Linear Discriminant Analysis and Classifier Assessment

ABSTRACT

PURPOSE: Use linear discriminant analysis (LDA), followed by a ROC curves assessment to produce an effective linear separation on two datasets.

METHODS: Comparing PCA plots, LDA scores plots, and ROC curve figures to evaluate the effectiveness of LDA classification.

RESULTS: The method using LDA successfully created a linear separation for Diabetes dataset compare to the Obesity dataset, with the AUC score of 0.9309 and 0.6071 respectively.

CONCLUSION: On a highly unbalanced dataset, using ROC AUC to assess prediction results is a favorable approach over Accuracy, thus providing an insight into the relative trade-offs between true positives and false positives.

Word count: 98

INTRODUCTION

The objective is to evaluate the linear separation on two different datasets with binary labels. In this study, the LDA classifier is the subject for implementation and evaluation.

Principle Component Analysis (PCA) is a mathematical method used for dimensionality reduction in linear data analysis. When working with binary labels, the convention in machine learning is to use values 1st and 2nd labels of observations. A powerful concept that manages the observations with labels is Linear Discriminant Analysis (LDA), which has been constructed using linear algebra and the Rayleigh quotient. LDA application can be exploring using PCA and scatter matrices. [1] By first producing a visualization of PCA, a 'preferred' coordinate frame for linear separation can be spotted. This location is centered at the mean of the data. This point is aligned with the eigenvectors of the scatter matrix of the zero-mean form of the data. The labels are then projected onto a hyperplane, and this can describe whether the axis does a good job at distinguishing the labels of observations.

The study works with a "tall and thin" data matrix A, with a larger number of rows to columns. If data matrix A has binary labels, it can be partitioned into 2 data matrices, one including observations with label y = +1 and one with observations with label y = -1. The original data matrix can be written, as $A = [\frac{A_1}{A_2}]$. The mean of the original data can be calculated using the mean observation of partition A_1 and A_2, as $\bar{A} = \overline{A_1} + \overline{A_2}$. Furthermore, the zero-mean matrices can be found for the individual partitions can be calculated as $M_1 = A_1 - \vec{1}^T \overline{A_1}$ and $M_2 = A_2 - \vec{1}^T \overline{A_2}$. The four scatter matrices that are associated with the 2 partitioning includes three within-label scatters, defined as

$$S_1 = M_1^T M_1$$

$$S_2 = M_2^T M_2$$

$$S_W = S_1 + S_2$$

The last scatter matrix is a between-label scatter, defined as

$$S_B = \begin{bmatrix} \overline{A_1} - \bar{A} \\ \overline{A_2} - \bar{A} \end{bmatrix}^T \begin{bmatrix} \overline{A_1} - \bar{A} \\ \overline{A_2} - \bar{A} \end{bmatrix}$$

LDA's goal is to maximize the ratio of the Rayleigh quotients, this is achieved by minimizing the within label-scatter and maximizing the between-label scatter. Assuming that the within-label scatter matrix S_W is symmetric positive definite. The vector LDA is this separating hyperplane in binary classification.

With the optimal classifier LDA, the confusion matrix is a tool to assess the performance of this hyperplane. To go beyond considering the accuracy based on the number of "right" and "wrong" instances, the confusion matrix considers 4 components [2]
True Positive (TP): true label is +1 and prediction is +1
False Negative (FN): true label is +1 and prediction is -1
False Positive (FP): true label is -1 and prediction is +1
True Negative (TN): true label is -1 and prediction is -1
In this work, we focus on constructing True Positive Rate (TPR) and False Positive Rate (FPR), defined as $TPR = \frac{TP}{P}$; $FPR = \frac{FP}{N}$, which measures the hit rate and false-alarm rate respectively.

The Receiver operating characteristic (ROC) curve is a graph showing the performance of classification at a variety of classification thresholds. The ROC curve utilizes FPR and TPR directly as its first and second entry respectively. To compute the scatter points in a ROC curve, a sorting-based algorithm called Area under the ROC Curve (AUC) is implemented. AUC measures the 2D area underneath the entire ROC curve from (0,0) to (1,1). Two possible reasons that make AUC a versatile tool are its scale-invariant and classification-threshold-invariant. [3] A higher value of AUC indicates that the performance of the classifier is better at telling apart positive and negative classes.

