LDA and Classifier Assessment to explore health data Classification

ABSTRACT

PURPOSE: To determine how well 15 provided health variables can be used to classify the diabetes and obesity status for 520 patients using LDA classification.

METHODS: Use of LDA dimensionality reduction and classification for the dataset using the Diabetes and Obesity label vectors, and analyzing the results produced using classification evaluation metrics like ROC curves and confusion matrices.

RESULTS: The results clearly demonstrate the effectiveness of LDA as a classification algorithm for the 15 health variables using Diabetes and Obesity features as classification label vectors shown in Figures 2, 3, 4 and 5, through the evaluatory metric values shown in Table 1 and Figure 1.

CONCLUSIONS: The use of LDA for dimensionality reduction and classification for an exploration into the merit of Diabetes and Obesity as features of classification revealed that the better classifying feature was Diabetes. The meta-analysis of results using confusion matrices and ROC curves supported the conclusion to a high accuracy.

Word Count: 150

INTRODUCTION

The purpose of this assignment is to determine how well 15 provided health variables can be used to classify the diabetes and obesity status for 520 patients using LDA classification.

LDA is a dimensionality reduction and classification algorithm which works on the principle of maximising between cluster separation and minimizing within cluster separation while modelling data using label vectors, making it an effective supervised algorithm to explore datasets where two features and their merit as classification bases needs to be explored. The classifier assessment is done using ROC curves which graphically evaluates the overall performance of binary clusterings produced by LDA. The assessment is supported by the confusion matrices for the features being analysed for classification as they summarize the performance of the classifications in terms of true / false positives and negative which can further be used to evaluate the accuracy, precision and F scores of a classification.

The question being explored is the effectiveness of a LDA classification algorithm applied to the exogenous data using Diabetes and Obesity label vectors to provide the basis of classification, its accuracy at modelling the dmrisk.csv dataset.

METHODS

This exploration makes use of two prime techniques to investigate how the features 'Obesity' and 'Class', which is the status of Diabetes diagnosis of a patient, model the classifications of the rest of the 15 other features.

The first part of the algorithm, part (a), focuses on computing the Linear Discriminant Analysis (LDA) scores of produced by the two features in question (Diabetes and Obesity) to using a PCA reduction to 2D.

The second part of the algorithm, part (b), is a meta-analysis of the LDA axes produced by both features that define the classifications present in the data. It does so by producing a Receiver Operating Characteristic (ROC) curve using values produced at each iteration of the values in the evolving confusion matrices for both features. We then compute the accuracy of the LDA axes for both the features of focus and compute a final confusion matrix using the optimal threshold for each of them.

The parent function a4_20292366 is the executory function where all the data is first initialized for exploration, and then relevant plots and output to the console is produced by calling other functions in the program to process the data.

Part (a) is relevant to functions a4q1 and lda2class wherein the LDA axes and their respective label classifications for the Diabetes and Obesity features are calculated.

This is done primarily in the a4q1 function which first initializes a return vector that will be returned upon being called in the parent function. It first creates a classifier for the first feature being processed by applying LDA on labels extracted in the yvec previously in the parent function, and the exogenous data present in the Xmat for the corresponding label. In essence, labels of 1 in the vectors are attributed to the 'positive' class whereas labels of negative 1 are attributed to the 'negative' class. These LDA classification vectors for the features are then 'projected' into a lower dimensional space to increase the accuracy of interpretation and general ease of exploration.

The lda2class function being called inside a4q1 is where a accurate classification algorithm is implemented where the principle of maximizing between-class scatter and minimizing within-class scatter is applied. A simple spectral decomposition method using the built-in 'eig' function to create a binary separation, while applying the principle of Fisher's linear discriminant, which is the largest vector of the Rayleigh's quotient, is implemented to create a vector that produces the proper classifications for the LDA axes.

Part (b) makes use of the roccurve, confmat and aucofroc functions, each of which are involved in the meta-analysis of the results produced by part (a) of the algorithm which are concerned with exploring the data through supervised binary separation using labels.

