INTRODUCTION TO DATA MATHEMATICS

CLASSIFICATION PROJECT

SWD Inc.

Shuang Guan, Shutong Luo, Zhengneng Chen, Ziao Yan

supervised by
Prof. Kristin BENNETT
Prof. Bruce PIPER

March 31, 2016

1 INTRODUCTION

To indicate the class of a list of data points, there are two ways to achieve it. The Mean method and Fisher LDA method. Mean method works by finding the mean of all the data points, and using this number to separate the points to either 1 or -1 class. The Fisher LDA method is to find a line in the axis and make a projection for each of the data points then get the best way to separate the points into two classes. Both methods will have a error occur when it applies to the data and it is important to calculate these errors to figure out which is the better way to separates the two classes of data points. This data analysis process is going to find the size, mean, covariances, normal, threshold and the errors for both methods. Then compare the error on both train and test data calculated by two different methods.

2 DATA ANALYSIS

The data set have attributes of 41 and 1055 points in each class. Then, we find the mean in $+1$ class and -1 class. The size of the positive class is 356, and for negative is 699. Then, the size of V is 400. In Figure 1 below is the heatmap of covariance of class 1, and Figure 2 shows the covariance of class -1. In Figure 3 it is the covariance of V.

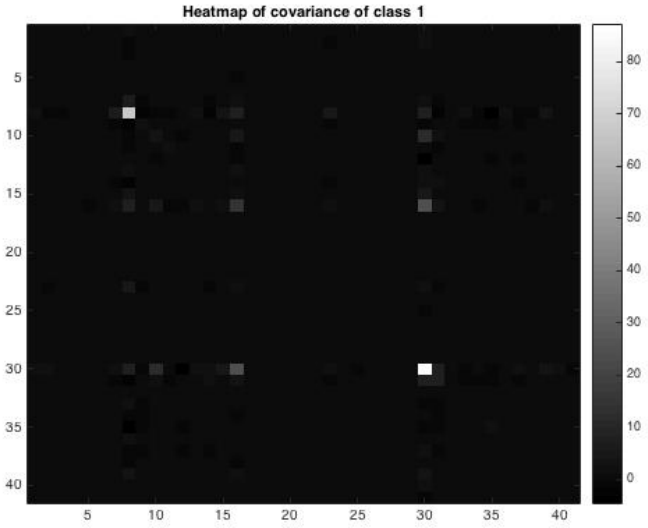


Figure 1: Heatmap of covariance of class 1

3 COMPARE WITH TWO MODEL

Using the predictive model, our company found that the total error of the mean method would be 33.37% on the training data and 41.90% on the testing data. On the contrary, using the Fisher LDA method it has a total error of 13.79% on training data and 16.19% on the testing

