Evaluating Performance of LDA Using ROC Curves

Matthew Sun

1 Abstract

The purpose of this assignment is to use perform LDA on the two-dimensional reduction of two variations of the given data and evaluate the effectiveness of the binary classifications using the AUC of their ROC curves. The first variation uses diabetes data as the binary classifier and the second variation uses obesity as the classifier. It was found that using diabetes data as a classifier produced an AUC of 0.9309 and using obesity data as a classifier produced an AUC of 0.6071. The higher AUC value of the diabetes label indicated a better-performing binary classification and visual observation of its PCA dimensionality reduction and LDA scores agreed with those findings.

2 Introduction

The objective of the assignment was to perform linear discriminant analysis (LDA) on a set of data containing 15 health variables which were used to classify the diabetes and obesity statuses of 520 patients. Additionally, this assignment aimed to determine which variable (diabetes or obesity) when used as the label vector, produced a better binary classification using LDA by evaluating their performance with the respective receiver operator characteristic (ROC) curve.

Dimensionality reduction can be performed on a given data set using principle component analysis (PCA). This process will produce the first n loading vectors for the zero-mean data which when scored with original data will provide an optimal n-dimensional reduction since those loading vectors are associated with the largest variance in the data. The motivation for LDA is that PCA is an unsupervised algorithm that does not consider any labels provided by the data. LDA on the other hand is supervised and can produce an optimal classification of the labelled data in the form of a hyperplane(s) of best separation. This hyperplane can be determined from Fisher's linear discriminant, also known as the LDA axis. Suppose there exists a data set where each observation has a binary label $y_i \in \{1, 2\}$.

Assume that the corresponding data matrix is a "tall-thin" matrix $A \in \mathbb{R}^{m \times n}$, where $m > n$ and the zero-mean matrix is $M := A - \vec{1}^T \bar{A}$. If PCA were performed on A, it would be observed that the first loading vector would provide a relatively good description of the data and a relatively poor separation of the labels while the last loading vector will be relatively good at separating the labels. Both of these vectors would be axes that are aligned with the corresponding eigenvectors of the scatter matrix $S = M^T M$. LDA accounts for existing labels within data by partitioning the observations of A into A_1 and A_2 where A_1 contains all observations with $y_j = 1$ and A_2 contains all observations with $y_j = 2$ where the goal is to simultaneously maximize the distance between the mean of each partition while minimizing the scatter within each partition. This is done by finding Fisher's linear discriminant which can be written as the constrained optimization problem

$$\vec{w} = \underset{\vec{u} \in \mathbb{R} : \vec{u} \neq 0}{\arg\max} \frac{R(S_B, \vec{u})}{R(S_W, \vec{u})} \tag{1}$$

where the function $R(S, \vec{u})$ is the Rayleigh quotient

$$R(S, \vec{u}) := \begin{cases} \frac{\vec{u}^T S \vec{u}}{\vec{u}^T \vec{u}} & \text{if } \vec{u} \neq \vec{0} \\ 0 & \text{if } \vec{u} = \vec{0} \end{cases} \tag{2}$$

S_W is the within-label scatter matrix that is defined as $S_W := S_1 + S_2$. $S_1 := M_1^T M_1$ and $S_2 := M_2^T M_2$ are the scatter matrices of each partition with zero-mean matrices $M_1 = A_1 - \vec{1}^T \bar{A}_1$, $M_2 = A_2 - \vec{1}^T \bar{A}_2$. S_B is

the between-label scatter matrix and is defined as

$$S_B = \begin{bmatrix} \bar{A}_1 - \bar{A} \\ \bar{A}_2 - \bar{A} \end{bmatrix}^T \begin{bmatrix} \bar{A}_1 - \bar{A} \\ \bar{A}_2 - \bar{A} \end{bmatrix} \tag{3}$$

The vector $\vec{w}$ provides a direction that maximizes the between-label scatter while minimizing the within-label scatter. If S_W is positive definite, then the equation for Fisher's linear discriminant can be simplified to $\vec{w} = \vec{v}_{MAX}(S_W^{-1} S_B)$. The LDA is score is found by multiplying the zero-mean data matrix with $\vec{w}$.