The dataset used in this study is from UCI Machine Learning Repository. This dataset contains the sign and symptoms of newly diabetic or would be diabetic patient, collected using direct questionnaires from the patients of Syhet Diabetes Hospital in Sylhet, Banglades and approved by doctors. This is a multivariate dataset with 17 attributes, and 520 instances. The associated task for this dataset is binary classification.

This study attempts to answer the scientific question of whether LDA classification can provide a linear separation for two different datasets. The result from LDA scores is then evaluated under the ROC curve method.

METHODS

This study produces an assessment of Linear Discriminant Analysis (LDA) classification using ROC curve and confusion matrices. The entire csv data file is loaded into the program, without the first-row storing column names. The last two columns of the dataset are used as the label columns. Two new datasets are created, both with all data columns and either with column 16 as the labels for

Obesity diagnostic or column 17 as the labels for Diabetes diagnostic. These two datasets will be used to analyze the effectiveness of LDA.

The first given task was to reduce the standardized data to 2D using PCA and then to compute the LDA axis and scores. This is to determine how efficient is LDA on the two sets of Obesity and Diabetes data.

To inspect the data with PCA, the program has pre-computed reduced dimensionality for both Obesity and Diabetes dataset. Two PCA graphs are created to visualize the data and how they scatter for two classes (positive and negative).

To find the LDA vectors and scores, the program first computes the LDA axis for each dataset. This is done using lda2classfunction, where the parameters are the partitions from an original matrix X. In this case, the partition is determined by the diagnostic label X1 for all positive cases (+1) and X2 for all negative cases (-1). Once inside lda2class, the program calculates the zero mean matrices for the parameters respectively. These two matrices can then be used to construct the first three within-label scattered matrices of this partition. S_1, S_2, and S_w. The function then continues to calculate the final between-class scatter matrix S_b. The Rayleigh quotient is calculated as the linsolve result of S_w on S_b. Using the eigs function, the largest eigenvector is extracted from from the Rayleigh quotient as the Fisher's linear discriminant vector, qvec. This vector is checked for correction to point towards mean of X1.

From the lda2class function, the returned qvec value is the LDA axes that can be used to calculate LDA scores by projecting the respective Xmat to the qvec. With the LDA axes and scores, we can plot the LDA scores against themselves to see the effectiveness of linear separation on the Obesity and Diabetes dataset. By using the scores on both x and y axes, an overlap diagonal line demonstrates how all data points are projected on to a single chosen hyperplane.

The second task is to compare ROC curves for classifiers and comparing confusion matrices to find the best choice of threshold for the LDA scores. The program implements the roccurve function, passing in the true labels and the LDA scores as parameters. Inside the function, the scores are first permuted and labeled accordingly. A unique subset of the LDA scores is stored and used for testing as threshold values. For every single threshold value, a confusion matrix is created to keep track of the threshold that create the highest accuracy. The confusion matrices are created using a separate confmat function. Within this function, true y labels, LDA scores, and a single threshold value are utilized as parameters. Every value in yvec will be compared against the scores to generate a confusion matrix for that specific threshold value. A two-by-two matrix is returned from this function. AUC score will then be calculated by passing sorted tpr and fpr arrays into the existing aucofroc function.

The best threshold values for Obesity and Diabetes data are then displayed along with their corresponding confusion matrices, and two ROC curve graphs for the two datasets using the results from the roccurve function.

This program was tested on "dmrisk.csv" dataset from the Machine Learning Repository of the University of California at Irvine. The dataset describes health-related values for 17 variables of

520 study participants. The dataset is propagated with categorical binary variables coded into +1 and -1. (Female vs. Male, Yes vs. No, Positive vs. Negative).

To evaluate the result from the study, the plotting of PCA, LDA scores, and ROC curves are generated for both datasets. For the Obesity data partition, the best threshold value is 3.7875 and the AUC score is 0.6071. For the Diabetes data partition, the best threshold value is -0.4998 and the AUC score is 0.9309. This demonstrates that LDA has a more sufficient application on Diabetes data compare to Obesity data, based on the AUC scoring.

