Investigation on Multi-Label Classification Using Multivariate Datasets

Hai-Niu (Johnny) Xu
Undergraduate Student (Statistics)
University of California, Davis
scxu@ucdavis.edu

Abstract

This midterm project focuses on studying method of conducting multi-label classification using multivariate dataset. To be specific, the intrinsic difference between variables of a dataset is studied and the result of this study is integrated with classification theories to come up with a classification pipeline that is suitable for the particularly datasets that are been studied. Futhermore, Principal Component Analysis (PCA) is employed as a dimension reduction technique to handle the multivariate datasets as well as to visualize the potential clustering of the data samples, which will in turn provide more intuition and information on conducting multi-label classification on the specific datasets. We see that model selection techniques such as Cross-Validation, which prevents our model from over-fitting the training data, are vital components of seeking the perfect predictive model from the multivariate dataset. Furthermore, we see that, as per the No-Free-Launch theorem, there is no universally one best classification method for multi-label dataset and the method to be employed highly depends on the characteristic of the particular dataset.

1 Introduction

With the advancement in technology, classification using computer algorithm and massive dataset has become a new trend and numerous methods have been proposed thereafter. Among the classification problems, multi-label and multivariate classification are one of the most difficult problems. While the binary classification problems can be easily handled with the Bayes optimal classifier, it does not have a good performance on multi-label classification due to the increased complication. The difficulty of handling multi-label classification is that the subtle intrinsic differences between labels are often hard to detect. Failure to detect the nuances between labels will result in labels being mis-classified to its similar kinds thus result in a low classification accuracy rate. In addition, the complication with multivariate classification comes with the issue of over-fitting. The popular and the most prominent classification methods such as logistic regression, decision tree, etc. are not immune to over-fitting thus such issue will also needs to be addressed while conducting classification. In practice, however, most classification problem are multi-label and multivariate. Some common classification scenarios includes classify images, service ratings based on user comments, performance of sport player based on their game statistics, etc. Therefore, investigate and gain more understanding in multi-label classification using multivariate datasets is essential to improve one's ability to utilize classification in real world problems.

In this study, the two multivariate, multi-label datasets to be studied are the Seeds dataset and the Automobile dataset both of which are obtained from the Machine Learning Repository of the University of California, Irvine (UCI). The major difference between the two datasets is that the variables in the Seed dataset are all numerical and continuous random variables whereas the Automobile dataset contain numerous categorical variable. Thus, the Seed dataset is used to study the

Midterm Project Report for STA160, Data Science Practices. Instructor: Prof.Fushing Hsieh

different aspects that we may see the intrinsic differences between different label classes since it is easier to compare the difference between numerical variables than the difference between numerical and categorical variables. Because of the fact that the study done with the two datasets are not closely related, the two datasets will be introduced separately in their own sections. Furthermore, depends on the characteristic of the two datasets, different classification methods are employed. For the seeds dataset, since the variables are all numerical, the K-Nearest-Neighbors algorithm was employed after studying the intrinsic differences between different labels. For the Automobile dataset, on the other hand, contains a mixture of numerical and categorical variables. Therefore, the two classification algorithms to be employed here are the multinomial logistic regression method and the AdaBoost Decision-Tree method. All of these procedures and discoveries will be analyzed in details in the later sections.

2 Analysis on the Seeds Data

2.1 Exploratory Data Analysis

The Seeds dataset obtained from the UCI machine learning repository includes 8 variables that documents the characteristics of kernels of three types of wheat: Kama, Rosa and Canadian. Of the 8 variables, 7 variables contains information on the physical trace of the different kernels and 1 variable indicates the label of the specific kernel. The sample size of this dataset is 210 in total with 70 samples from each wheat. This indicates that the label of wheat kernel follows a uniform distribution in this dataset. While there is no information on the popularity of the three type of wheat, the fact that this dataset contains equal number of samples from each wheat suggests that theses three types of wheat are equally popular. Furthermore, to get a comprehensive grasp of the dataset, the basic summary statistics of each variable are listed in Table 1.

	area	parameter	compactness	length	width	asymmetry	groove length
MEAN	14.85	14.56	0.87	5.63	3.26	3.70	5.41
STD	2.91	1.31	0.02	0.44	0.38	1.50	0.49
MIN	10.59	12.41	0.81	4.90	2.63	0.77	4.52
25%	12.27	13.45	0.86	5.26	2.94	2.56	5.05
50%	14.36	14.32	0.87	5.52	3.24	3.60	5.22
75%	17.31	15.72	0.89	5.98	3.56	4.77	5.88
MAX	21.18	17.25	0.92	6.68	4.03	8.46	6.55

Table 1: The mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and max of the variables in the Seeds dataset.

By looking at the mean and the standard deviation, we see that the variable *Compactness* have the most condensed distribution across labels. This suggest that this variable may provide very few information on intrinsic information of the wheat kernels. On the other hand, we see that *Asymmetry* has the largest variation across labels, which suggest that this variable could be very helpful with providing intrinsic information. To further examine the distribution of the 7 variables a complete density plot for each variable of each label is plotted. For the sake of the conciseness of the report, this figure is included in the Appendix, Figure 7, in the back of this report. Among the 7 variables across different labels, we see that 4 variables' distributions are rather distinctive among different labels as shown in Figure 1.