The roccurve function is the prime function of interest as this is where the accuracy of each classification produced by either Diabetes or Obesity is calculated in two ways, the first of which has to do with the computation of the area under curve (AUC) value, which aims to be maximal to show an effective LDA classification and separation. The second measure of accuracy has to do with the values calculated in the confusion matrices at optimal threshold for each prime feature being used to explore the data.

The function first initializes vectors to be processed by sorting the scores into a unique subset from the LDA and permuting the scores to their respective labels. It then iterates through the size of the unique vector initialized previously and at each iteration, calculates the confusion matrix and the optimal threshold to initialize the qvec against in the confmat function. It stores the accuracy and the optimal threshold values in variables pre-initialized to the smallest value possible in MATLAB, -inf, for increased accuracy. It then calculates the AUC values for each feature using the aucofroc function part of the pre-given code, which basically calculates the integral of the ROC of each feature.

The confmat function is used to calculate the true negative, true positive, false negative and false positive values in the confusion matrices using a threshold calculated at each iteration of unique vector in the roccurve function. It does so using the qvec calculated previously and the label vectors for Diabetes and Obesity.

RESULTS

Table 1 contains the summary of results using the Linear Discriminant Analysis and Classifier assessment on the dmrisk.csv dataset and calculating the confusion matrices for the optimal threshold for Diabetes and Obesity features' LDA axes.

Figure 1 shows the ROC curve for the label vectors for Diabetes and Obesity at their optimal thresholds using the values in the confusion matrices in Table 1.

Figure 2 shows the PCA scatterplot for the reduced data in the Diabetes data vector.

Figure 3 shows the PCA scatterplot for the reduced data in the Obesity data vector.

Figure 4 shows the LDA axes for the classifier scores using the Diabetes label vector.

Figure 5 shows the LDA axes for the classifier scores using the Obesity label vector.

Table 1: Summary of results using the Linear Discriminant Analysis and Classifier assessment on the dmrisk.csv dataset and calculating the confusion matrices for the optimal threshold for Diabetes and Obesity features' LDA axes.

Diabetes:

AUC ≈ 0.93

		+1	**-1**
Label	**+1**	270	50
	-1	16	184

Obesity:

AUC ≈ 0.61

		+1	**-1**
Label	**+1**	1	87
	-1	2	430

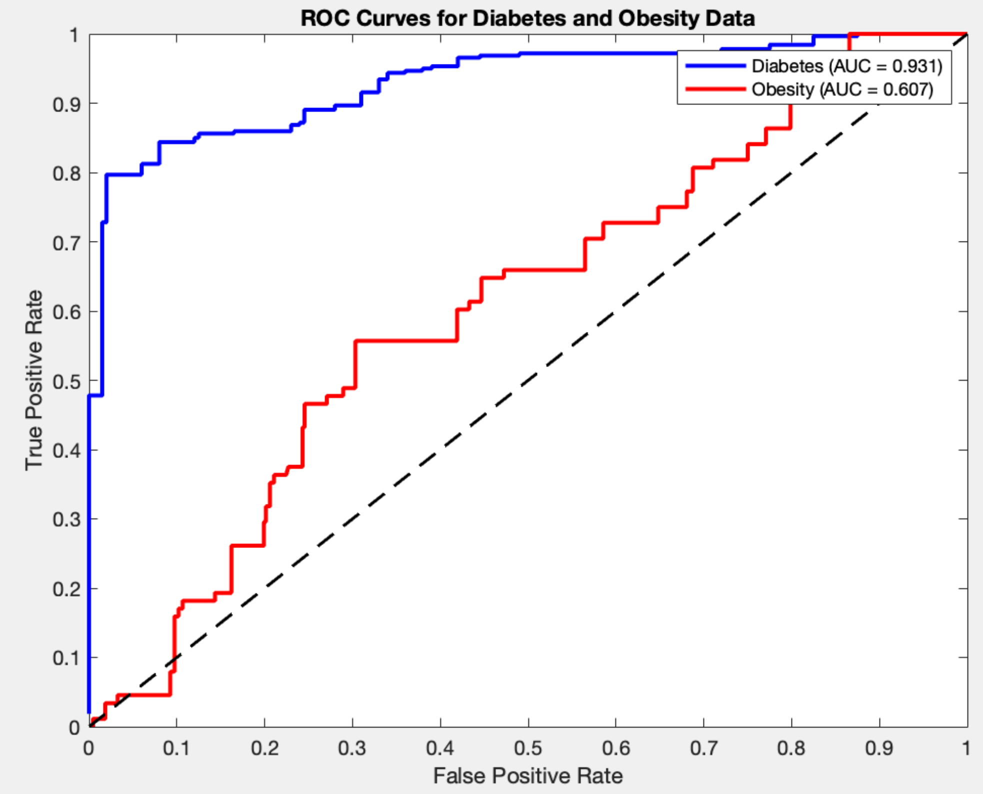

Figure 1: ROC curve for the label vectors for Diabetes and Obesity at their optimal thresholds using the values in the confusion matrices in Table 1.

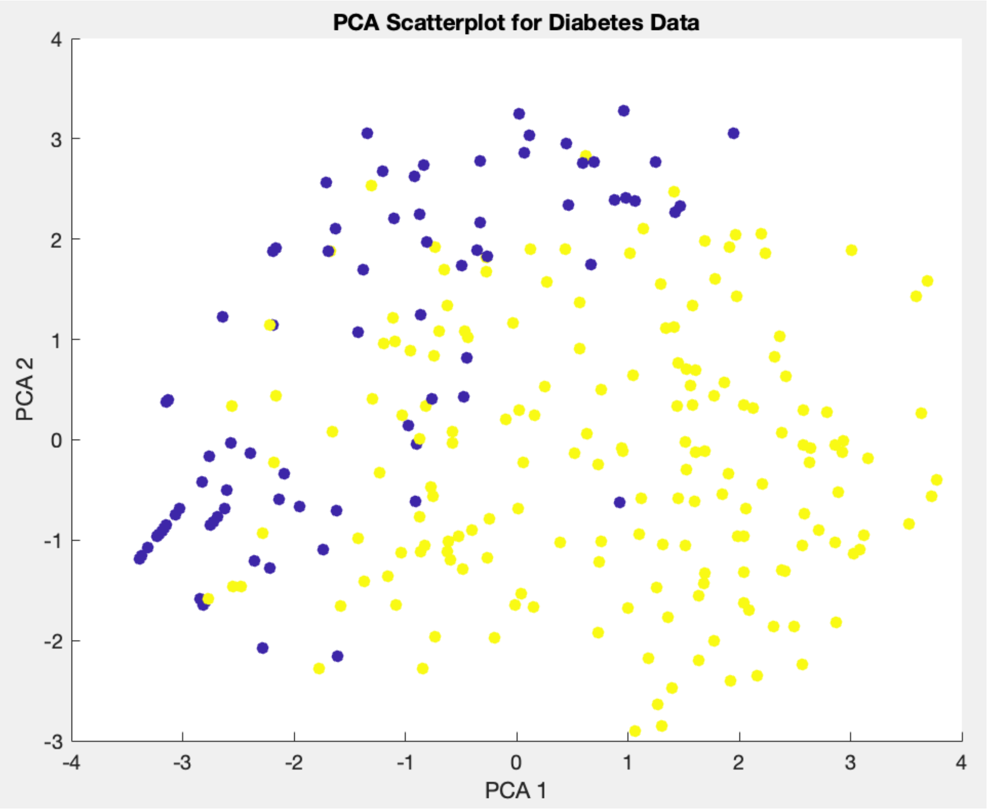

Figure 2: PCA scatterplot for the reduced data in the Diabetes data vector.