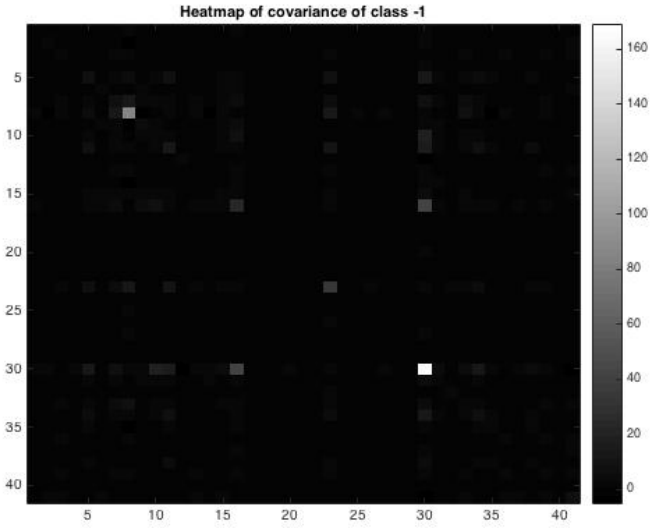


Figure 2: Heatmap of covariance of class -1

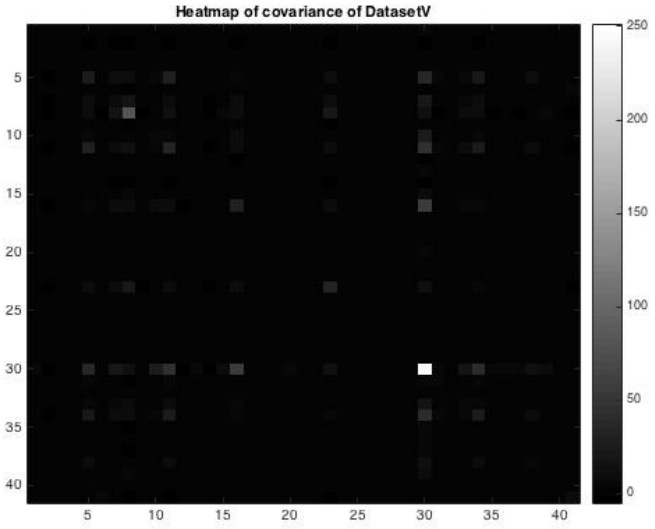
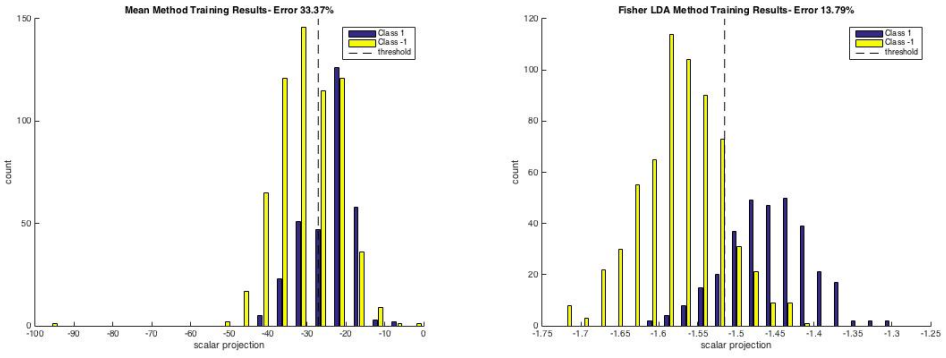


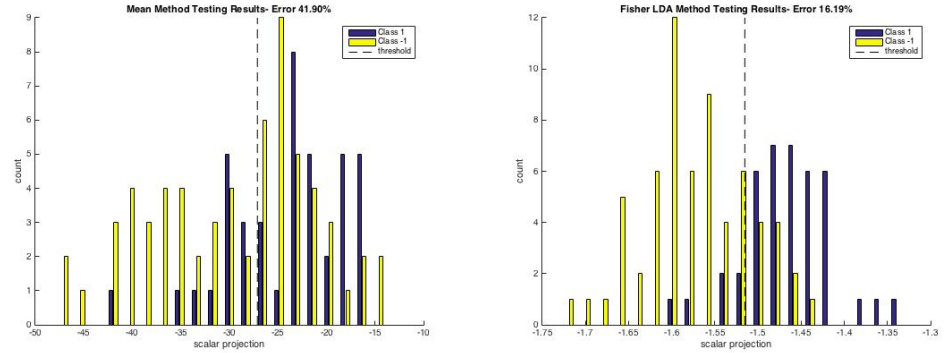
Figure 3: Heatmap of covariance of DatasetV

data. The error is much fewer in Fisher LDA method than it is in mean method, thus we believe that the Fisher LDA method is preferable. The two methods are different in calculating the hyperplanes. The mean method uses the vectors as the difference between the means of the two classes which gives us a threshold of -27.1221 and the fisher LDA method uses the vector created by Fisher Linear Discriminant which gives us a threshold of -1.5153.



