LINEAR DISCRIMINANT ANALYSIS AND CLASSIFIER ASSESSMENT

ABSTRACT

PURPOSE: The purpose of this experiment is to explore linear discriminant analysis by finding a linear separation of the data and performing methods of assessing the accuracy of this separation.

METHODS: Methods to implement this objective included principal component analysis as well as computations for Fisher's linear discriminant. Linear discriminant analysis was assessed by finding optimal threshold values with their corresponding confusion matrices and plotting a receiver operator characteristic curve for both of the labels.

RESULTS: Results show two plots describing the data with reduced dimensionality from PCA, and two plots showing the label grouping of the data after projecting the data onto the Fisher's linear discriminant axis. Two ROC curves of the TPR and FPR values from different thresholds are shown along with two tables that correspond to the confusion matrices of the optimal threshold values for each of the classifiers.

CONCLUSIONS: Results show that the diabetes classifier does an overall better job at classifying the labelled data on a range of threshold values. An optimal threshold value can be found for both sets of labelled data that produces a reasonably accurate classifier.

INTRODUCTION

The objective in this assignment is to use labelled data to find a linear separation of the classification variables and further to implement methods to assess the model used to separate the labelled data.

Principal components analysis (PCA) is a method used to take a higher dimensionality of data and reduce to a lower dimension to better interpret the data while minimizing the loss of information. PCA generally does not consider any labels within the data, it tries to explain the data given without any interpretation or prediction, which serves little use for labelled data. In PCA with labelled data, we want to combine the use of PCA with all the data with PCA using the means of partitioned data to find the best separation of labels. For a tall-thin data set matrix A, we can find the zero-mean matrix $M = A - \vec{1}A$. The scatter matrix S for the entire dataset is defined as $S = M^T M$. We can partition the data by its labels into two subsets A_1 and A_2, where we can then calculate the means M_1 and M_2 and scatter matrices S_1 and S_2 in an analogous fashion to the original data set. These subsets will allow us to calculate the scatter within the labels $S_W = S_1 + S_2$ and the scatter between the labels $S_B = \begin{bmatrix} \bar{A}_1 - \bar{A} \\ \bar{A}_2 - \bar{A} \end{bmatrix}^T \begin{bmatrix} \bar{A}_1 - \bar{A} \\ \bar{A}_2 - \bar{A} \end{bmatrix}$. Once we have these scatter matrices, we then want to find Fisher's linear discriminant. Fisher's linear discriminant has two goals, one is to minimize the within label scatter and to maximize the between label scatter. These two goals are achieved by maximizing the ratio of Rayleigh quotients for S_W and S_B and finding the largest eigenvector. When we use this vector, we are performing linear discriminant analysis (LDA). Using this axis, we are then able to find a separating hyperplane of the data.

Given a hyperplane, we may now want to assess how well this hyperplane separates the data. One method of doing this is to compute a confusion matrix. Moreover, if we change a bias scalar, we can further optimize the classifier. We can compare the labels of the data to the predictions of the classifier to find the positive and negative instances as well as the true positives, false positives, true negatives, false negatives. These values can then be arranged into a confusion matrix to better visualize the model. The confusion matrix allows us to calculate the true positive rate $TPR = \frac{TP}{P}$, otherwise known as the sensitivity, and the false positive rate $FPR = \frac{FP}{N}$, otherwise known as the specificity or Type 1 error. We can also find the accuracy of the model, which is the number of classifications that were predicted correctly. The accuracy is computed as $ACC = \frac{TP+TN}{P+N}$.

The receiver operator characteristic (ROC) is another method to analyze a classifiers performance visually. The ROC operates by plotting sensitivity (TPR) with the specificity (FPR) by using threshold values. Using the TPR and FPR will give us a score, which is a ROC vector. To create a curve using data, we need to use varying threshold values named theta, in which the classifier will say that a point is +1 if the score is above the threshold value, and -1 if the point is below the threshold value. We can then create a relative confusion matrix after using these different threshold values to find the TPR and FPR. After plotting scores for different threshold values, we can assess the efficacy of the classifier by finding the area under the curve (AUC). The closer the AUC is to 1, then the closer the classifier was able to perfectly separate the data.