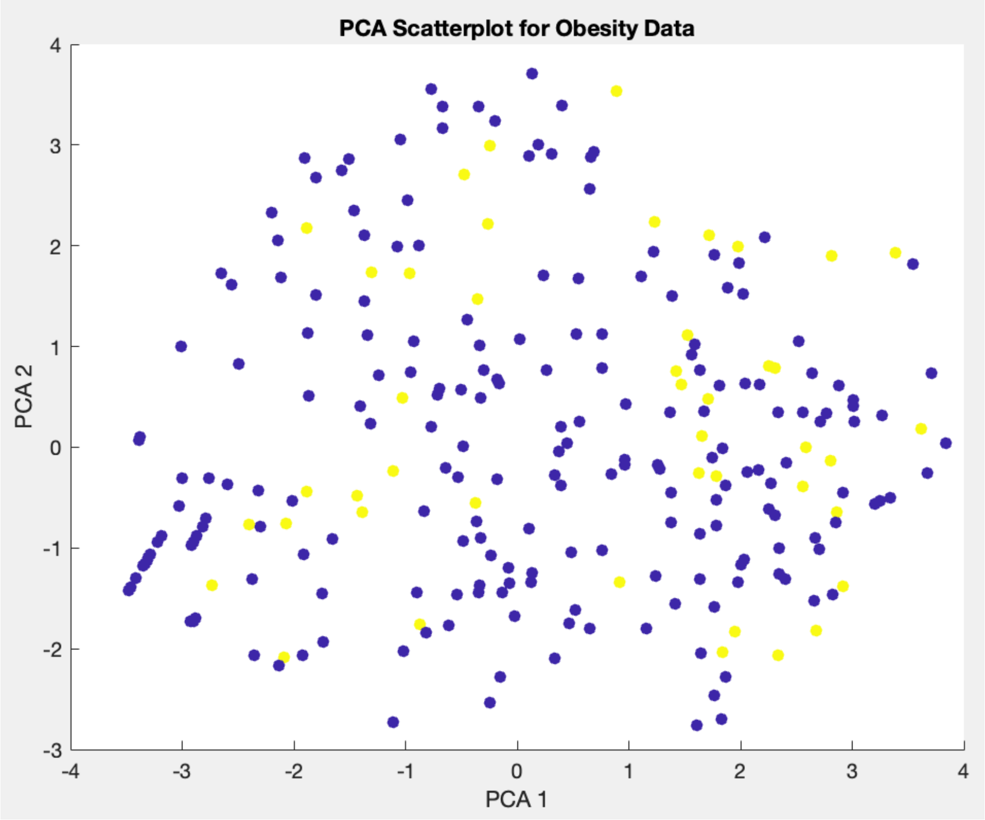

Figure 3: PCA scatterplot for the reduced data in the Obesity data vector.