(a) Mean method (b) Fisher LDA method

Figure 4: Training result



(a) Mean method (b) Fisher LDA method

Figure 5: Testing result

4 THE PREDICTIVE MODEL

Our predictive model would be Fisher LDA method. First we need to find the mean of two classes and use that to compute the covariance. Using the covariance of one of the two classes multiply the transpose of itself and add the covariance of another one multiply itself, we could find the within class scatter matrix S_w. Divide it by the difference of the mean of two classes we could find the hyperplane and find the normal w easily. Using the summation of the mean of two classes and dividing it with $2w$ we could find the threshold t.

5 ERROR RESULT

ON DATASETA Using the Fisher LDA method, we have 14.3258% error in class 1 and 13.877% error in class -1. The total error is 14.028%.

TOTAL ESTIMATE We tested a set of data in 'DatasetA.csv' by using Fisher LDA method and have a conclusion that Fisher LDA method has a total error in range of $10-20\%$ varying due to data set's size. This estimation is found by multiple attempts in analyzing different sizes of data from 'DatasetA.csv'.

CLASS 1 AND -1'S ERROR ESTIMATE As we tested before, training set has a size of $1055-105=950$ and has an error of 13.79%. Data in 'DatasetV.csv' has a size of 400 which is fewer than 950 but greater than 105. It's reasonable to estimate that the total error of using Fisher LDA method to test 'DatasetV.csv' is around 15% which between 13.79% and 16.19%. As we already tested that class 1's error is 14.3% and class -1's error is 13.9%. We can reasonable induct that class 1 and class -1's error on 'DatasetV.csv' is around 14% by using Fisher LDA method.

6 CONCLUSION

Our project provides a data analyzing model using Fisher LDA method. Employing Fisher LDA method, we can find a different hyperplane and threshold with Mean method and use these to analyze the testing data. Compared to Mean method, Fisher LDA method could provide more accuracy with nearly 20%. Our model could make the prediction of biodegradability more reliable.

To improve the future performance of the model, more properties of the chemicals should be studied, as the biodegradability could be influenced by complicated factors.

7 A USEFUL CLASSIFIER-SVM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

For a linearly separable set of 2D-points which belong to one of two classes, find a separating straight line.

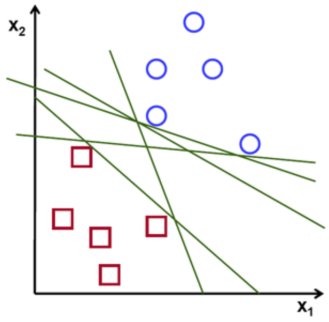


Figure 6: There are multiple choices to classify

In the above picture you can see that there exists multiple lines that offer a solution to the problem. Is any of them better than the others? We can intuitively define a criterion to estimate the worth of the lines:

A line is bad if it passes too close to the points because it will be noise sensitive and it will not generalize correctly. Therefore, our goal should be to find the line passing as far as possible from all points.

Then, the operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. Twice, this distance receives the important name of margin within SVM's theory. Therefore, the optimal separating hyperplane maximizes the margin of the training data.

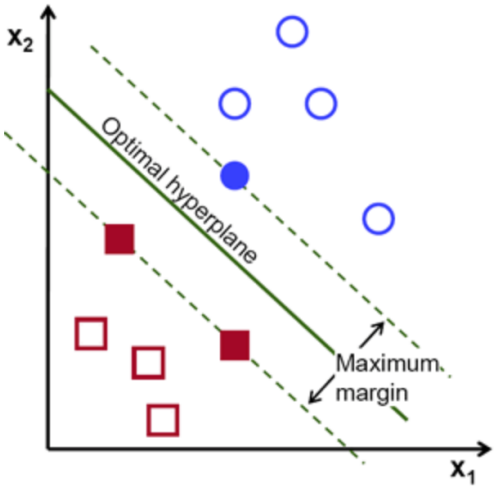


Figure 7: To find maximum margin

8 TESTING RESULT OF SVM

Based on our Matlab result, we produce a lower error rate of 13.33% on testing set of DatasetV. This rate is about 3% lower than 16.19% of Fisher LDA method. Thus we can tell that to estimate DatasetV, Linear SVM Classifier will make a better result than Fisher LDA method.