining analysis with 16 important predictors.

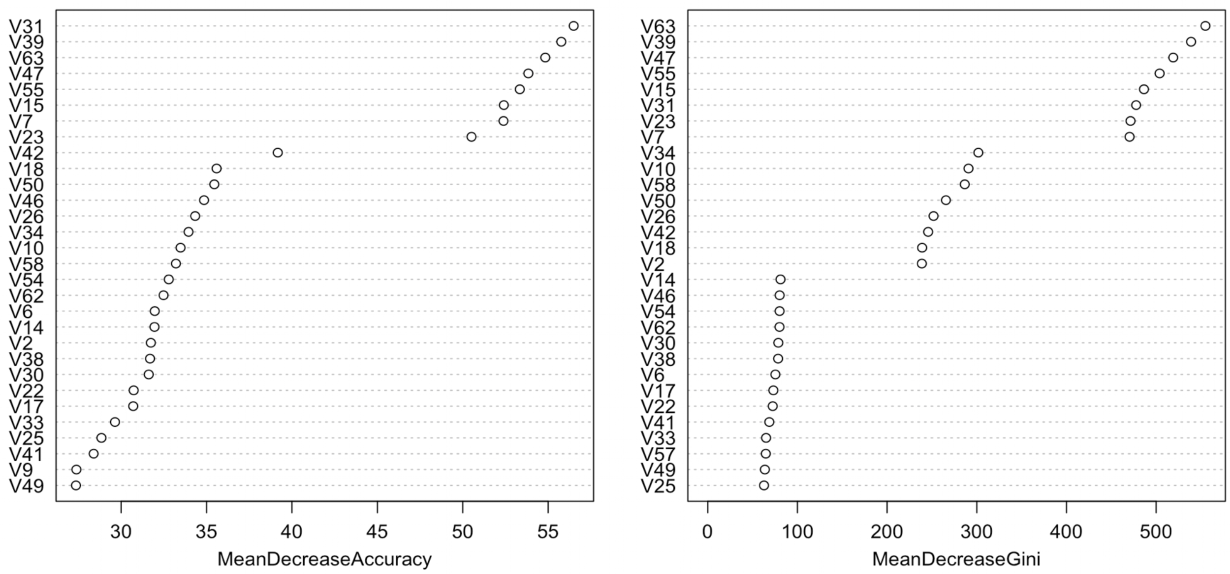

Figure 5: The Importance of the Variables

3. Classification Methods

In this part, we fitted various classification methods. In order to compare their prediction performance, we split the dataset into training and testing two parts.

3.1 Logistic Regression

The first model we tried was the logistic regression. We fitted the original logistic regression which had all its terms linear in the model and also fitted one which included quadratic terms of the important variables.

Table 3: Logistic Regression (Linear & Quadratic) Prediction

Truth vs Logistic Prediction	Linear 0	1	2	3	Quadratic 0	1	2	3
0	317	60	90	135	560	1	7	34
1	18	22	200	329	0	556	4	9
2	84	158	205	153	13	38	523	26
3	119	33	152	261	15	103	12	435

From the logistic regression prediction in Table 3, we saw that adding quadratic terms in the logistic regression model largely increased the prediction accuracy from 34.5% to 88.8%. This confirmed our expectation before that the decision boundary between classes was not linear.

3.2 LDA & 3.3 QDA

We then fitted LDA and QDA classifiers to the dataset.

Table 4: LDA & QDA Prediction

Truth vs	LDA 0	1	2	3	QDA 0	1	2	3
0	294	63	108	137	559	1	4	38
1	4	17	204	344	0	544	8	17
2	54	164	222	160	9	20	559	12
3	84	28	172	281	21	72	10	462

Here we could see that the classifier QDA with quadratic classification boundary had an overwhelming advantage over the classifier LDA with linear boundary regarding the prediction accuracy. And QDA had a very good prediction performance with its prediction accuracy over 90.9%.