The ROC curve describes the binary classification abilities of a system as the threshold for separation increases. The curve is constructed by plotting the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis. The FPR and TPR for each point on the curve are determined by a confusion matrix. The confusion matrix is an instance of a contingency table where the top left corner is the number of true positives, the top right corner is the number of false positives, the bottom left corner is the number of false negatives, and the bottom right corner is the number of true negatives. Plotting the ROC curve requires the output scores from a classifier to compare to the labels. In a binary classification problem, there is the positive and negative class, denoted by +1 and -1 respectively. By changing the hyperparameter θ to different values along the given scores where all scores $< \theta$ are assigned to -1 and scores $\geq \theta$ are assigned to +1. The values in the confusion matrix can then be found where the true positives (TP) are occurrences where the score and class are +1, the false positives (FP) are occurrences where the score is +1 and the label is -1, the false negatives (FN) are occurrences where the score is -1 and the label is +1, and the true negatives (TN) are occurrences where the score and label are -1. FPR can be computed as $FPR = \frac{FP}{FP+TN}$ and TPR as $TPR = \frac{TP}{TP+FN}$. The confusion matrix that each θ produces can be evaluated by its accuracy $acc = TPR \cdot TNR - FPR \cdot FNR$. The area under the ROC curve (AUC) is an overall measure of the performance of the binary classification across all possible θ values. An AUC of 0.5 indicates that using the given binary classifier is no better than randomly guessing while an AUC of 1 indicates that the classification makes no faults.

The scientific question to be explored is: Which variable when used as the label vector, diabetes or obesity, does LDA provide a more effective binary classification and why? The effectiveness of each binary classification was determined through visual observation and by comparing the AUC of each ROC curve. A higher AUC value indicates a better-performing binary classification system.

3 Methods

To compare the performance of diabetes and obesity as labels, the data was first processed into two distinct subsets of the original data. The first data matrix is formed with all variables except the column containing diabetes and uses the entries of the diabetes column as the labels. The first data matrix was formed with all variables except the column containing obesity and uses the entries of the obesity column as the labels. Dimensionality reduction to two dimensions was performed on both data matrices using PCA. LDA was then performed on both PCA scores with the respective label vector. The LDA axis was found by first partitioning each of the data sets where A_1 contained the observations with the first label and A_2 contained observations with the second label. The zero-mean matrices M_1 and M_2 for each data matrix were then computed which was then used to find the within-label scatter matrix $S_W = M_1^T M_1 + M_2^T M_2$. The between-label scatter matrix $S_B = [A - \bar{A}]^T [A - \bar{A}]$, where A is a matrix containing A_1 as the first row and A_2 as the second row. The eigenvector corresponding to the largest eigenvector of $S_W^{-1} S_B$ is the LDA axis. The LDA score was found by multiplying the original zero-mean data matrix with the LDA axis.

The ROC curve for each data matrix and corresponding label vector was found by iterating through every unique LDA score and using each entry of the LDA score as the hyperparameter then finding the confusion matrix. The confusion matrix was calculated by counting the occurrences of true positives, false negatives, false positives, and true negatives. The score that acts as the best hyperparameter based on the accuracy measure described previously and the corresponding confusion matrix was stored. The AUC was

calculated using the Trapezoidal rule.

The AUC of each ROC curve was used to evaluate the effectiveness of using diabetes or obesity as binary classifiers. The scatter plots of the PCA dimensionality reduction and the visualization of the LDA scores were assessed by visual observation to complement the evaluation provided by the AUC.

