Evaluating LDA dimensionality reduction for seperability and classification for effectiveness

Abstract

PURPOSE: Testing the effectiveness of LDA dimensionality reduction for the purpose of separation, and LDA classification for the purpose of accuracy.

METHODS: LDA dimensionality reduction is performed on both the datasets and displayed. Then LDA classification is performed on both the datasets, and tested via ROC curve and confusion matrix examination.

RESULTS: The LDA regarding the diabetes dataset performed slightly above average on separability and accuracy, the LDA regarding the obesity dataset did not provide any significant changes from randomly generated diagnosis.

CONCLUSIONS:The diabetes dataset has significant enough results in order to create diagnosis, and thus the doctors gather enough information to make correct diagnosis. The obesity dataset is not significant enough to create diagnosis, thus it needs more information to be able to become reliable.

Introduction

Health is a complex topic with many unknown variables. Data is expensive to gather, and thus doctors are usually working with too little information due to these costs. A method to test if enough information has been gathered would be any supervised classification algorithm. A very popular one is Linear Discriminant Analysis (LDA). LDA uses scatter matrices in order to calculate the Fisher's Linear Discriminant, which is used to reduce dimensionality and produce vector scores. LDA can be used to reduce the dimensionality of data and separate it, or it can be used to classify data. In this case LDA will be used in order to test how well the health variables can detect Diabetes and Obesity within a dataset of 520 patients. The accuracy of the classification will be tested to see if it is statistically significant, when compared to random chance . The accuracy of the LDA can be evaluated in many ways. The tested ways will be analysis of ROC curve and a confusion matrix. The ROC curve is a display of the true positive rate in comparison to the false positive rate, depicting the overall correctness of the classification.

Methods

The data was loaded and prepared by separating the data into two groups, diabetes and obesity. The label vectors are also extracted. The data is then reduced using the PCA method to get data down to two dimensions. For each of the groups, scatter matrices were created for both the between label scatter matrix and within label scatter matrix. From these scatter matrices, the fisher discriminant is created, then LDA axes are created from those. LDA scores are created

from multiplying the original matrix and the LDA axis, then plotted for visibility. After this, the LDA effectiveness is tested via the creation and plotting of ROC curves. This was done for each of the sets by iterating through each of the elements within the set, gathering the TPR and FPR via equations taught in class and saving that to a list. The area under the curve was gathered by the given function, passing in the TPR and FPR lists. The thresholds are tested via the measure of accuracy, saving the largest accuracy and the corresponding false positive rate, true positive rate and the threshold value.

Results

Table 1: AUC values and confusion matrices of the datasets

	AUC Value	Confusion Matrix
Diabetes Data	0.7697	35 50 12 184
Obesity Data	0.5332	0 35 1 250