2.2 Search for Intrinsic Differences Among Distinct Labels

Observe that in Figure 1, the density plot for variable 0 and variable 1 shows that the density of these two variables of label 2 and label 3 are almost disjoint. This observation is confirmed by the Wasserstein distance between the three distributions. Notice that the Bayes Optimal Classifier for binary classification, which is what are are facing right now, is defined as

$$f^*(x) = \begin{cases} 1 \text{ if } \eta(x) \geq \frac{1}{2} \\ \\ 0 \text{ otherwise} \end{cases}$$

where $\eta(x) := P(Y = 1|X = x)$

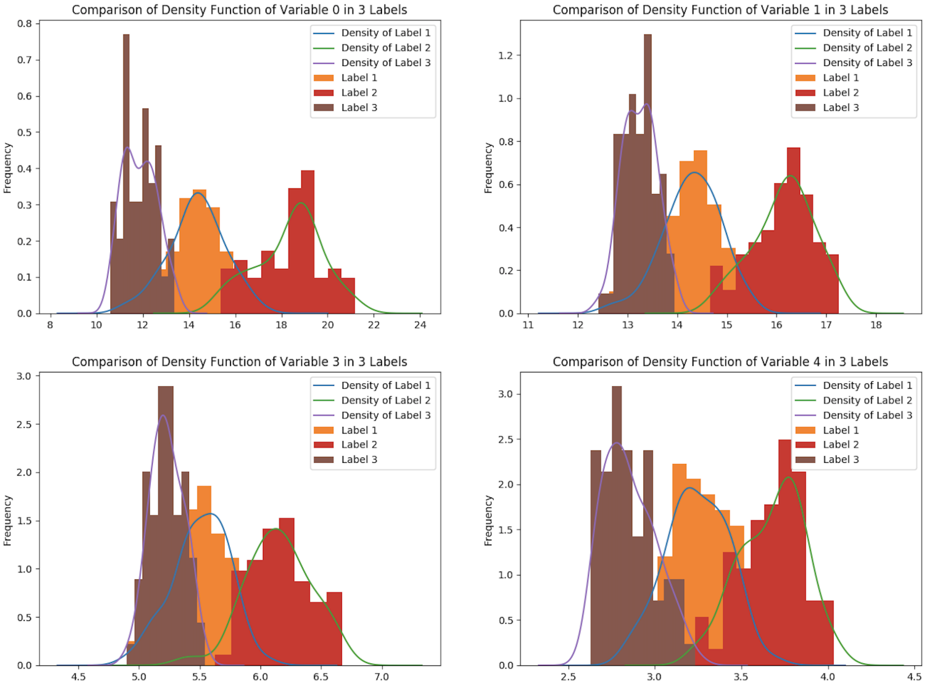


Figure 1: Variables that have distinctive distribution among different labels. 0: area, 1: parameter, 3: kernel length, 4: kernel width

Therefore, this observation allows us to conclude that the intrinsic difference between label 2 and label three is vibrant and a binary classification based solely on these two variable shall yield a nearly 100% accuracy rate in theory. In addition, the density of variable 4 of label 2 and label 3 is also roughly disjoint but not as much as that of variable 0 and 1. Therefore, although variable 4 would also do a good job classifying label 2 and 3, we still go with either variable 0 or variable 1 and the reason that we do not use both variable 0 and variable 1 will be discussed through their correlation.

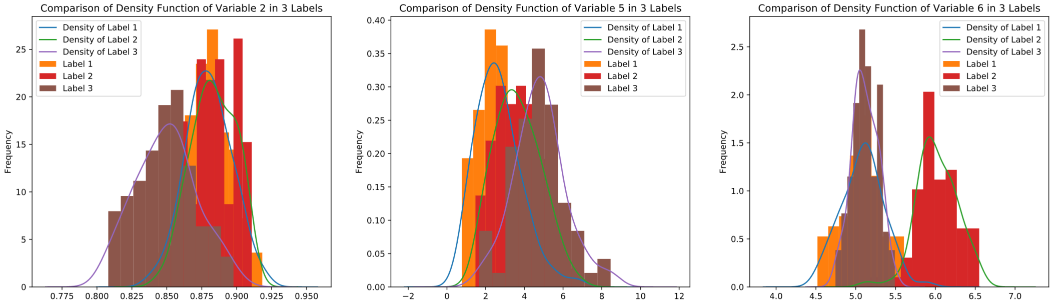


Figure 2: Variables that have intertwined distribution among different labels. 2: compactness, 5: asymmetry, 6: groove length

Besides the variables that have rather distinct density, Figure 2 shows another 3 variables which of whom have densities that are severely intertwined with each other. From Figure 2, we can see that the density of these three variables severely overlaps with each other making it hard to determine the density from which the variable was drawn. Therefore, as per the Bayes Optimal Classifier, which most binary classification methods are based on, these three variables shall not be used for classification.