4 Results

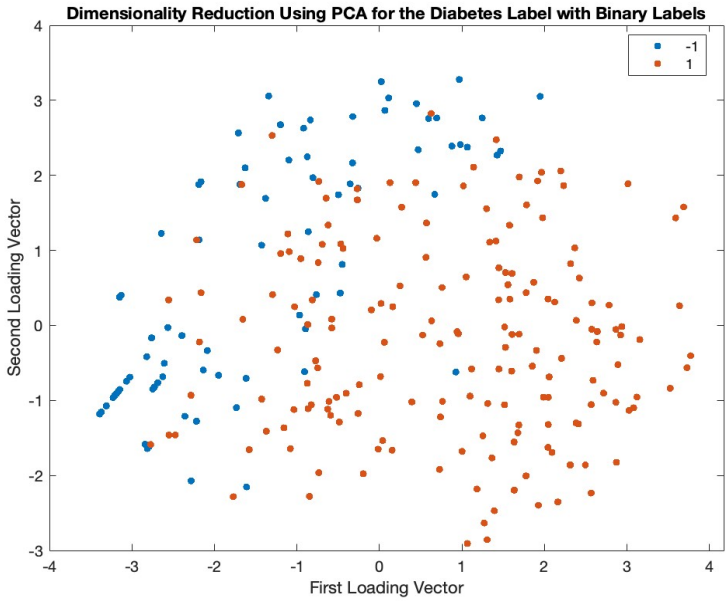

Figure 1: Scatter plot of the dimensionality reduction produced by PCA of the data excluding the diabetes variable while using its entries as labels. Positive group (1) in red, negative group (-1) in blue.

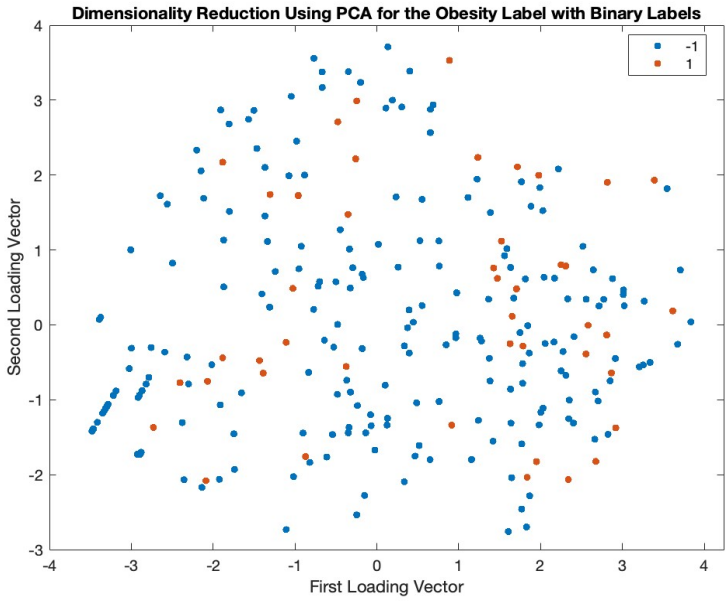

Figure 2: Scatter plot of the dimensionality reduction produced by PCA of the data excluding the obesity variable while using its entries as labels. Positive group (1) in red, negative group (-1) in blue.

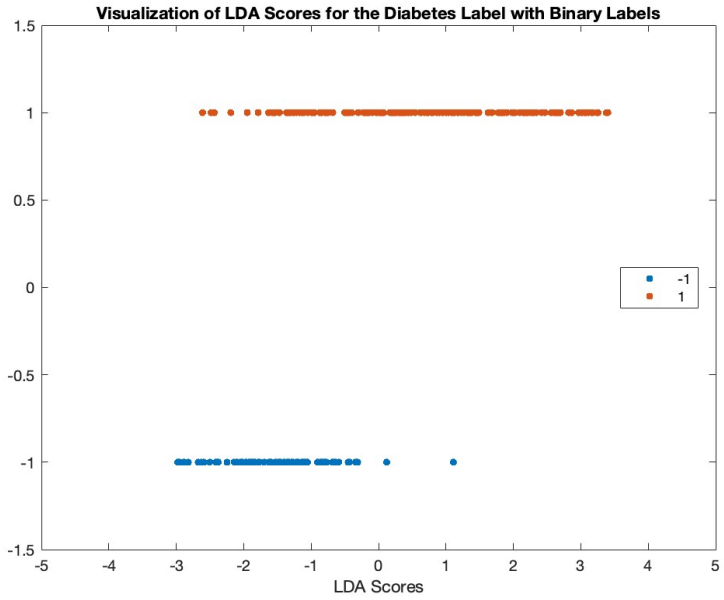

Figure 3: Visualization of the LDA scores for the diabetes label. Vertical separation exists for visualization purposes only. Positive group (1) in red, negative group (-1) in blue.