3.4 KNN

Next we tried KNN on this dataset. We selected K as 3 by finding the corresponding K with the smallest 5-fold cross validation error in the training set, which was shown in Figure 6.

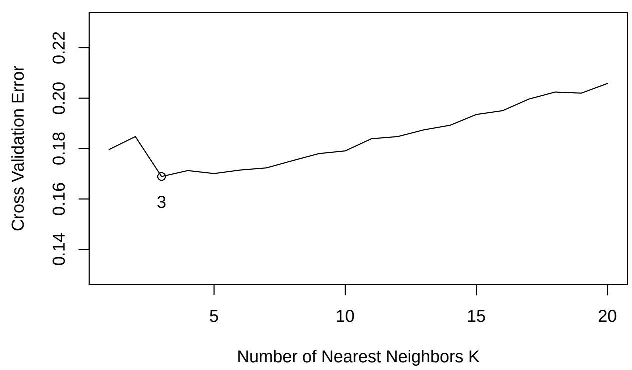

Figure 6: The 5-fold Cross Validation Errors for Each Choice of K.

Table 5: KNN (K=3) Prediction

Truth vs KNN	0	1	2	3
0	517	1	12	72
1	0	553	3	13
2	2	112	435	51
3	7	78	8	472

Table 5 indicated that KNN (K=3) also performed well on this dataset, although its performance was little worse than the logistic regression with quadratic terms and QDA.

3.5 SVM

Fit the SVM classifiers with radial kernel and polynomial kernel to the dataset. The optimal parameters such as cost, gamma, degree were chosen by minimizing the validation errors. Note that here we used validation instead of cross validation under consideration of the computational cost. For the SVM with radial kernel, the optimal parameters chosen were that cost = 1, gamma = 0.5 . For the SVM with polynomial kernel, the optimal parameters cost = 100, degree = 2.

Table 6: SVM with Radial & Polynomial Kernel Prediction

Truth vs SVM	Radial 0	1	2	3	Polynomial 0	1	2	3
0	571	0	3	28	552	1	15	34
1	0	549	12	8	0	548	9	12
2	208	20	364	8	7	38	537	18
3	25	56	20	464	16	80	11	458

Table 6 showed that SVM with these two kernels both performed well, with the one with polynomial kernel having higher accuracy rate than the one with the radial kernel. We saw that the SVM with polynomial kernel had optimal degree 2 and relatively better prediction performance, indicating that the decision boundary between classes probably did have quadratic curves.

3.6 Classification Tree & 3.7 Random Forest

Finally we fitted classification tree and random forest to the dataset.