Among the density comparison plot of all 7 variables of 3 different labels, there is one thing in common, which is the density of label 1. As the figures show, regardless whether the density plot of the variable among label 2 and label 3 can be perfectly separated or not, the density of variables of label 1 always have intersection with the density of variables of label 2 or label 3 making label 1 easily mis-classified as either label 2 or label 3 and the same logic applies to other variables. Therefore, we shall change a perspective in searching for the intrinsic difference between label 1 and the other two labels.

As shown in Figure 1, the correlation matrix of the 7 variables suggest that some variables are strongly correlated. For instance, the area, parameter, kernel length, kernel width, and groove width are all pairwise strongly correlated which makes perfect sense as they all measure the physical feature of the wheat. Due to the strong correlation between variables, we can already seen that it is unnecessary to utilize all variables to do classification. Therefore, this finding justifies the use of the Principal Component Analysis (PCA), which is the most frequently used dimension reduction tool in multivariate statistics.

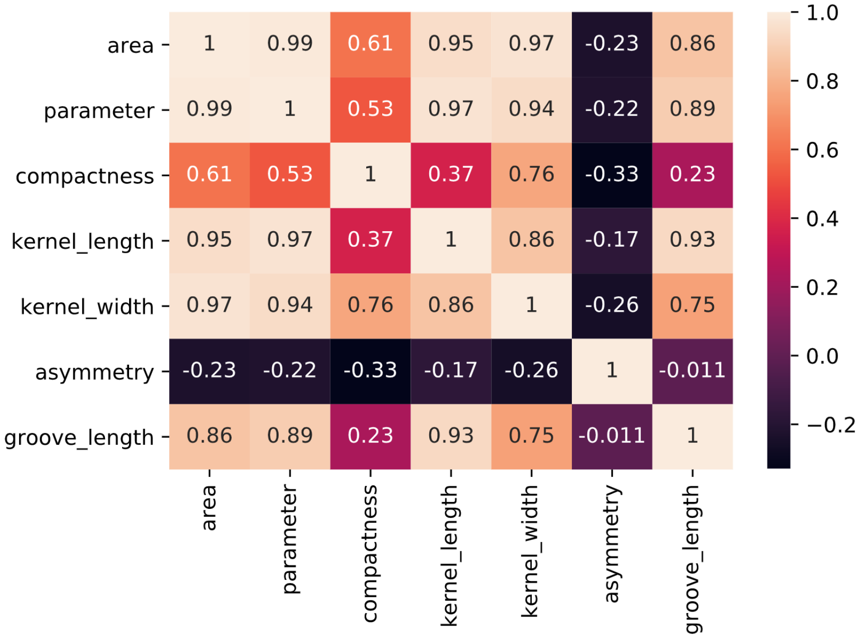


Figure 3: Correlation between variables of Seed dataset.

The number of Principal Components to be kept is set to 3 and PCA is conducted on the standardized dataset. First we look at the clustering of the data after PCA.

Like what we have discussed based on the distribution of the variables between different labels, the PCA clustering result of component 1 VS component 2 as well as the 3D plot of all three components shown in Figure 4 further confirms that data from label 2 and label 3 are indeed linearly separable and label 1, which located in between label 2 and label 3, is the obstacle here. Also from the clustering result, we see that although label 1 is not linearly separable from other labels. Although not linearly separable, label 1 does have its own cluster and it is most distinctive than the other two labels on the direction of component 2. This observation is further proved by the comparison of density of the three components across different labels.

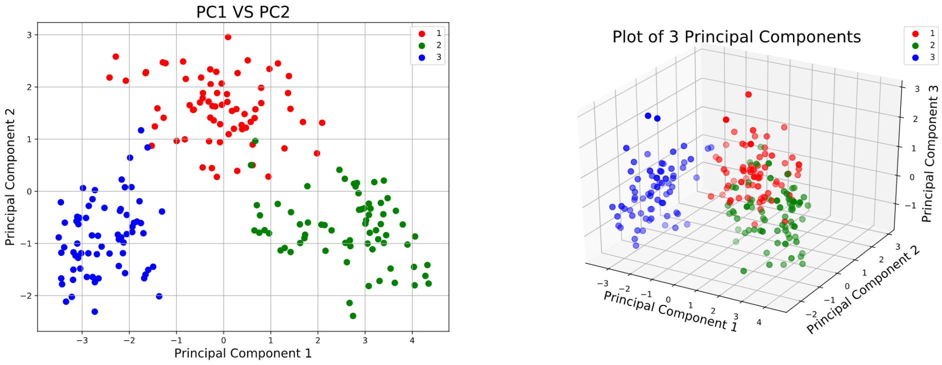


Figure 4: Visualization of PCA result.

