Effectiveness of LDA for Diabetes and Obesity Classification

Student Number: 10219553

0.1 Abstract:

PURPOSE: Investigate & evaluate the effectiveness of Linear Discriminant Nalaysis (LDA) to partition data pertaining to diabetes and obesity through the use of the ROC curve, and the area under this curve (AUC).

METHODS: The data was first reduced to 2D through Principal Component Analysis (PCA), then the LDA axes and scores were calculated alongside the confusion matrices for a subset of thresholds. These values were parsed for the most accurate result, for which ROC curves and the AUC were found.

RESULTS: Through analysis of the PCA, LDA, and AUC, the method was highly effective for the classifying diabeties, but only moderately effective for obesity. LDA was found to be significantly more effective than PCA on its own.

CONCLUSIONS: The algorithm implemented over the given dataset met the metrics evaluation criteria specified in the scientific question, with both AUC values surpassing 0.6. This method is effective for this use-case.

Word Count: 145

0.2 Introduction:

The objective of this study was to investigate the effectiveness of Linear Discriminant Analysis (LDA) to separate and classify data pertaining to diabetes and obesity.

Principal Component Analysis (PCA) is a method of reducing the dimensionality of a dataset. To perform PCA, the first two right singular vectors from the V matrix are taken from the Singular Value Decomposition (SVD), and multiplied with the Zero-mean data matrix. In our case, this will reduce the dimensionality of the dataset to 2 Dimensions.

Linear Discriminant Analysis (LDA) identifies a linear combination of features that best separates multiple classes or clusters. This combination can either be used as a direct classifier, or as a method of dimensionality reduction, which is useful for further classification methods. The LDA process requires the mean vectors for each class in the dataset, and it calculates two scatter matrices. The first is the within-class scatter matrix, which measures how much the classes are spread out from their respective means. The between-class scatter matrix shows the separation between the class means from the overall mean. The LDA Axis signifies the vector direction that maximizes the separability of the classes. It is a critical component of enhancing classification performance [1].

Binary-label confusion matrices visualize and evaluate the accuracy of classification algorithms. They calculate and organize the outcomes of a classifier into a matrix that compares the predicted class labels against the actual labels. This provides a snapshot of the model's performance. It does this by calculating the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [2]. These values are organized as follows:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Receiver Operating Characteristic (ROC) Curves are graphical plots that illustrate the ability of a binary classification system as its discrimination threshold changes. The curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) [3].

Area Under the Curve (AUC) evaluations measure the 2D area under the ROC curve, which exists between (0,0) and (1,1). It provides a single scalar value that summarizes the performance of a binary classification value, ranging from 0 to 1. An AUC of 0.5 implies that the classification is just as effective as picking a class at random every time, while higher values indicate better predictions.

The scientific question to be looked at is, given a dataset pertaining to diabetes and obesity, can LDA be applied to the dataset such that the AUC values are above 0.6, indicating performance above that of random chance.

0.3 Methods:

The analytical framework of this study is structured into four distinct sections. This approach ensures a comprehensive understanding of the data characteristics and underlying patterns.

The preliminary requirements are to access the dataset and perform PCA using built-in functions for simplicity. These results are then plotted separately for Diabetes and Obesity.

The study commences with the calculation of the LDA axes and scores for the dataset labels, previously mentioned to be Diabetes and Obesity. The first step in this process is to zero-mean the matrices, denoted by Xmat1 for diabetes and Xmat2 for obesity. Subsequently, the within and between-label scatter matrices (SW and SB, respectively) are calculated, allowing us to evaluate the Rayleigh Quotient. SW is found by adding the scatter matrices of the two labels, while SB is the scatter of zero-mean means. The Rayleigh Quotient Matrix is calculated by taking the product of the inverse of SW with SB. Finally, the largest eigenvector in this new matrix is taken as the LDA axis for the dataset.

To determine the LDA scores of the dataset using these new values, the product of each Xmat is taken with the corresponding qvec.

The final part of the analysis is to calculate the ROC curve and the AUC evaluation. A subset of unique scores is found and used as threshold values. For each of these values, a confusion matrix is calculated. The True/False Positive rates are calculated, and the optimal threshold is found to be the confusion matrix with the highest accuracy. The plots for the ROC curves of the two labels have the TPR and FPR as axes. Finally, the Area under each curve is calculated, showcasing the accuracy of the predictions and the effectiveness of LDA to classify data.