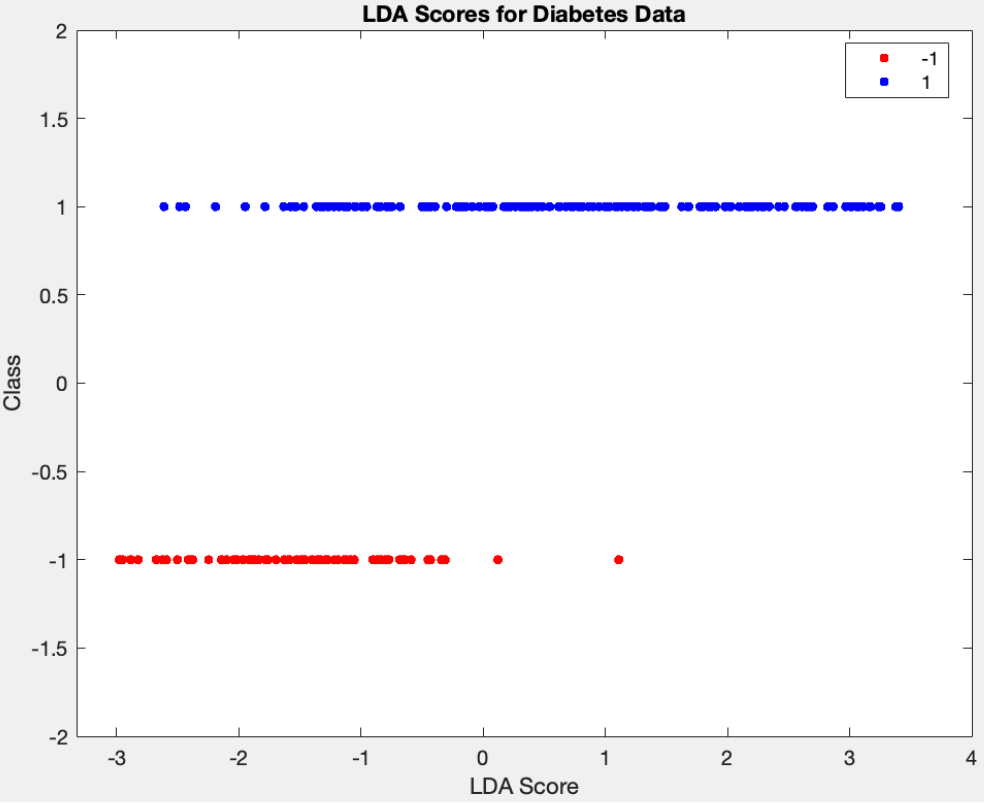

Figure 4: LDA axes for the classifier scores using the Diabetes label vector.

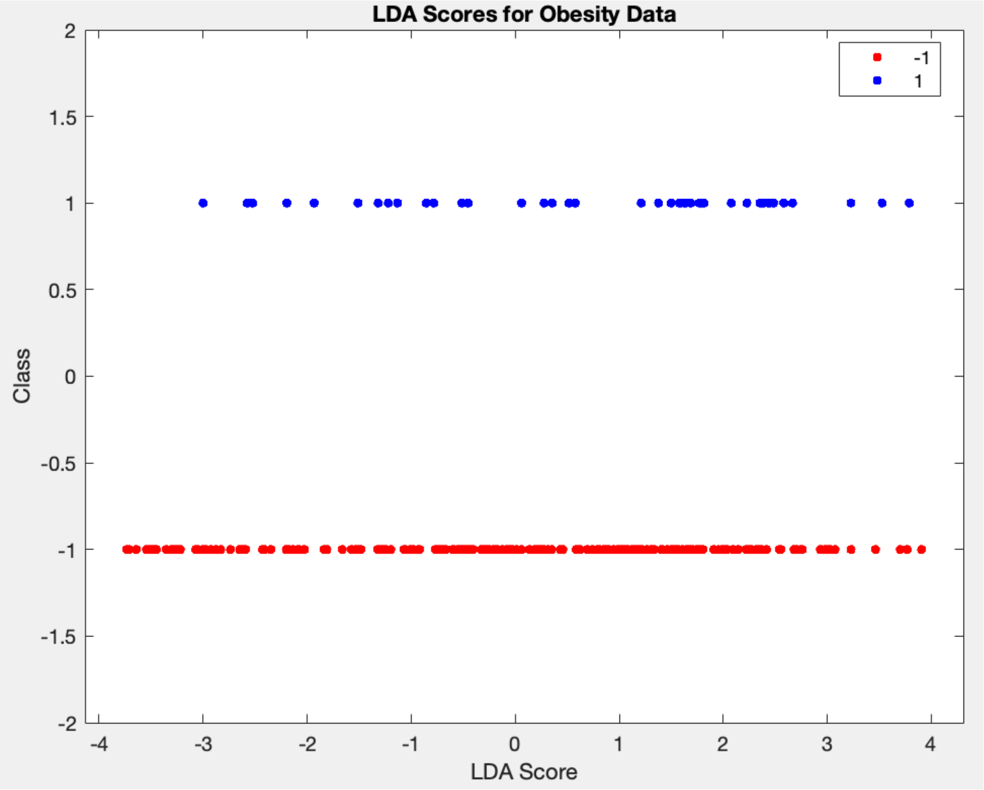

Figure 5: LDA axes for the classifier scores using the Obesity label vector.

DISCUSSION

Figures 2, 3, 4 and 5 are the output relevant to the first part of our exploration which investigates the effectiveness of LDA to separate the data present in dmrisk.csv with the labelled vectors for Diabetes and Obesity. The pre-processing of the data done by PCA dimensionality reduction makes it optimal for effective interpretation since interpretation done in 2D is exponentially more convenient than in 15D, the graphical output for which can be seen in Figures 2 and 3 for Diabetes and Obesity respectively.