A testable hypothesis for this experiment is to first reduce the data using PCA, then perform linear discriminant analysis. After a classifier is established, we can assess the accuracy by various methods such as a confusion matrix or an ROC curve.

METHODS

To implement linear discriminant analysis, we first need to load the data set that we are going to use. In this data set, the last column of the dataset corresponds to whether the patient was diagnosed with early-stage diabetes of type two, and the second to last column is whether the patient has clinical obesity. We will be using these variables to carry out our linear discriminant analysis. Using these variables, we fist reduce the dimensionality of the data down to two dimensions using principal component analysis (PCA) with the pca function in MATLAB.

Once we have the reduced dimensionality, we can then find the linear discriminant analysis on both of the data variables. We achieve this in the a4q1 function, where we pass through the two-dimensional data, and the labels of the classifier column. In this function, we first need to find the Fisher's linear discriminant analysis axis with each of the variables, so the lda2class function is called. In this function, the means of the datasets are calculated so that we can then find the within-class means and scatter matrices, and the between-class scatter matrix. Using MATLAB's eigs function passing through the matrix division of the within scatter and between scatter matrices. The largest vector then corresponds to the largest eigenvector of the Rayleigh quotient. Once the axis is found, then the scores of the LDA can be computed by a projection of the dataset to the axis.

After calculation of the important information needed to describe the data, we can then plot the findings. For the diabetes and obesity variables, first the reduced dimensionality from PCA is plotted, then the LDA scores are plotted grouped by their labels after projection. These plots are found using MATLAB's gscatter function.

The next task is to find the ROC curves for the data, as well as finding the confusion matrices for optimal threshold values. The roccurve function is used to achieve this. A label vector and score vector are inputs to the function, which are then sorted. This function iterates through different threshold values in bvec which are found by taking all the unique score values from the score vector zvec. For each threshold value, a confusion matrix is calculated as described later in this section, followed by the computations of the true positive rate and the false positive rate, which are appended to two separate vectors. From the confusion matrix we can also calculate the accuracy of the threshold value. With a conditional statement, the accuracy value is compared to the accuracy of the previous threshold value and updated if the accuracy increases. After updating the accuracy, the corresponding threshold value is stored since it will be the optimal value. Once the optimal threshold value is found, the corresponding confusion matrix is then displayed. Using the true positive rate vector and the false positive rate vector, the area under the curve is computed using the aucufroc function, which implements the trapezoidal rule to find the area under the curve.

Within the roccurve function, the confmat function is called to find the confusion matrix for a given threshold value, labels, and scores. First, the quantization vector qvec is found by using the threshold value theta. If a score is above a threshold value, then it receives a label of +1, otherwise it will receive a label of -1. Once the quantization vector is found then the confusion matrix can be calculated by comparing qvec to the labels of the data yvec. In a for loop, each entry of the vectors are compared, and each comparison will fall in one of four cases; true positive, false positive, true negative, and false negative. These cases are gradually incremented until the iteration stops, and the confusion matrix is then set up using these values as the entries.

After the roccurve function is called, then the relevant findings are then plotted and displayed. For the diabetes and obesity data, the optimal threshold values are shown with their confusion matrices, as well as the scalar value for the area under the curve. Then, the ROC curve is plotted by false positive rate along the x-axis and true positive rate along the y-axis. After this plotting, all of the desired results have been displayed.