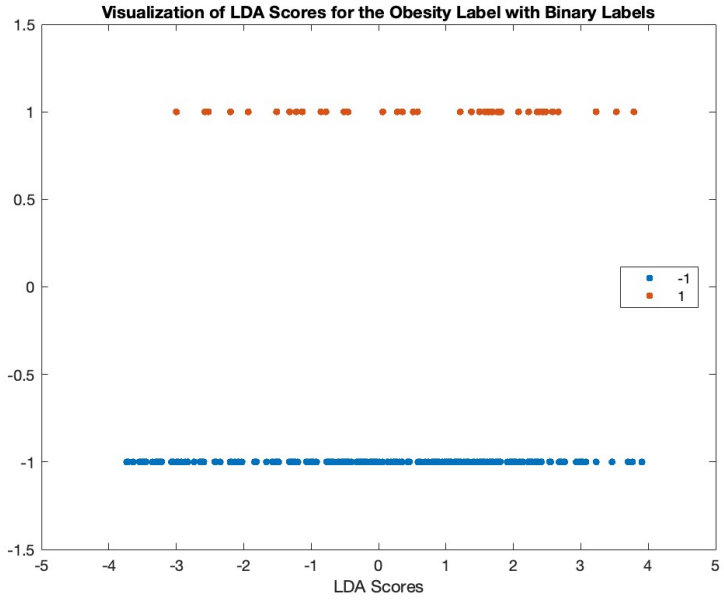

Figure 4: Visualization of the LDA scores for the obesity label. Vertical separation exists for visualization purposes only. Positive group (1) in red, negative group (-1) in blue.

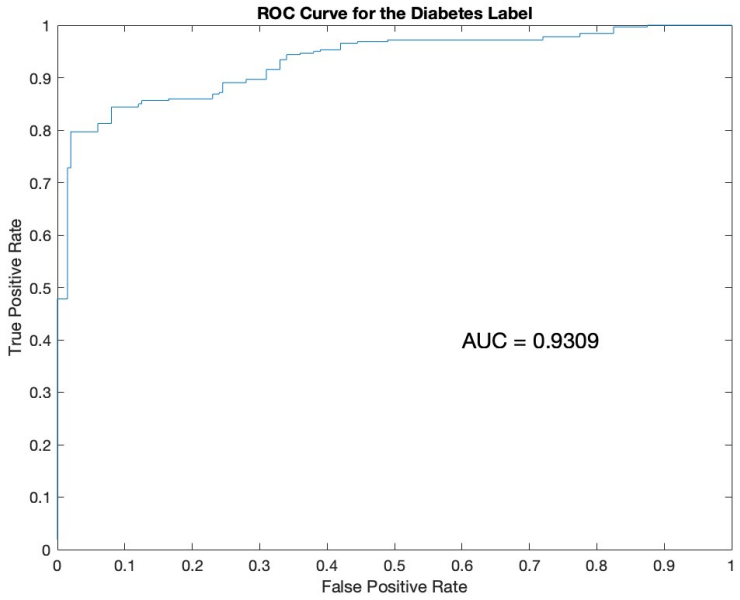

Figure 5: ROC curve plotted using the LDA scores of the data with the diabetes label and the AUC of this curve rounded to four significant digits.