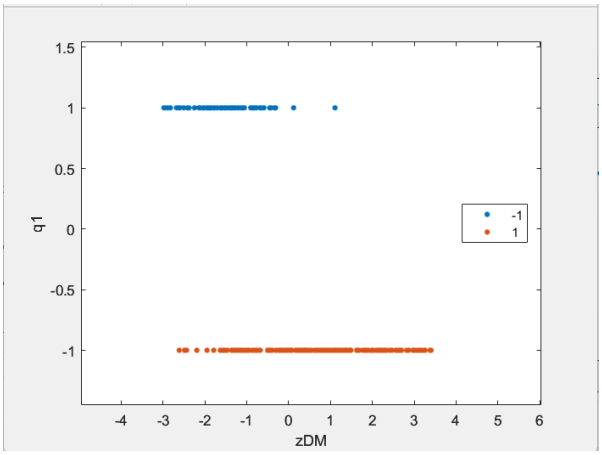

Figure 1: LDA separated Diabetes Dataset,

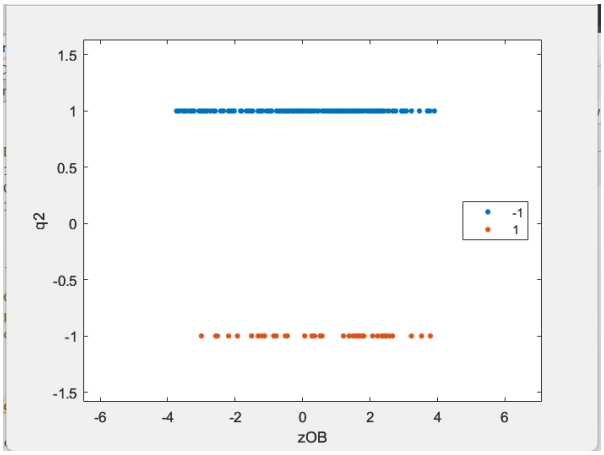

Figure 2: LDA separated Obesity Dataset

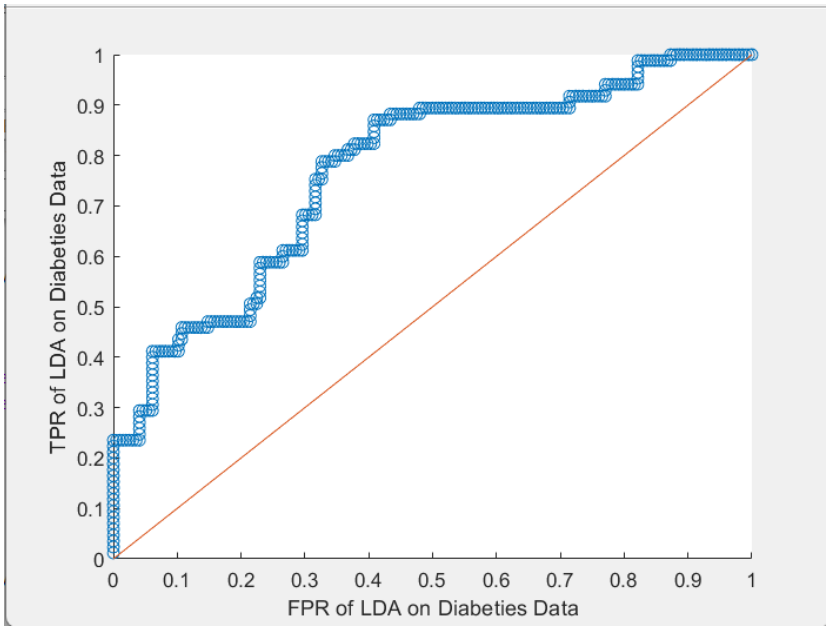

Figure 3: ROC curve depicting LDA on Diabetes Dataset

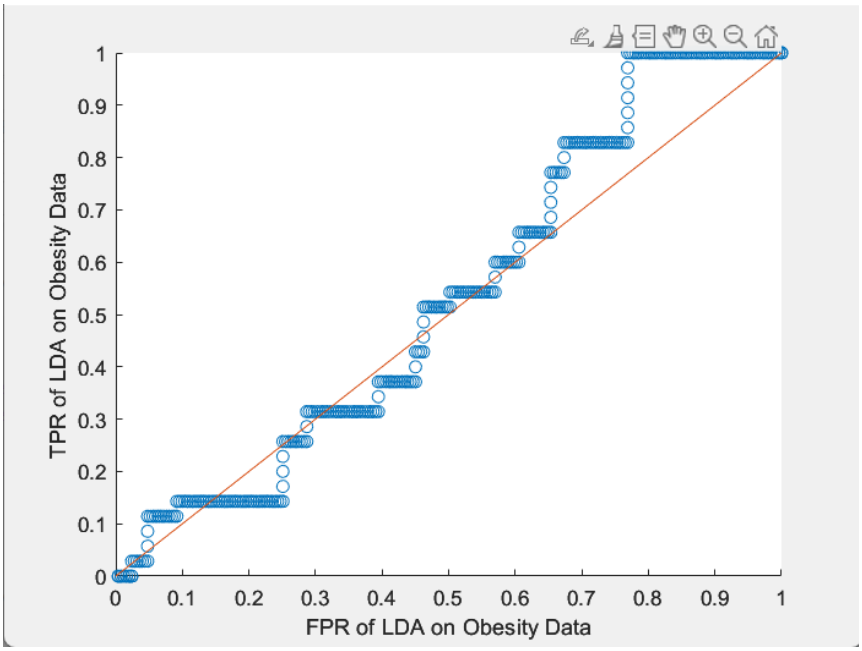

Figure 3: ROC curve depicting LDA on Obesity Dataset

Discussion

Visually inspecting the separated two datasets, LDA does a much better job separating the diabetes dataset compared to the obesity dataset. This can be seen in how each set spans the

entire row in the obesity graph, showing a lot of overlap. The diabeties data has a much more distinct seperation between the sets, the diabeties negative group being biased towards the left, and the positive group being swayed towards the right.. This can reflect about the data itself, showing how the dataset contains enough information to diagnose diabetes, but not enough to diagnose obesity. Thus, with the data available, doctors are able to read and diagnose if someone has diabeties or not, the same cannot be said for obesity. LDA has done a very good job expressing separability, where it exists within the data, however, for separability that does not exist within the data, it is much less effective. Looking at the trends of the ROC curves for both of the datasets. The ROC for obesity follows a linear path of approximately y = x. This is about the same error as randomly guessing, and thus is a not very effective model. The ROC curve for diabetes though does increase quite quickly and goes above the y = x line, proving much more accuracy than the obesity dataset. With the most accurate obesity confusion matrix, there are no TP observations. This is likely due to a flaw in the dataset, not providing enough information, and thus the most accurate way to diagnose people is to say they are mostly negative.. This model can then create issues in the medical force, where creating false negatives has very severe reprocussions. The diabetes set was a more accurate dataset. It has true positives, so it was able to diagnose diabetes relatively well. The issue is with the proportionally large amount of false negatives that the model produced. The AUC values describe the overall accuracy of the model. The obesity data has around 50% accuracy which is not significant enough for diagnosis, further proving that it needs more data. The diabetes data is much more significant at 77%, resulting in a much more reliable model, however a 23% incorrect rate is much to high and should not be used. The LDA classified the data well for the diabetes dataset, and did not classify the data well for the obesity dataset. Neither of these datasets are accurate enough for medical purposes.