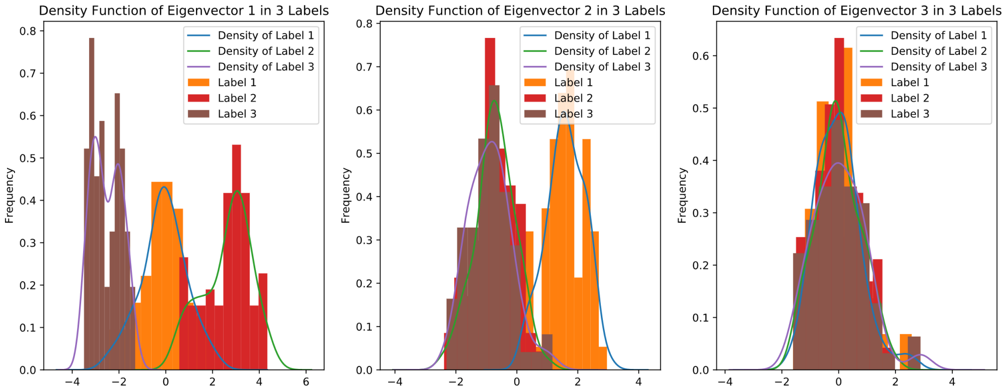


Figure 5: comparison of the distribution of 3 principal components of different labels

2.3 Conclusion of Analysis on the Seeds data

Hence, we have shown that there are indeed numerous intrinsic differences between the three labels. The difference between label 2 and label 3 are the most well-present as the comparison of the density of variables 0, 1, 3, and 4 of label 2 and label 3 shows that they are almost disjointed. As for label 1, we see that although there is no distinctive difference in terms of the observed variables, the difference between label 1 and the two other labels is shown in the PCA result, especially in the direction of the 2^{nd} principal component. To further examine these intrinsic differences, classification algorithms are used to test the distinction, which can be reflected by the accuracy rate.

While using the K-Nearest-Neighbor (KNN) method to classify label 2 and label 3 based on variables 0, 1, 3, 4 independently, all 4 classifiers yielded a 100% accuracy rate including variable 4, which of whom, although have slightly intersected densities, still give a perfect distinction between label 2 and label 3. As for label 1, the KNN algorithm is also used with only the eigenvectors corresponds to the second Principal Component and the classification has a 95% accuracy rate, which makes sense as the density of label 1 indeed has some overlap with the density of the other two labels in the case of the direction of the 2^{nd} principal component. Therefore, both classification results proves that the intrinsic differences discerned in this study is valid.

3 Analysis on the Automobile Data

3.1 Exploratory Data Analysis

The Automobile dataset obtained from the UCI machine learning repository includes 26 variables that documents the characteristics of 7 insurance risk ratings. Of the 26 variables, 25 variables contains information on both the numercal and categorical features of different cars and 1 variable indicates the label of the specific insurance risk. The sample size of this dataset is 205 in total with an unequal amount of labels. The distribution and number of samples of each label is displayed in Figure 6.

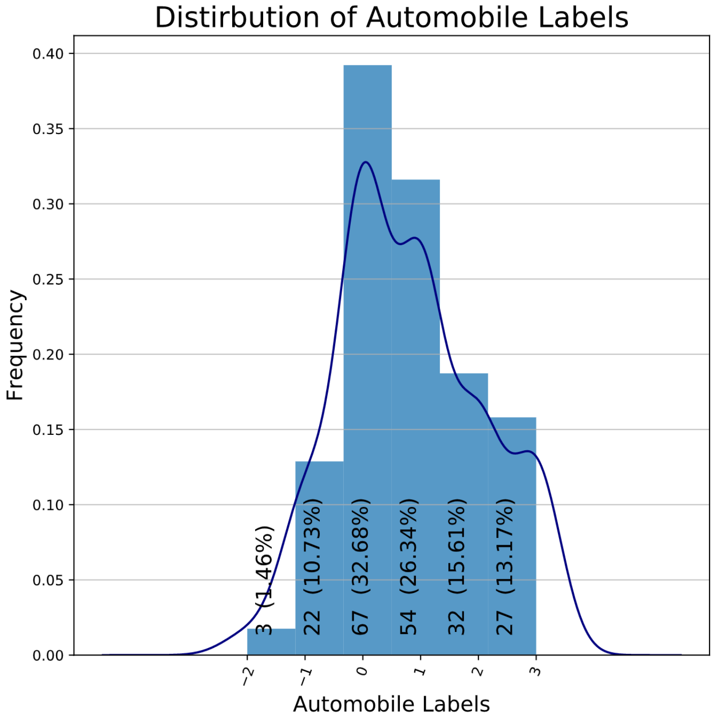


Figure 6: Distribution and number of samples of each label in Automobile data.

We can see that although there are 7 labels ranging from -3 to 3, the dataset actually contains 6 labels with no sample from label -3 and very limited sample from -2. Now we are facing a different scenario than that of the Seeds data where the labels are uniformly distributed. Here, the labels are roughly normally distributed, which also makes sense. Since the label here represents the insurance risk rating, therefore we would expect most cars fall under the category of a moderate rating and a few cars being extremely low risk or extremely high risk.