RESULTS

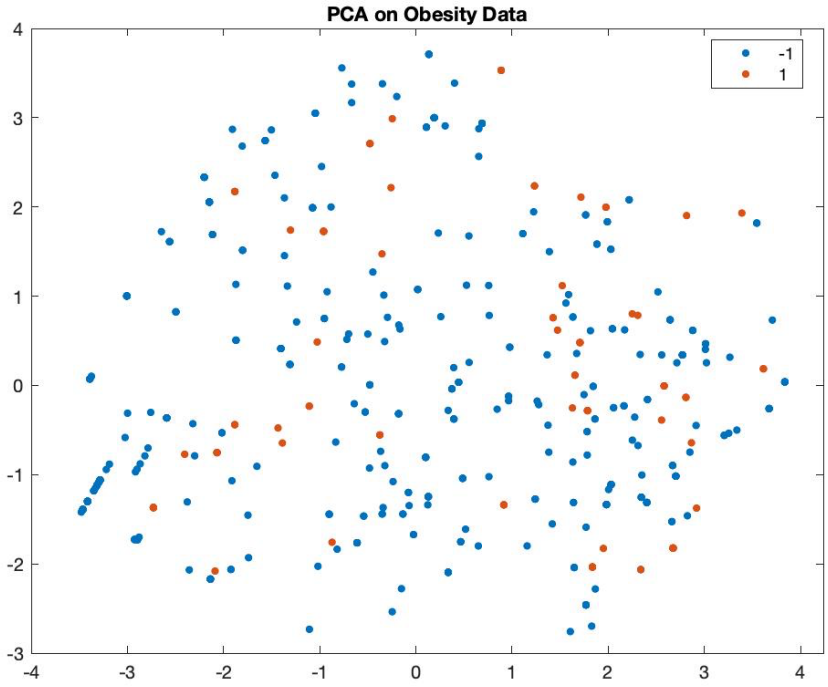

Figure 1: Plot of PCA on Obesity data. The figure demonstrates clustering into 2 groups of Obesity diagnosis (Positive as +1, Negative as -1)

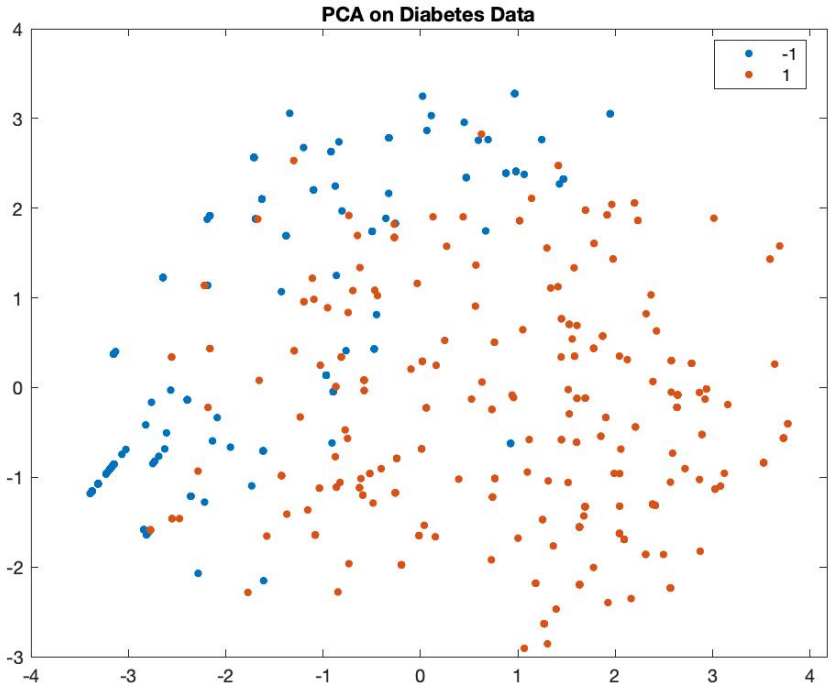

Figure 2: Plot of PCA on Diabetes data. The figure demonstrates clustering into 2 groups of Diabetes diagnosis (Positive as +1, Negative as -1)