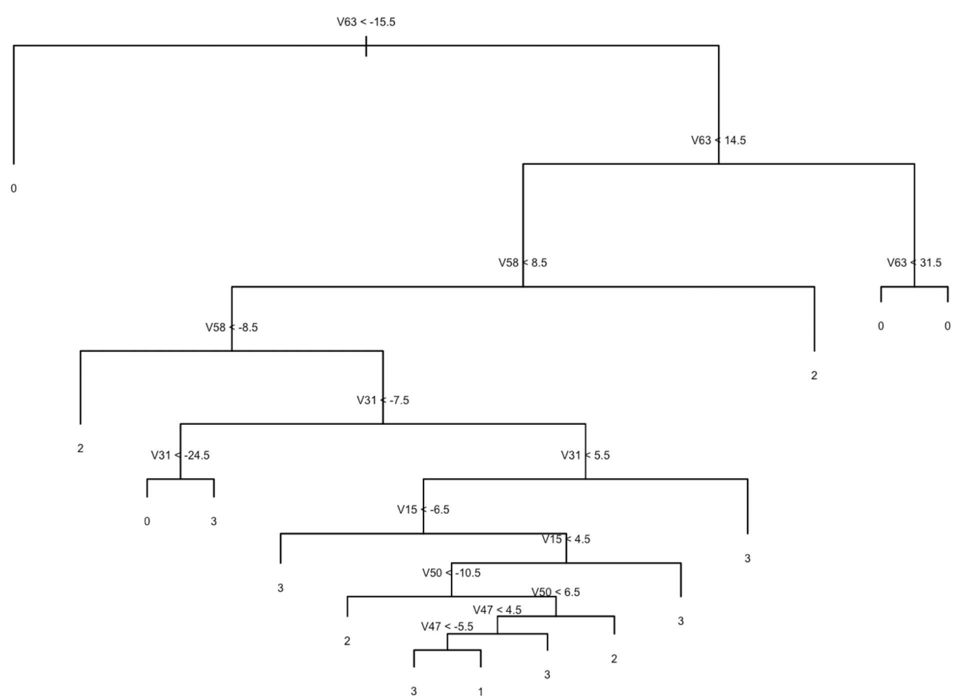

Figure 7: Decision Tree Plot

The ultimate decision tree model was shown in Figure 7. We saw that it used the variables ‘V63’, ‘V58’, ‘V31’, ‘V15’, ‘V50’ and ‘V47’. The decision tree had accuracy slightly larger than 70%, which was hardly counted as good performance. This intrigued us to do random forest to better the prediction performance.

The random forest had the best prediction accuracy so far which was over 91.1%. The partial plots of “V7” and “V18” shown in Figure 8 represented the effects of 7th sensor and 2nd sensor in each reading respectively. The influence of the 7th sensors to the gesture classes varied when they were closed to 0 and away from 0.

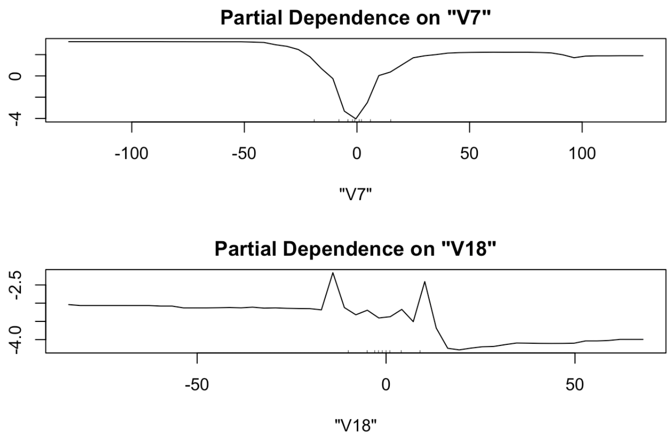

Figure 8: Random Forest Partial Plots

Table 7 showed the prediction results of the classification tree and random forest. For each class category, random forest outperformed the classification tree model.

Table 7: Decision Tree and Random Forest Prediction

Truth vs	Tree 0	1	2	3	Random Forest 0	1	2	3
0	463	1	49	89	569	0	9	24
1	2	441	52	74	0	503	18	48
2	37	35	365	163	2	15	569	14
3	62	92	23	388	33	24	20	488

Prediction Accuracy Comparison

Table 8: Prediction Accuracy of Various Classification Methods

Method	Prediction Accuracy %
Logistic Regression	34.46
Logistic Regression with quadratic terms	88.78
LDA	34.85
QDA	90.92
KNN	84.63
SVM (radial)	83.39
SVM (polynomial)	89.68
Classification Tree	70.93
Random Forest	91.14

Table 8 collected the prediction accuracy rates of all the classification methods we went through. We can see that random forest had the best prediction performance, QDA, SVM with degree-2 polynomial kernel and logistic regression with quadratic terms of the predictors had the next best performance.

4. Conclusion and Discussion

This specific dataset was balanced and clearly had non-linear, probably quadratic decision boundary between gesture classes.

In order to reduce dimensionality, we used the random forest and stepwise forward selection method to do variable selection. We found that the motion of the muscles corresponding to the 2nd and 7th sensors are almost decisive for these 4 gestures.

Among all the classification methods we tried, random forest performed the best, the classifiers with quadratic decision boundary such as QDA, SVM with degree-2 polynomial kernel and logistic regression with quadratic terms all performed quite well.