RESULTS

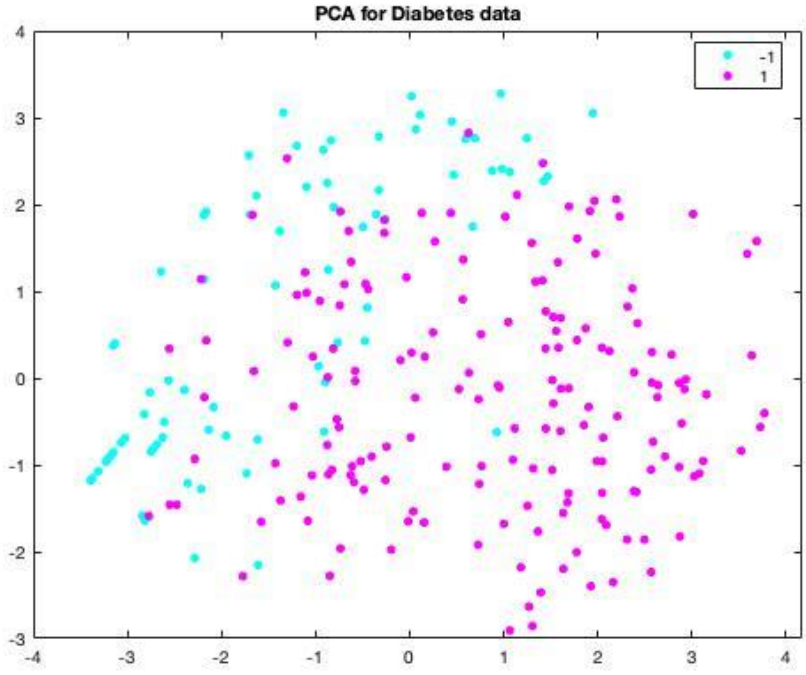

Figure 1. Shows the principal component analysis for the diabetes data. Reduced dimensionality down to two dimensions. Data is grouped by label, magenta points represent positive diabetes labels and cyan points represent negative diabetes labels.

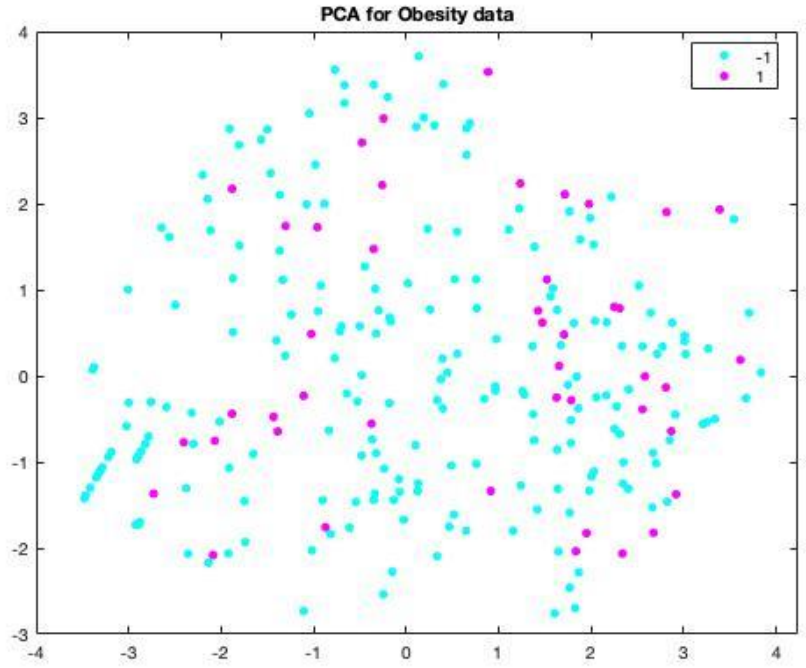

Figure 2. Shows the principal component analysis for Obesity data. Reduced dimensionality down to two dimensions. Data is grouped by label, magenta points represent positive obesity labels and cyan points represent negative obesity labels.