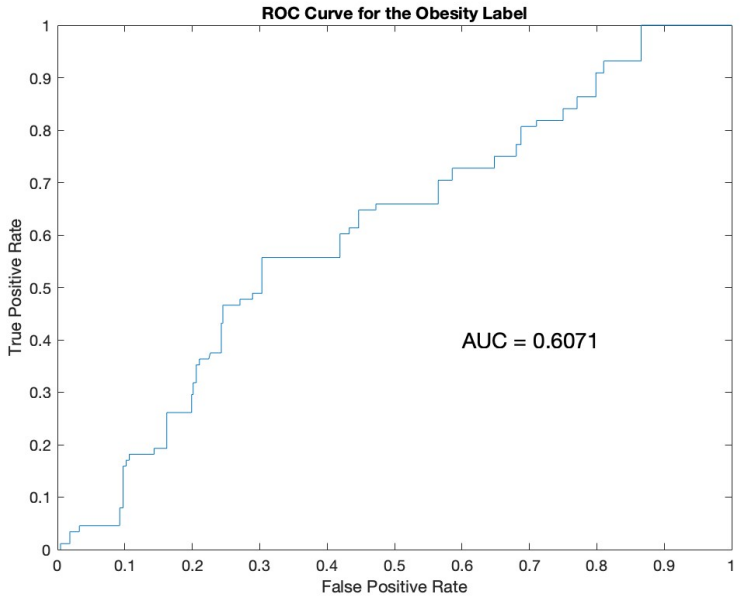

Figure 6: ROC curve plotted using the LDA scores of the data with the obesity label and the AUC of this curve rounded to four significant digits.

Table 1: The AUC and "optimal" confusion matrix, computed using the LDA scores, for the diabetes and obesity labels.

Table 1.1:
Diabetes, AUC ≈ 0.9309

Label \ Score	**+1**	**-1**
+1	255	65
-1	4	96

Table 1.2:
Obesity, AUC ≈ 0.6071

Label \ Score	**+1**	**-1**
+1	49	39
-1	131	301

5 Discussion

For the given data set containing samples of patients and health variables with either diabetes or obesity as the binary classifier, an evaluation of the performance of LDA on the PCA two-dimensional reduction found that diabetes is more suitable as a binary classifier for the given data set. It was found using diabetes as a classifier resulted in an AUC value of 0.9309, an optimal hyperparameter at the LDA score value of -0.2985. Using obesity as a classifier resulted in an AUC value of 0.6071, an optimal hyperparameter at the LDA score value of 1.2116. Visual observation of the scatter plot of the PCA dimensionality reduction and the visualization of the LDA scores would agree with the evaluation provided by the AUC values. Observe Figure 2 where the positive and negative classes are very dispersed while neither of them occupies any certain region. Contrast this with Figure 1 where although the two classes have many points within close proximity of each other, the points in the negative class are much closer together. From this observation, it should be expected that neither of these reductions would have a hyperplane that perfectly separates the two classes but it would be expected the reduction with the diabetes label (Figure 1) could be better classified. A similar observation can be made for the visualizations of the LDA scores. Observe Figure 4, the LDA scores between the two classes have much more "overlap" compared to the LDA scores in Figure 5. In this sense, the AUC is a general indicator of the effectiveness of a label as a binary classifier. Comparing the accuracy of the "optimal" hyperparameter provides further evidence that diabetes produces a better classification. The LDA scores with the diabetes label produced a peak accuracy of 0.7769 while the LDA scores with the obesity label produced a peak accuracy of 0.2536.

A limitation to this process is that the sample size provided in the given data set is relatively small considering the complexity of these health conditions. Furthermore, this process only evaluates diabetes and obesity as classifiers but the data set includes many other conditions and symptoms which could be used as classifiers. Doing so could provide a much more comprehensive evaluation of all variables and how they compare to each other as classifiers which could in turn reveal something about the relationship between these variables. The results could be further verified by repeating this process of finding the AUC for each classifier and then performing k-fold cross-validation for assessment. A 2020 study [Gur+20] executed a similar process as suggested on a data set containing results from a Surface-enhanced Raman scattering (SERS) spectroscopy, an analytical method used in chemistry often used to detect proteins in bodily fluids. In this study, researchers used what they called a "repeated double cross-validation" where they first optimized the number of loading vectors from PCA to use in each LDA model with multiple iterations of 7-fold cross-validation then multiple iterations of 3-fold cross-validation were performed to assess the LDA models. They then calculated the ROC curve for each of the LDA models. A similar process could be applied to this data set to provide more comprehensive results.

For the given data set it was found that performing LDA on the data set with the diabetes data as a binary classifier was superior to obesity as a classifier based on the AUC of their respective ROC curves. Visual observation of the plotted PCA dimensionality reduction and the visualization of the LDA scores agrees with the difference in AUC values.

References

[Gur+20] Elisa Gurian et al. "Repeated double cross-validation applied to the PCA-LDA classification of SERS SPECTRA: A case study with serum samples from hepatocellular carcinoma patients". In: *Analytical and Bioanalytical Chemistry* 413.5 (2020), pp. 1303–1312. DOI: 10.1007/s00216-020-03093-7.