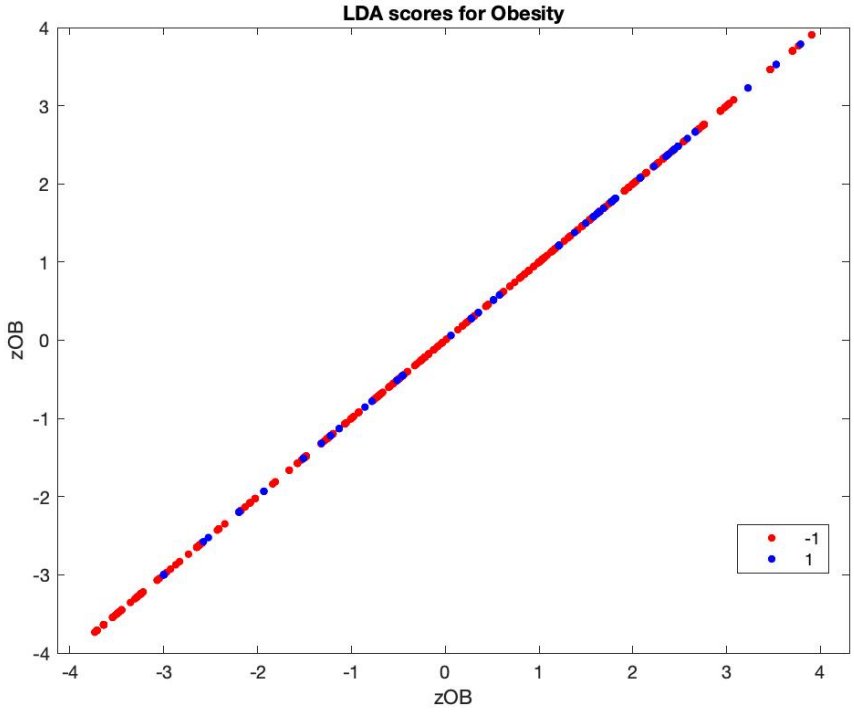

Figure 3: Plot of LDA on Diabetes data. The figure demonstrates linear separability of the two class of Obesity diagnosis

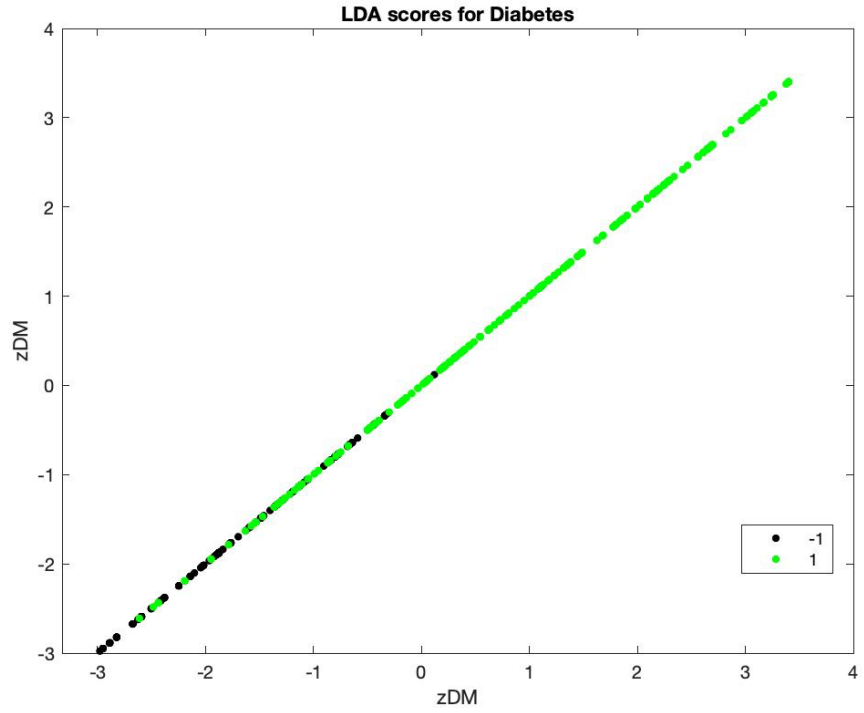

Figure 4: Plot of LDA on Diabetes data. The figure demonstrates linear separability of the two class of Diabetes diagnosis