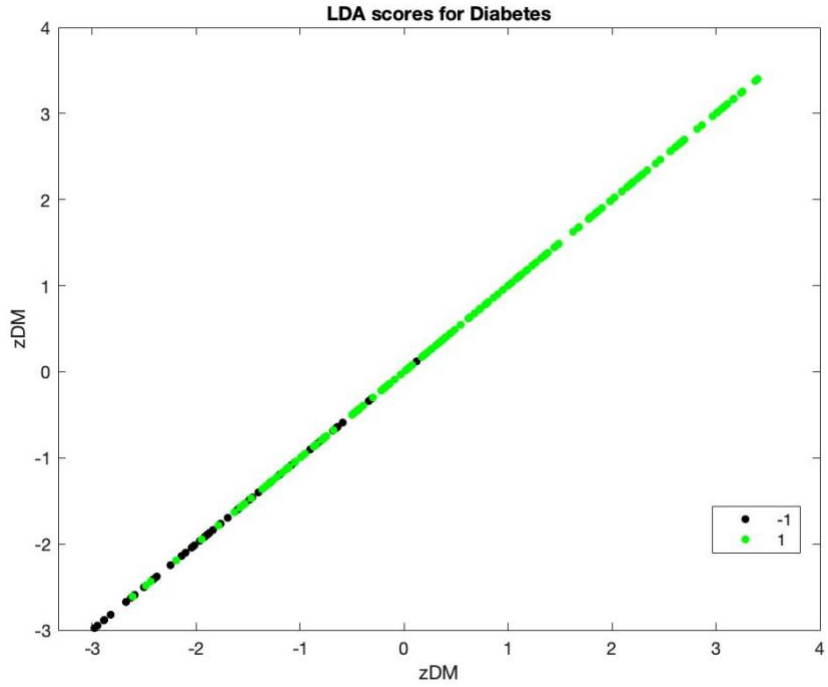

Figure 3. Shows the projected scores of the LDA grouped by label for diabetes data. Data is grouped by label, green points represent positive diabetes labels and black points represent negative diabetes labels.

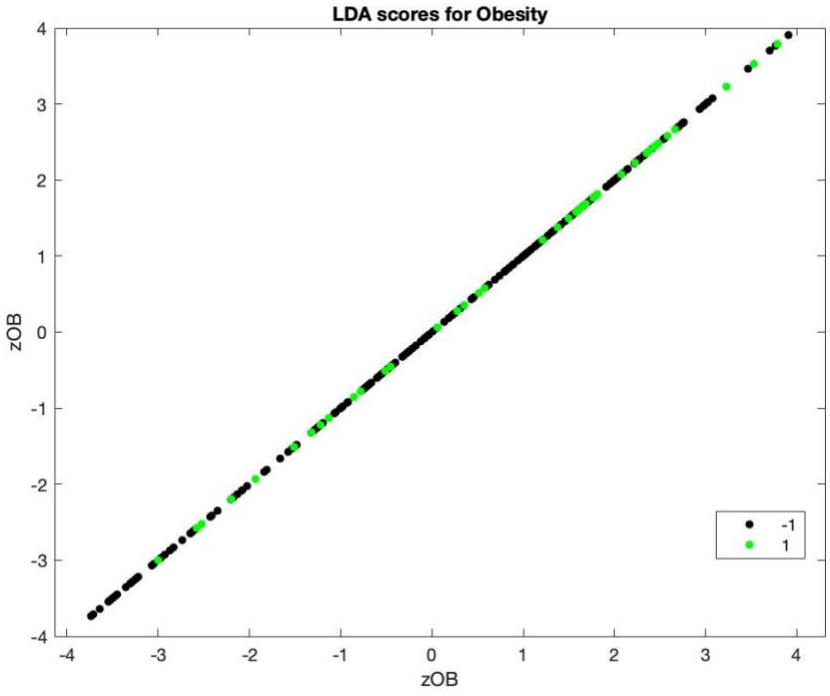

Figure 4. Shows the Projected scores of the LDA grouped by label for obesity data. Data is grouped by label, green points represent positive obesity labels and black points represent negative obesity labels.