The PCA reduced plot for Diabetes in Figure 2 shows two clusters that, when compared to the PCA reduced plot for Obesity in Figure 3, are relatively well separated. This unsupervised clustering indicates the fact that the Diabetes label vector is likely going to create an LDA axis which shows a better binary clustering of ‘positive’ and ‘negative’ labels which would indicate a better ability to define the classification attributes present in the overall dataset dmrisk.csv.

Doing this unsupervised algorithm before subjecting the data to a supervised classification-based algorithm helps us find any natural patterns in the data which is invaluable for exploration especially in higher dimensional data reduced to lower dimensions. The relatively good clustering of data in Figure 2 for Diabetes is supported by the LDA axes for the LDA scores calculated for Diabetes in Figure 4 which, relative to the overlap present between the LDA axes in Figure 5 for Obesity, is well separated. Since LDA works on the principle of maximizing between scatter distribution whilst minimizing within-scatter distribution, we can intuitively say that the labelled vector for Diabetes status is a better basis for classification for the data present in dmrisk.csv.

This mutual support between the results of the supervised LDA and the unsupervised PCA inherently back the claim that Diabetes is a better modelling feature relative to Obesity, but to mathematically evaluate this claim, we implement the part (b) meta-analysis of our results produced in part (a).

We do so in two distinct but interconnected ways both of which have been described in the Methods section. The first is using the ROC curves for both the features to explore the trade-off between the true positive rates (TPR) and the false positive rates (FPR) for the classifications produced by the Diabetes and Obesity features and their overall performance as characterized by the AUC values for them.

Since TPR is the proportion of actual positive cases that are correctly identified as positive by the classifier, whereas FPR is the proportion of actual negative cases that are incorrectly classified as positive by the classifier, a good classifying feature has a relatively higher TPR and a low FPR. The TPR for Diabetes and Obesity is easily calculated using their respective confusion matrices computed using the algorithm in Table 1. The TPR and FPR for Diabetes are approximately 84.4% and 8% respectively, whereas the TPR and FPR for Obesity are 1.1% and 0.46%, which mathematically show the fact that Diabetes is a better classifying feature.

The proportion that classification algorithms aim to maximize is that of TPR/FPR which for Diabetes is 10.55 compared to that of Obesity, which is 2.39, supports the fact illustrated in the ROC curves for both the features shown in Figure 1. Another intuitive measure of accuracy can be gleaned from the confusion matrices for the Diabetes and Obesity in Table 1, since the feature with a higher raw value of true positives and true negatives would be the better basis for classification. This is true since they account for roughly 87.3% of all outcome values in Diabetes but relatively less for Obesity at 82.9% of all outcome values, otherwise known as accuracy. This in isolation though, is not an effective measure of accuracy.

The ROC curves in 1 Figure 1 are clearly above the 0.5 AUC benchmark which is better than random, which indicates that both features are valid selections for a classification algorithm. Consequently, the fact that the AUC measure, which is a metric to measure the overall classification performance of the binary separations created by the label vectors of the two features, was higher for Diabetes further supports Diabetes to be the better classifying vector and have a higher ‘discriminative power’ in the overall dataset.

Even though this exploration produces mathematically sound conclusions, it is not without its limitations since LDA classification is limited to binary classification as it can only be used for explorations into a binary clustering. Though it is true that it can be extended to take on multiple features into a single clustering, this does not produce meaningful results since the results do not speak to the effectiveness of a single feature, but rather a random combination of them, which depending on the size of the dataset might be huge and turn out to be computationally expensive and interpretationally futile.

LDA may also be suboptimal for use in datasets that assume a normal distribution or non-standardized raw data, which would likely cause the algorithm to overfit the data due to the unpredictable amount of variance and noise in the dataset. This would likely lead to poor generalization and performance on new datasets which make the whole point of using a supervised classification algorithm redundant. Though this is not the case for our dataset, our algorithm may not be guaranteed to work on datasets that were initially non-standardized with high variance.

In conclusion, the use of LDA for dimensionality reduction and classification for an exploration into the merit of Diabetes and Obesity as features of classification revealed that the better classifying feature was Diabetes. The meta-analysis of results using confusion matrices and ROC curves supported the conclusion to a high accuracy.