In addition, unlike the Seeds dataset which is perfectly dense, the Automobile dataset contains several missing values. There are 41 missing values in the feature *Normalized Losses*, which is the most among all feature, and several negligible missing data in other features. To see the detailed count of missing data, the table of the number of missing data is included in the Appendix, Table 2. Other than the missing values, here are some basic sample statistics of the numerical variables of the Automobile dataset are listed in Table 4. Further, since most data analysis tools employed here requires a dense dataset, the missing data are imputed based on the mean of each variable. Imputation based on the mean may not be the best option here but it is used to give a rudimentary result and different imputation method shall be tested in future study.

Variable	Number of Missing Data
normalized losses	41
number of doors	2
bore	4
stroke	4
horsepower	2
peak RPM	2
price	4

Table 2: Number of missing values and the corresponding variables of the Automobile data.

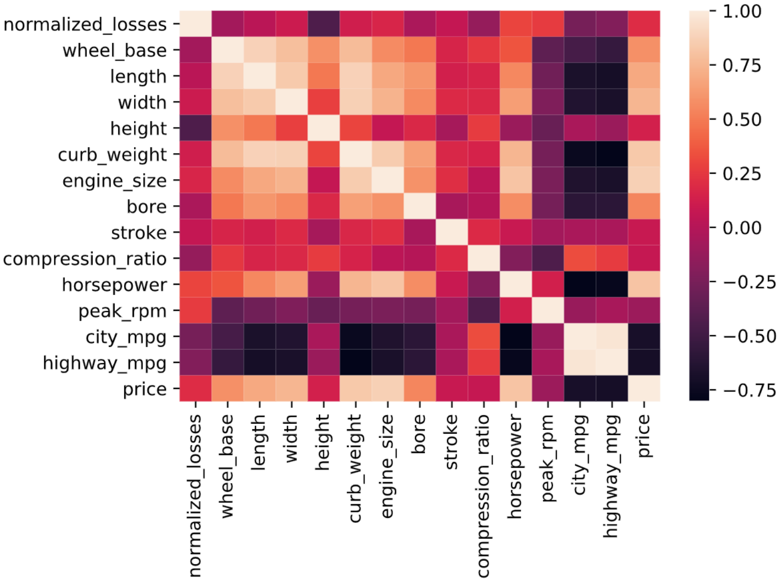


Figure 7: Correlation matrix of the numerical variables in Automobile Dataset.

Variable	Mean	STD	Median
normalized losses	122.000000	35.442168	115.00
wheel base	98.756585	6.021776	97.00
length	174.049268	12.337289	173.20
width	65.907805	2.145204	65.50
height	53.724878	2.443522	54.10
curb weight	2555.565854	520.680204	2414.00
engine size	126.907317	41.642693	120.00
bore	3.329751	0.273539	3.31
stroke	3.255423	0.316717	3.29
compression ratio	10.142537	3.972040	9.00
horsepower	104.256158	39.714369	95.00
peak rpm	5125.369458	479.334560	5200.00
city mpg	25.219512	6.542142	24.00
highway mpg	30.751220	6.886443	30.00
price	13207.129353	7947.066342	10295.00

Table 4: Test Statistic of Automobile Variables.

From the mean and standard deviation, we can see that some potential good variables for classification includes curb weight, engine size, compression ratio, horsepower, peak RPM, city MPG, and price. These variables have large variation which indicates that the within label distribution of these variables could be distinctive as we discovered from some variables in the Seeds dataset. Further, the correlation matrix of the numerical variables is shown in Figure 7.

We can see, there are still numerous variables that have strong correlation, which suggests that dimension reduction and model selection are options to handle this dataset. Furthermore, looking into the 10 categorical variables, we have the following categories:

Body Style: sedan (96), hatchback (70), wagon (25), hardtop (8), convertible (6).

Number of Doors: four (114), two (89).

Fuel System: mpfi (94), 2bbl (66), idi (20), 1bbl (11), spdi (9), 4bbl (3), spfi (1), mfi (1).

Aspiration: std (168), turbo (37).

Number of Cylinders: four (159), six (24), five (11), eight (5), two (4), twelve (1), three (1).

Drive Wheels: fwd (120), rwd (76), 4wd (9).

Engine Location: front (202), rear (3).

Make: toyota (32), nissan (18), mazda (17), mitsubishi (13), honda (13), volkswagen (12), subaru (12), volvo (11), peugot (11), dodge (9), bmw (8), mercedes-benz (8), plymouth (7), audi (7), saab (6), porsche (5), isuzu (4), jaguar (3), chevrolet (3), alfa-romero (3), renault (2), mercury (1).

Fuel Type: gas (185), diesel (20).

Engine Type: ohc (148), ohcf (15), ohcv (13), l (12), dohc (12), rotor (4), dohcv (1).