0.4 Results:

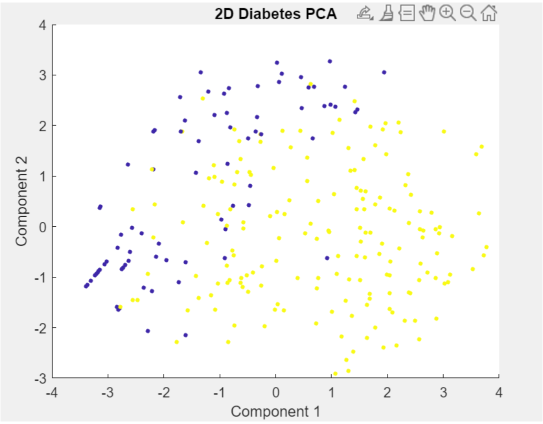

Figure 1: The scatter plot demonstrates the 2D PCA of the Diabetes Label. The data is fairly well clustered into two classes.

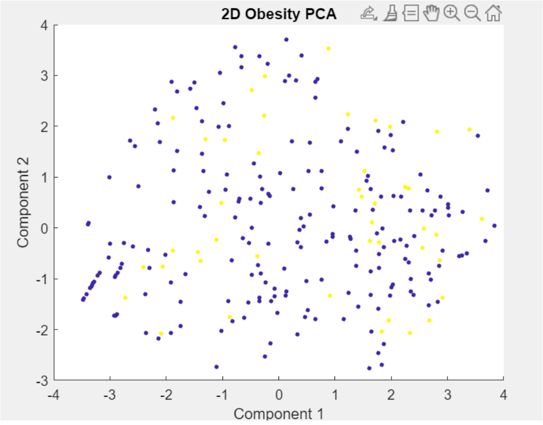

Figure 2: The scatter plot demonstrates the 2D PCA of the Obesity Label. The data is poorly clustered into two classes, with lots of overlap.

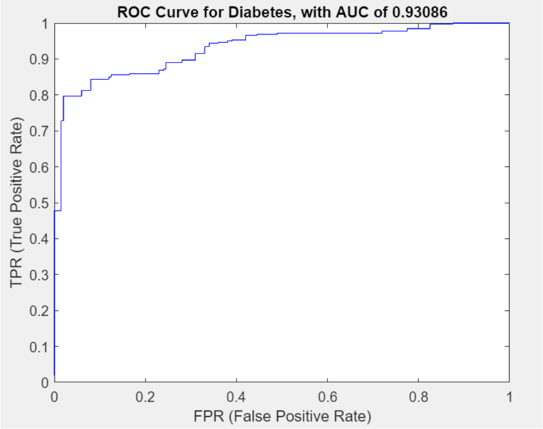

Figure 3: Graph showing the ROC curve for Diabetes.

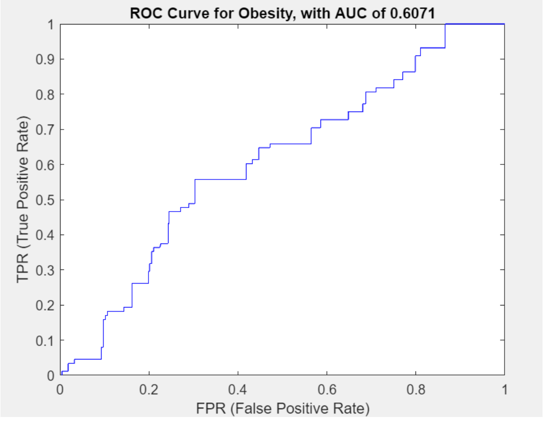

Figure 4: Graph showing the ROC curve for Obesity.

Table 1: The AUC value, alongside the confusion matrix with the optimal threshold, calculated through LDA, for both Diabetes and Obesity.

	Diabetes (AUC = 0.9309)			Obesity (AUC = 0.6071)	
	+1	-1		+1	-1
+1	270	50	+1	1	87
-1	16	184	-1	2	430

0.5 Discussion:

The findings of this experiment are mostly contained within the figures and table in the results section. Figures 1 and 2 visualize the effects of PCA on the data labels, Diabetes and Obesity, while Figures 3 and 4 plot the ROC curves, respectively. Table 1 shows the relevant confusion matrices and AUC values for the ROC curves in question. The optimal threshold for the diabetes label is -0.4998, with an AUC of 0.9309. The optimal threshold for the obesity label is 3.7875, with an AUC of 0.6071.

