STA 6707

HOMEWORK 2: Classification

Introduction:

The original problem is to predict the type of cardiac arrhythmia by considering multiple features including but not limited to EEG signals, age, sex, heart rates and so on. This problem is a multiclassification problems to predict the type of cardiac arrhythmia within the existing 16 groups using Linear discriminant analysis (LDA), Quadratic discriminant analysis and decision tree. The original dataset consists of 453 instances and 279 variables including numerical and nominal ones.

Data cleaning

Data cleaning procedures handled issues related to missing values, constant columns, high correlated variables, minority classes. As suggested by Dr.George Michallidis, the missing values which were shown by "?" replaced by the mean value of each column. One of the columns has been removed as more than 95% of the information in that column was missed. To investigate the collinearity of the variables as most of them are constant 0 and 1s, the correlation matrix has been used, and the highly correlated variables were removed. At the end, 187 features including the labels have been used for the task of prediction. The original dataset is imbalanced with the majority to minority class ratio of 2:245(classes 1-16).

Table 1: Distribution of classes

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
245	44	15	15	13	25	3	2	9	50	0	0	0	4	5	22

Linear Discriminant Analysis:

As we discussed in the previous section, some of the constant features with high correlation to other variables have been discarded from the data. The collinearity could lead to singular covariance matrix which raise an error in defining the LDA model. LDA uses the mean of each class and covariance to build the decision boundaries to discriminate different classes. Priors of the existing class labels are as follow:

Table 2:Prior probabilities of groups

1	2	3	4	5	6	9	10	16
.572	.087	.036	.036	.036	.051	.021	.117	.042

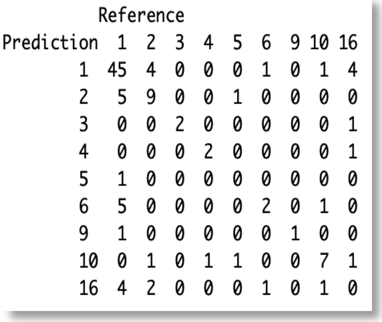
The prior probabilities once again confirm that the dataset is not balanced. As it is shown in Table 2, %57.2 of observations in the training data corresponds to class 1 which is significantly greater than others. The second thing that you can see is the Group means, which are the average of each predictor within each class. The high correlation between the variables could show the influence of each variable for a specific category. The proportion of traces for all LDs shown in Table 3 suggest that around %50 of variance explained through the first LD which has a great difference with other LDs.

Table 3:Proportion of trace

LD1	LD2	LD3	LD4	LD5	LD6	LD7	LD8
.5105	.1985	.0768	.0675	.0489	.0426	.0318	.0233

Dataset split into 70/30 split in which %70 is the training data, and the rest is the holdout test data. The confusion matrix of predicted and actual labels shown in Figure 1. The model can predict the first class good which could be related to the high number of observations from this class. On the other side, our model has a hard time to predict class 16 and 5. Our model has achieved an overall accuracy of 0.6415.

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 9	Class 10	Class 16
Sensitivity	.7377	.5620	1	.6666	.0	.5	1.0	.7	0
Specifity	.7778	.9333	.99038	.99029	.990385	.94118	.990476	.95833	.91919



```
          Reference
Prediction  1  2  3  4  5  6  9 10 16
        1  45  4  0  0  0  1  0  1  4
        2   5  9  0  0  1  0  0  0  0
        3   0  0  2  0  0  0  0  0  1
        4   0  0  0  2  0  0  0  0  1
        5   1  0  0  0  0  0  0  0  0
        6   5  0  0  0  0  2  0  1  0
        9   1  0  0  0  0  0  1  0  0
        10  0  1  0  1  1  0  0  7  1
        16  4  2  0  0  0  1  0  1  0
```

Figure 1: Confusion Matrix

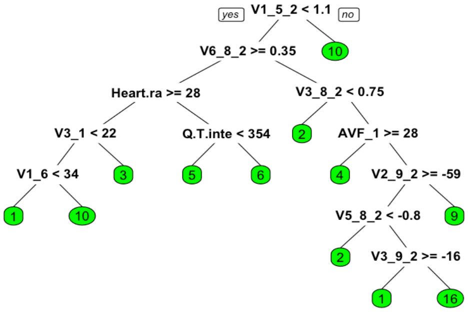
	Low	High
Low	179	64
High	58	77

Quadratic Discriminant Analysis

QDA aims to find the decision boundaries by considering the mean and covariance of each class. While it is a more intuitive method in comparison to LDA, there is a need for enough observations within each class to compute the covariance matrix. As the number of variables is large concerning the total number of instances and there are not enough observations for each level, the original data could not be used with QDA. One way to solve this issue is to use Diagonal Quadratic Discriminant Analysis (DQDA) since this method has less number of parameters to determine relative to ordinary QDA by making the off-diagonal elements equal to zero. However, DQDA package works for the early versions of R and not the most up to date ones. The alternative is to process the data so that the assumptions of the QDA method has been preserved. The classes divided into observations with “low” and “high.” Merging classes into two groups solved the original problem about enough observations for each category. Cross-validation metric within QDA method in MASS package has been used to produce more accountable results. Due to the limited number of instances we were not able to split the data into training, and holdout test data and all the training data used in all phases. The confusion matrix has shown in Table 5. QDA achieved the accuracy of 0.6772 percent, and the misclassification error rate is about 0.33 percent. The misclassification error could be related to the fact that we have to merge different classes into the same category which could make it challenging to do the classification. Moreover, the lack of enough observation to test our model with is another point which could impact our results.

Decision Tree

As the decision tree could handle different types of data and it does not have strict prior assumptions like QDA or LDA, we used the complete dataset without removing any of the classes to build our decision tree. 10-fold cross-validation has been used to construct the tree model. The final tree model has shown in Figure 2. The leave nodes show the different class labels. Just like QDA and LDA, data has been split into training and testing data (70/30) to assess the performance of our model more accurately. It is possible to prune our tree to make it more stable. However, for this project it has been decided to continue without pruning. The main reason for this decision is that by pruning the tree, it is possible that some of the classes, such as 16, 2 and 1, are not classified accurately. The decision tree achieved %75 accuracy and misclassification error around %25.



	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 9	Class 10	Class 16
Class 1	70	5	0	3	0	0	0	3	3
Class 2	0	6	0	0	0	0	0	0	0
Class 3	0	0	1	0	0	0	0	0	0
Class 4	0	1	0	1	0	0	0	0	1
Class 5	0	0	0	0	1	0	0	0	0
Class 6	0	0	0	0	2	7	0	2	0
Class 9	0	0	0	0	0	0	1	0	0
Class 10	2	0	0	0	0	0	0	9	1
Class 16	1	1	3	0	0	0	1	1	1