3.2 Multinomial Logistic Regression

In this case, since we have a mixture of numerical and categorical variables, classification methods designated to numerical classification such as linear regression is no longer applicable here. Therefore, one natural approach to handle such a mixture of data is logistic regression. Specifically, the first algorithm to be tested is the multinomial logistic method that is suitable for classifying multiple labels and can take on both numerical and categorical variables simultaneously. To make sure that the classification algorithm works on the dataset, the categories of categorical data are first converted from plain text to integer numbers that represent their categories. After that, $\frac{2}{3}$ of the data are used for training and the other $\frac{1}{3}$ of the data are used testing. The full model is fitted and tested using the testing dataset which yielded the following confusion matrix:

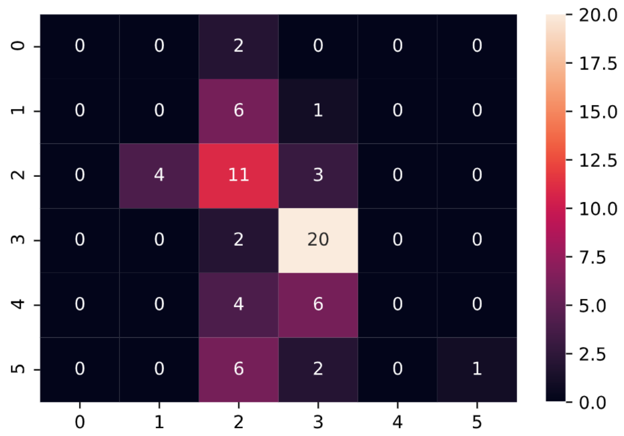


Figure 8: Confusion matrix of classification result based on the full logistic model.

From the confusion matrix, we see that the classification result is not prominent. In fact, the accuracy rate of classification using the full multinomial logistic model is only 47.0588%. Such a low accuracy rate suggests that there might be severe over-fitting problem with the full model. Therefore, model-selection using recursive feature ranking is conducted. To determine the best number of variables to be used, a grid search is conducted and the number of variable and the corresponding accuracy rate is

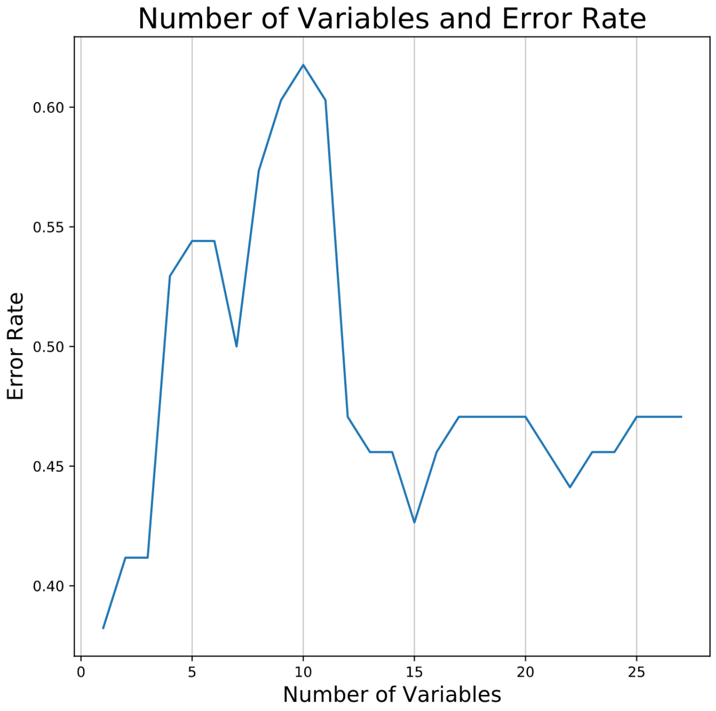


Figure 9: Number of variables in the logistic regression model and the corresponding accuracy rate.

From the grid search, we see that the multinomial logistic regression model performs the best when it contains 10 variables and the best accuracy rate it can achieve is 61.7647%, which still has a lot of room for improvements. Taking a look at the algorithm of the logistic regression method, which is basically fitting a linear model based on multiple log-odds ratios, it is noticeable that the primary goal of this model is also to minimize the least squares of the regression model to the sample data. Therefore, the hard-to-classify data points are not emphasized in the learning process.

3.3 AdaBoost Decision Tree

In the case of multi-label classification, as discussed in the Seeds dataset, the boundaries between different labels are often blurred. Therefore, to improve the accuracy of this method, the data points on the boundary, which are the hardest to be classified in this case, shall be emphasized in the learning process. With this intuition, Ada-boosted Decision Tree comes naturally to my mind. The Adaboost Decision Tree re-enforce the learning of hard-to-classify data points in the learning process thus allocates the computing power to the points that demand to be taken care of the most. With Decision Tree model, grid search method is used again to determine the optimal depth of the decision tree. As shown in Figure 10, we see that the Adaboost Decision Tree generally out-performs the multinomial logistic regression method. It reaches its peak performance when the depth of the tree is 4 or 5 and the corresponding accuracy rate is 82.3529%, which is a significant improvement comparing to that of the multinomial logistic regression.