A glance at the first two figures gives us a general idea of how the LDA might progress. The PCA for the Diabetes label shows some separation between the two clusters. While there is significant overlap, there are few significant outliers and it is clear how the LDA may help separate the classes further. However, the PCA graph for the Obesity label shows almost zero significant partitioning, indicating that there may be some difficulties in classifying the data, even after using LDA.

These results fall in line with the ROC curves and AUC values. An AUC value above 0.5, approaching 1, indicates that the classification is better than randomization. For the diabetes data, the AUC value was a very healthy 0.9309, which is unsurprising considering the PCA graph. Similarly, the results for the obesity data are as expected. The AUC was barely beyond the evaluation criteria of 0.6, sitting at about 0.6071, indicating poor overall performance. The ROC curve for obesity almost follows a straight line with $y = x$, which would be an auc value of 0.5. Overall, this means that the classifier for the obesity data is far less accurate than for the diabetes data, making it difficult to predict if someone is obese.

Further analysis of the confusion matrices show similar patterns. The obesity confusion matrix has very few true positive and false positives, with some false negatives but mostly true negatives. This indicates that the performance of the model is strong for negative predictions. The AUC should theoretically be higher, but this is clearly not the case. Figure 4, the ROC Curve, gives clues as to why this is happening. An ROC curve plots the True positive rate against the false positive rate. The TPR is low in the graph, while the FPR is high, indicating that the classification process has high type 1 error, and low sensitivity. (A type 1 error, also known as a "false positive," occurs when a statistical test incorrectly rejects a true null hypothesis. This means that the test indicates a significant effect or relationship when in reality there is none, erroneously inferring that an observed result is not due to chance.)

The confusion matrix of the optimal threshold for the diabetes matrix is much more even. There are numerous true positive and negative values, with very few type 1 and 2 errors, indicating that the classification is strong. This is in line with the ROC curve in figure 3 and the AUC value above 0.9. The TPR is accurate and the FPR is insignificant, implying that the model has high sensitivity.

Future work could involve exploring the quality and reliability of the data. It is mentioned in the project description that the dataset is relatively new and has not been explored substantially. The potential for bias within the data also merits attention, as it could significantly skew the classification models' precision. Biases in data collection or labeling, or a lack of representativeness in the training dataset, could undermine the model's validity. Addressing data bias, possibly through more heterogeneous sampling, represents a crucial direction for future research.

Exploring alternative classification techniques such as neural networks might yield better results than LDA. Future initiatives could include gathering additional data to enhance classification accuracy. Neural networks are increasingly used in medicine for tasks such as predicting diabetes, leveraging their ability to detect complex patterns in large datasets, including electronic health records. They are also applied in image analysis for diagnosing conditions from X-rays and MRI scans, demonstrating versatility in various medical diagnostic procedures [5].

The consequences of LDA's less-than-perfect data separation warrant consideration. In the realm of diabetes, where the classifier demonstrates proficiency, the ramifications are profound for medical diagnostics and treatment strategies. Precise predictions could trigger timely preventative and interventional measures, potentially elevating patient prognoses and curbing medical expenses. Conversely, LDA's suboptimal separation in obesity classification might influence public health strategies and efforts to mitigate obesity prevalence.

In conclusion, the Linear Discriminant Analysis met the evaluation criteria proposed in the scientific question, where the AUC values of both labels were above 0.6, indicating that the classifier had better performance than randomization. Further exploration into the dataset and possible bias, as well as applications of other methods such as neural networks may prove more fruitful.

0.6 References:

1. Patterns – Linear Discriminant Analysis, or LDA. Available from: `https://research.cs.queensu.ca/home/cisc271/pdf/Class24.pdf`

2. Classification – Assessment With Confusion Matrix. Available from: `https://research.cs.queensu.ca/home/cisc271/pdf/Class25.pdf`

3. Classification – Assessment With ROC Curve. Available from: `https://research.cs.queensu.ca/home/cisc271/pdf/Class26.pdf`

4. Using a Convolutional Neural Network to Predict Remission of Diabetes After Gastric Bypass Surgery. Available from: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8414302/`

5. Types of Biases in Data. Available from: `https://towardsdatascience.com/types-of-biases-in-data-cafc4f2634fb`