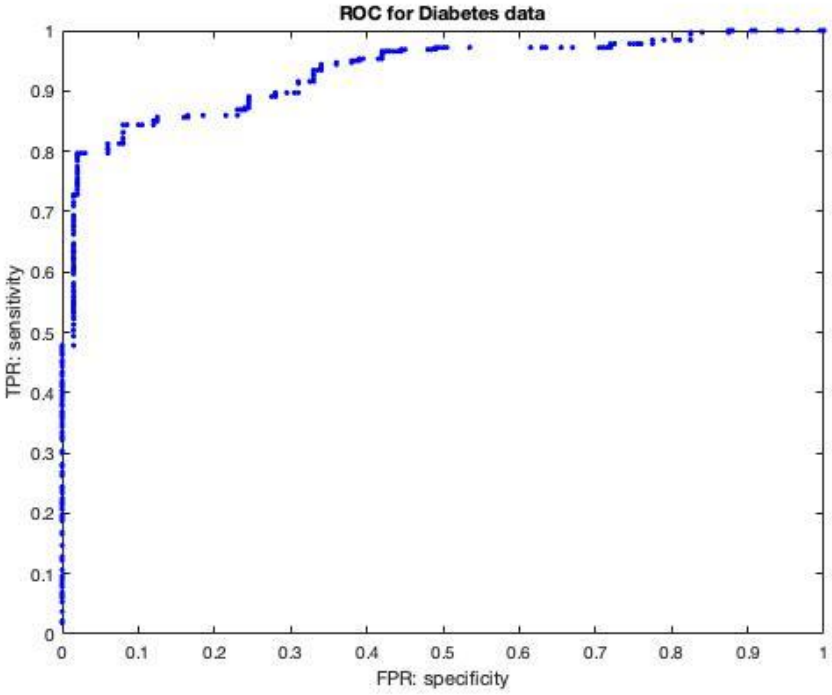

Figure 5. Shows the ROC Curve for diabetes data. Area under the curve for this ROC curve is calculated to be AUC = 0.9309.

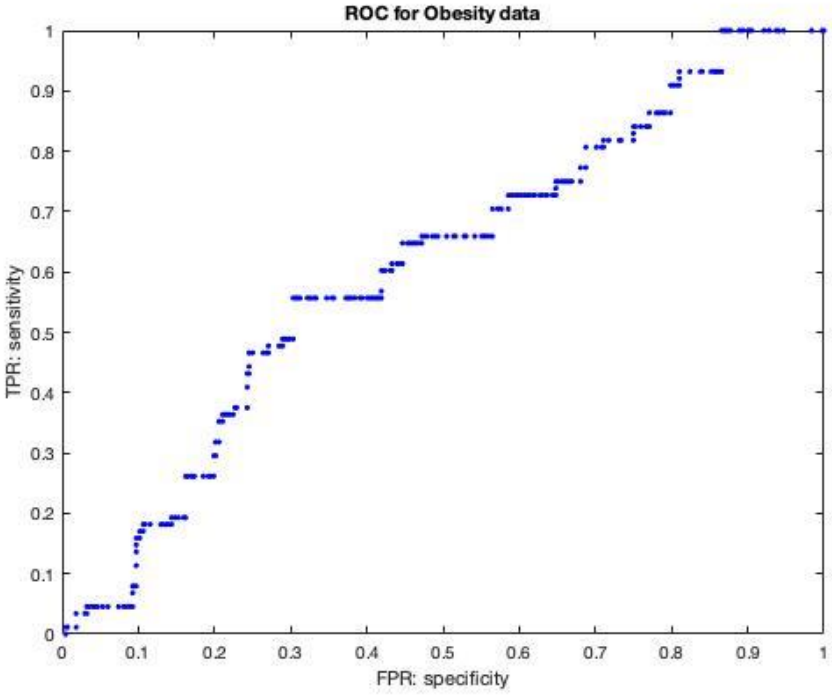

Figure 6. Shows the ROC curve for Obesity data. Area under the curve for this ROC curve is calculated to be AUC = 0.6071.

Table 1. Shows the confusion matrix for the optimal threshold value of the diabetes label. Computed using LDA which produced a threshold value of -0.4998.

TP: 270	FN: 50
FP: 16	TN: 184

Table 2. Shows the confusion matrix for the optimal threshold value of the diabetes label. Computed using LDA which produced a threshold value of 3.7875.

TP: 1	FN: 87
FP: 2	TN: 430

DISCUSSION

From our implementation of the hypothesis, we first needed to reduce the data down to two dimensions for each of the two labels of interest. After performing PCA on the diabetes label, our results are shown in Figure 1. By looking at the graph, we can visually see that there is a separation of the labels for diabetes. After performing PCA on the obesity label, out results are shown in figure 2. By looking at this graph, we can see that there is also a separation of the labels for obesity. When we compare Figure 1 and Figure 2, it is clear to see that the labels in the obesity data are more scattered throughout the plot relative to the labels in the diabetes data. From this information, we can tentatively predict that the separating hyperplane will perform better on the diabetes data, however we need to find a concrete method to prove this claim.