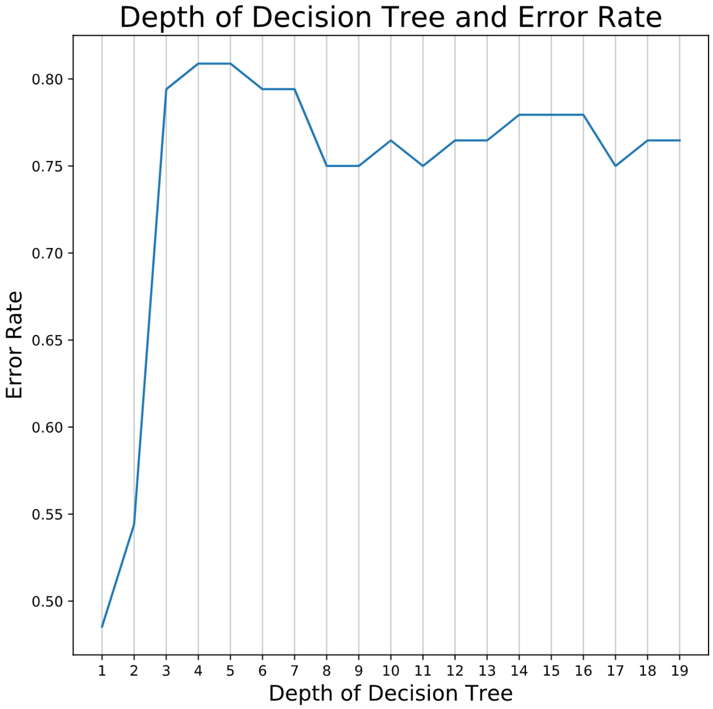
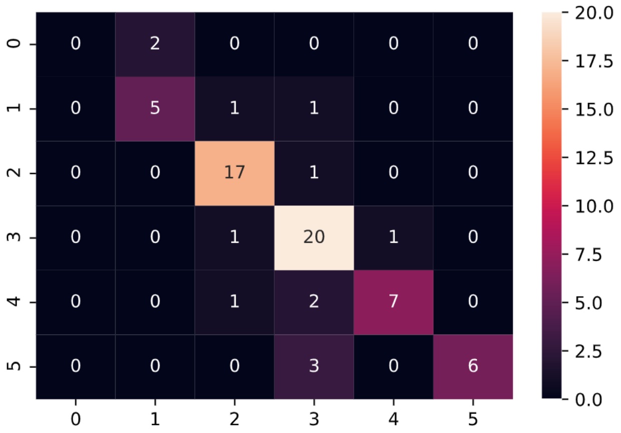


Figure 10: Depth of Decision Tree and the corresponding accuracy rate.

The corresponding confusion matrix shown in Figure 11 shows that the most difficult label to be classified correctly are label 0 and label 1. Since they are located in the middle of all variables, they will have trace of all the neighbor labels thus very easily mis-classified. Yet, we do see a huge accuracy leap from the "fair" multinomial logistic model that learn all the data points equally to the "unfair" Adaboost Decision Tree model that emphasize the data points that are hard to be classified. The Take away from this dataset is that model selection are always a must-have step for classification regardless of the numbre of labels and Adaboost Decision Tree is more suitable for multi-label dataset as it emphasizes learning on the hard-to-classify data points.



	0	1	2	3	4	5
0	0	2	0	0	0	0
1	0	5	1	1	0	0
2	0	0	17	1	0	0
3	0	0	1	20	1	0
4	0	0	1	2	7	0
5	0	0	0	3	0	6

Figure 11: Confusion matrix of the optimal Adaboost Decision Tree

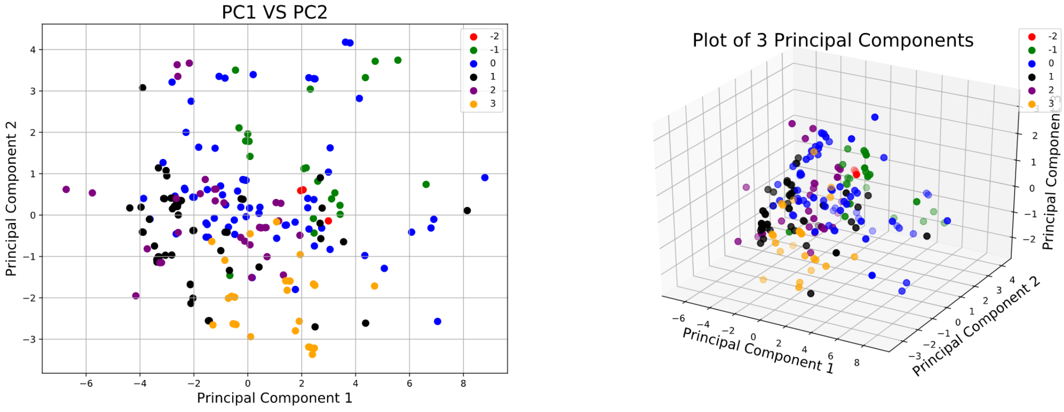


Figure 12: PCA result of Automobile data.