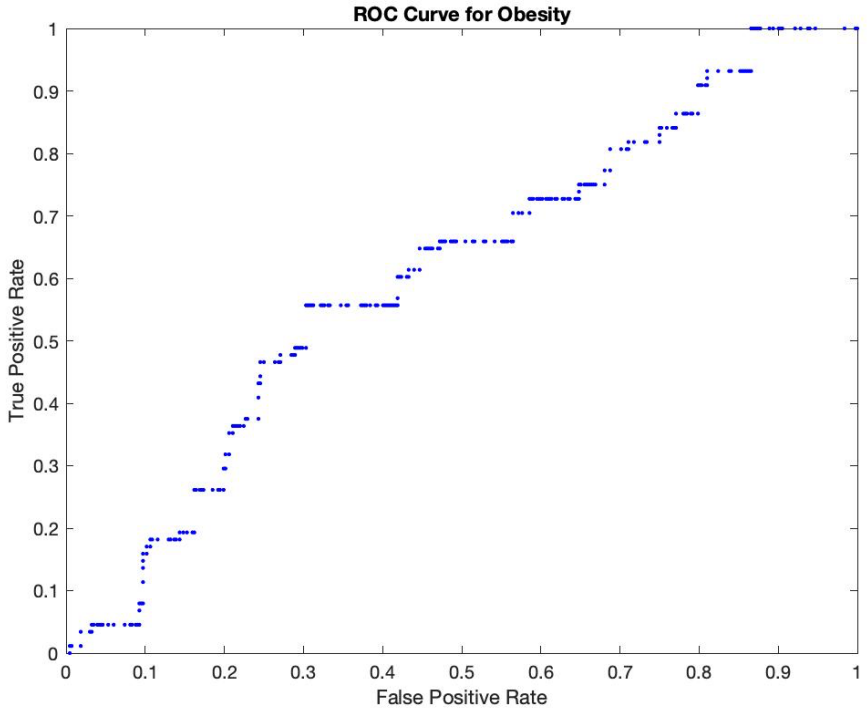

Figure 5: Plot of ROC Curve on Obesity data. The figure demonstrates the relationship between clinical sensitivity (True Positive Rate) and specificity (1-specificity)

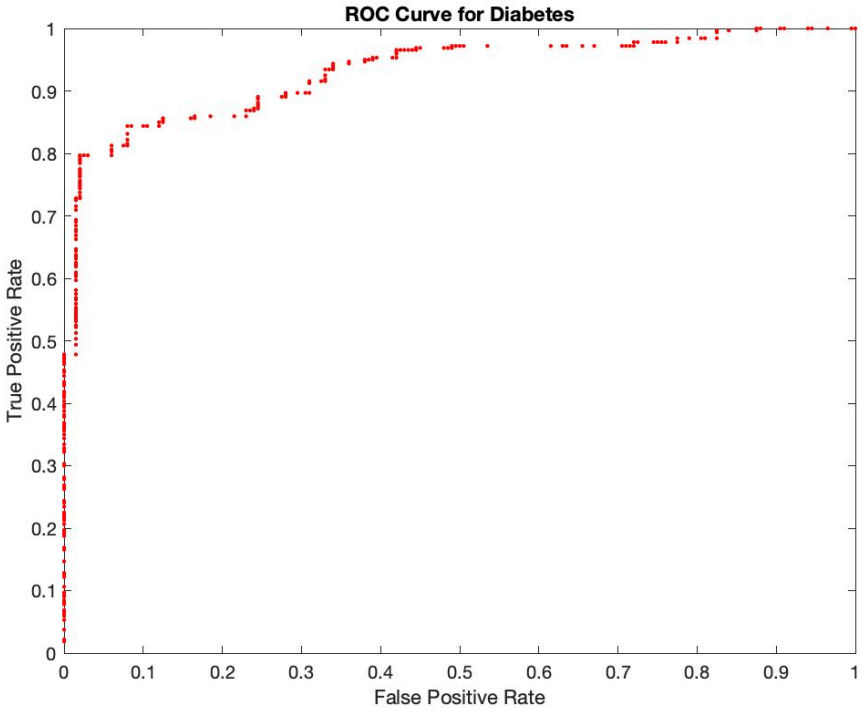

Figure 6: Plot of ROC Curve on Diabetes data. The figure demonstrates the relationship between clinical sensitivity (True Positive Rate) and specificity (1-specificity)