Our second form of analysis was to compute the Fisher's linear discriminant to perform linear discriminant analysis. The largest eigenvector of the ratio of Rayleigh quotients from the scatter matrices will allow us to find the separating hyperplane by projecting the data onto this axis to find the score values. These score values were then be plotted as shown in figures 3 and 4 grouping the data by label. For the diabetes data in figure 3, we can see that the labels are separated fairy accurately since there is little overlap between the labels. In Figure 4, the labels are also separated, but it is clearly visible to see that there is more overlap between the classifications, indicating that LDA performed well for diabetes and poorly for obesity.

Our next task was to formally implement a method to assess the efficacy of LDA. To do this we produced two ROC curves as shown in figures 5 and 6. We know that a perfect classifier would have curves that are closer to the top-left of the graph. In Figure 5 we can see that the diabetes classifier is close to this case, indicating that it was able to accurately predict the labels from the separating hyperplane. The diagonal of the ROC space indicates a random classifier. In Figure 6 we can see that the obesity classifier is closer to having random predictions than having correct predictions. The area under the curve summarizes the performance for the diabetes and obesity classifiers, and are found to be 0.9309 and 0.6071, respectively. Comparing these values gives us substantial evidence that the diabetes classifier performs more effectively relative to the obesity classifier.

Our methods also included a confusion matrix for an optimal threshold value that is calculated from the highest accuracy. From these matrices, we can extrapolate the accuracy that was found for the threshold value. For the diabetes data, the accuracy is 87.3%, and for the obesity data, the accuracy is found to be 82.9%. These accuracies show that for an optimal threshold value, the separating hyperplane does an adequate job at separating the data for both the diabetes and obesity labels.

Having different methods to assess the linear discriminant analysis is important because the ROC curve shows that the obesity classifier is relatively poor at classifying the data while the confusion matrix shows the obesity classifier was moderately successful at labelling the data. In such cases we need to decide if we are looking to test a classifier over a range of threshold values or if we just need to test the classifier on a fixed threshold value. The area under the curve does not necessarily show that the LDA is accurate or not, but rather the overall performance of the classifier over a continuous range of threshold values. By keeping track of the accuracy for each threshold, we were able to find the value which creates a classifier that is accurate. The choice to use the area under the curve or the confusion matrix is thus dependent on if we need to consider the sensitivity and specificity of the classification. For example, if we have a desire to have a high sensitivity with neutral consideration for the specificity then we can use the optimal threshold value for the diabetes data and create a single confusion matrix, since the sensitivity for this confusion matrix is 0.99. If we need to take into account the sensitivity and specificity, then the AUC would be a better measure of the classifier.

LDA is remarkably useful in many cases to help with supervised learning and provides a way to label data based on a separation of binary classifications. As we have seen in this assignment, we can use LDA in conjunction with PCA to classify diagnoses using labelled data that is already available to us. In real world applications, we can collect a large amount of classified data from an experiment which we can then use to create models that can further predict new classifications on incoming data.

In summary, our application of linear discriminant analysis and evaluation shows that classifiers are able to find a linear separation of the data to varying degrees of accuracy, depending on how we choose to interpret the data. The area under the ROC curve indicates that the diabetes classifier is overall more accurate over a range of threshold values while the obesity classifier is accurate for high sensitivity. To conclude, our testable hypothesis was capable of producing our desired objectives.

REFERENCES

Ellis, R. Class 23 Linear Discriminant Analysis – LDA. Retrieved April 1, 2021, from https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class23.pdf

Ellis, R. Class 24 Classification – Assessment by Confusion Matrix. Retrieved April 1, 2021, from https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class24.pdf

Ellis, R. Class 25 Classification – Assessment by ROC Curve. Retrieved April 1, 2021, https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class25.pdf