Further, as we have observed that the correlation between certain variables in the Automobile dataset is quite large. Therefore, PCA is employed here to reduce the dimension here. Due to the mixture of numerical and categorical data, the PCA is only conducted on the standardized numerical values of the Automobile data. The PCA result shown in Figure 12 explains why the logistic regression is having a hard time classifying the auto data. As we can see, even after dimension reduction, there is no obvious clustering between the 6 groups and they are all entangled with each other. Further, suppose we look at the data on the direction of each principal component individually, we are still unable to discern any distinction between different groups. Such scenario makes the logistic regression, which is essentially a linear classifier based on log-odds ratio and Bayes optimal classifier, very hard to distinguish between different classes.

Appendix

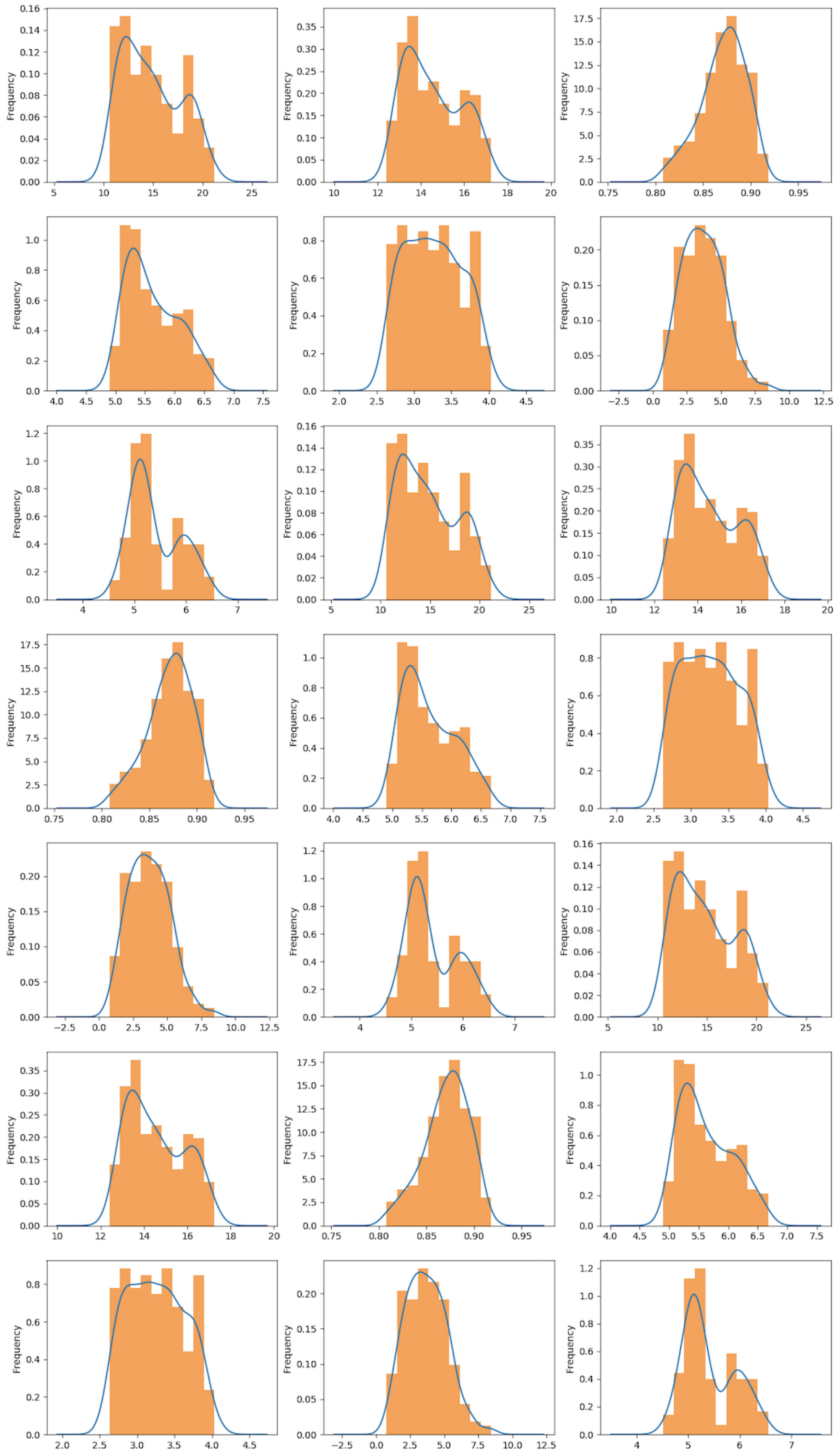


Figure 13: Distribution of each variables of different labels. The columns from left to right correspond to class 1 to 3. The rows from top to bottom corresponds to variables area ,parameter, compactness, kernel length, kernel width, asymmetry, and groove length