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

Diabetes:
AUC = 0.9309

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

Obesity
AUC = 0.6071

		+1	**-1**
Labels	**+1**	1	87
	-1	2	430

DISCUSSION

The results demonstrate successful application of LDA classification to the Diabetes dataset compared to the Obesity dataset. Upon inspecting the two PCA plots, PCA on Diabetes (Figure 2) data shows a clearer distinction between the two classes compare to PCA on Obesity (Figure 1) data. Figure 2 suggests a linear separation line can be drawn diagonally with a slope around 1 between the two classes. Figure 1 has the data points from two classes are more intertwined, making it more challenging to intuitively assign a lienar separation line upon looking at it. It is important to note that any king of projection to a lower dimension will result in some loss of information, in this study it's the PCA projection. [4] Upon inspecting the two LDA scores plots, LDA scores on Diabetes (Figure 4) data shows a distinct transition area from one color to another. Black data points are shown to concentrate more on the left end of the axis, and green data points on the right. LDA scores on Obesity (Figure 5) data has the two color of data points (red and blue) overlapping all throughout the

axis. Based on the first four figures, LDA seems more promising as a classifier for Diabetes data. Considering Figure 5 and Figure 6, the initial inspection is proved correct as ROC Curve for Diabetes (Figure 6) shows significant AUC compared to ROC Curve for Obesity (Figure 5). In the calculation, AUC values are 0.9309 and 0.6071 for Diabetes and Obesity data respectively.

Taking a closer look at the result, the accuracy with the best threshold for Diabetes data is 87.3 % and 82.9% for Obesity data. Considering solely the accuracy, the two results seem deceivingly comparable. This is due to the fact that in Obesity dataset, the problem itself is highly imbalanced. There are only 88 positive instances compared to 432 negative instances. Thus, a high accuracy of 82.9% is achieved simply by predicting more -1 labels, which is the majority class here. AUC deals with situations where the problem is highly skewed sample distribution, and the goal is trying to avoid overfit most predictions to a single class. It measures how true TPR to FPR trade off, by evaluating the classifier as the threshold varies through all possible values.

In health risk prediction, with the same 15 attributes, it's apparent that predicting Diabetes is a more obtainable classification task comparing to predicting. These attributes are collected through a questionaire format, making them accessible to gather from patients comparing to running formal test. We can conclude that Diabetes diagnosis can be run in a more efficient manner on patient collected data. Obesity diagnosis are more challenging to predict with simple health factors attributes that can be collected from patient questionaires.

LDA limitations come with the initial assumptions that must be satisfied to implement it. For example, every feature should have a Gaussian distribution, with a bell-curved shape. Feature should hold the same variance, and has values varying around the mean. Feature should be collected and sampled randomly, without any bias or influence. The features should not have correlation else prediction power can decrease.

In real world application, there are various techniques that can take advantage of LDA reduction approach, among those PCA is also a great alternative. As long as the problem is a classification problem, LDA can be implemented, such as speech/ text recognition, face recognition, biometrics. With facial recognition, faces are displayed in a large number of pixel values, LDA then reduce the number of features to a managagble count prior to classification process. In medical application, LDA is used for diagnosis purpose, much like this study. The task can be classifying patients' conditions and severity based on parameters and the mediacal treatment the patients are receiving in order to adjust the treatment accordingly.

REFERENCES

[1] Ellis Randy E. Class 23: Linear Discriminant Analysis – LDA. [unpublished lecture notes]. CISC 271: Linear Data Analysis, Queen's University lecture given 2021 March. Lecture notes can be found at: https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class23.pdf

[2] Ellis Randy E. Class 24: Classification – Assessment By Confusion Matrix. [unpublished lecture notes]. CISC 271: Linear Data Analysis, Queen's University lecture given 2021 March. Lecture notes can be found at: https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class24.pdf

[3] Classification: ROC Curve and AUC|Machine Learning Crash Course [Internet]. Google. Google; [cited 2021Apr2]. Available from: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[4] Silva TS. An illustrative introduction to Fisher's Linear Discriminant [Internet]. An illustrative introduction to Fisher's Linear Discriminant - Thalles' blog. 2019 [cited 2021Apr2]. Available from: https://sthalles.github.io/fisher-linear-discriminant/