CISC 271, Winter 2021
Assignment #4: Linear Discriminant Analysis and Classifier Assessment
Due by 4:00PM on Wednesday, March 13, 2024

The subject matter for this assignment is health data, specifically risk of early-stage diabetes.

Coding for this assignment is relatively modest. This assignment requires multiple uses of linear discriminant analysis (LDA), followed by an assessment of LDA classification using ROC curves. You may use basic `MATLAB` builtin functions. You are expected to write a small amount of code to compute confusion matrices and ROC curves, as part of learning about these prominent methods for assessing a classifier.

Please read the details and instructions carefully before you begin to work on the problem. There must be a single results section and a single discussion section on your report. The results section of the report must contain one table and between two and six figures; more or fewer, of either tables or figures, may produce deductions from your grade on this assignment.

Statement of Academic Integrity

This assignment is copyrighted by the instructor, so unauthorized dissemination of this assignment may be a violation of copyright law and may constitute a departure from academic integrity.

Sharing of all or part of a solution to this assignment, whether as code or as a report, will be interpreted as a departure from academic integrity. This includes sharing of the assignment after the due date and after completion of this course.

Learning Outcomes

On completion, a successful student will be able to:

- Implement a method for LDA of data that have binary labels
- Compute scores of data from an LDA axis
- Implement a computation for a binary-label confusion matrix
- Implement a computation of an ROC curve from a set of confusion matrices
- Evaluate a data set by comparing two classifications of the data

Preliminary: The Data

The data were gathered from the Machine Learning Repository of the University of California at Irvine and processed for use in this class. The original data, and a full description, were from

`https://archive.ics.uci.edu/ml/datasets/`
`Early+stage+diabetes+risk+prediction+dataset.`

The data set describes health-related values for 17 variables of 520 study participants. Categorical binary variables were coded into ±1 as:

- "Female" as $+1$, "Male" as -1
- "Yes" as $+1$, "No" as -1
- "Positive" as $+1$, "Negative" as -1

The final column indicates whether the participant was diagnosed with early-stage diabetes milletus type 2, here abbreviated as "diabetes". The second-to-final column indicates clinical obesity.

The data are in the file `dmrisk.csv` that is in the ZIP file for this assignment. The data are loaded by the "starter" code.

Preliminary: The Code

The instructor has provided "starter" code that you will modify. Of note is that the "area under Curve", or AUC, computation is provided in full; certain technical details of AUC computation, related to effective use of the Trapezoidal Rule for numerical integration, are needed but are beyond the scope of this course. You can use the AUC function without modification.

The "starter" code has:

- A main function that will invoke the functions to answer the two question in this assignment. The main function will dimensionally reduce the data to 2D for your visualization and analysis. You will create your relevant plots and display values to the `MATLAB` console.
- A function `a4q1` that you will modify for Question #1. This will find the LDA axis and LDA scores for the "diabetes" label and for the "obesity" label. These vectors will be returned to the main function.
- A function `lda2class` that you will modify to compute the LDA axis for a given label. It will receive as input two sets of data, possibly with different numbers of observations and definitely with the same number of corresponding variables. It will return the LDA axis that is directed from data that have the second label and towards data that have the first label.
- A function `confmat` that you will modify to compute a confusion matrix for a binary label. It will receive as input a label vector, a score vector, and a threshold value; it will return a 2×2 confusion matrix. Please note that *the confusion matrix must have the structure in the class notes*; there is no universal convention for the presentation of a confusion matrix and you must follow the convention from the notes.
- A function `roccurve` that you will modify to compute an ROC curve. It will receive as input a label vector and a score vector; it will return a vector for the True Positive Rate (TPR), a vector for the False Positive Rate (FPR), the AUC of the ROC curve, and a threshold for the optimal accuracy of the scores. The TPR and FPR should be computed from the confusion matrices of thresholds, which are one threshold for each unique score in the input score vector.
- A function `aucofroc` that you will *not* modify. This function computes the AUC value from the ROC vectors.

Question 1: Linear Discriminant Analysis 10% of Final Grade

For this question you will need to modify the main function, the function `a4q1`, and the function `lda2class`, all in the starter code.

The conceptual problem for this question is: how well does LDA separate the data when labelled for diabetes and for obesity? The method is to reduce standardized data to 2D, then to compute the LDA axis and scores.

In `lda2class`, you will need to implement LDA as described in the notes. This can be done with 10 lines of `MATLAB` code that calculate the scatter matrices and find the dominant eigenvector for the ratio of Rayleigh quotients.

In `a4q1`, you will need to understand the instructor's starter code and add a couple of lines that compute the LDA scores from the data and the LDA axes.

In the main function, you may want to plot your data and results. For up to two plots for each of the diabetes labels and the obesity labels, you might plot the 2D versions of the data and the LDA scores. These are optional and the choice depends on how you want to answer the conceptual problem for this question. The plots, up to four in total, can be used to explain what you found and can augment your explanation of your findings in the second question for this assignment.

Question 2: ROC Curve And Assessment 5% of Final Grade

For this question you will need to modify the main function and the functions `roccurve`, both in the starter code.

The conceptual problem for this question is: how well does LDA perform when it is used to classify the data for diabetes and for obesity? This assessment should be done in two ways: comparing the ROC curves for the classifiers, and comparing confusion matrices for "best" choices of thresholds for the LDA scores.

For each data label, diabetes and obesity, a classifier will have variable performance that depends on the selection of a threshold for the LDA scores of the data. Each threshold produces a confusion matrix, and each confusion matrix can be represented by its relative TPR and FPR values. Together, these rates can be plotted as an ROC curve. The Area Under Curve, or AUC, is one way to assess the overall performance of a classifier. Another way is to examine a confusion matrix at an "optimal" threshold.

You should calculate and plot separate ROC curves for the diabetes and obesity labels. As part of the calculation, you should estimate an "optimal" choice of a threshold; one measure is the product of the TPR and the TNR, minus the product of the FPR and FNR.

Your data should be summarized in a single table that displays the AUC and example confusion matrix for the data, according to the labels. An example is Table 1. You should assess and discuss the properties of the data, as described by linear discriminant analysis. These data are imperfectly separated by LDA and your discussion can include limitations, which are one way of describing potential future work. This will be slightly challenging because the data set is relatively new and is not well explored by others.

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

		Diabetes: AUC≈0.50 +1	 –1			Obesity: AUC≈0.50 +1	 –1
Label	**+1**	TP	FN	**Label**	**+1**	TP	FN
	–1	FP	TN		**–1**	FP	TN

3: Grading Guide

We will test your code using `MATLAB`. Your grade will be reduced if: you plot more or fewer than the specified number of figures; your code outputs anything other than the specified values; or you otherwise deviate in your implementation from these requirements.

The TAs will use this guide when they mark your assignment. Your grade will be based on your code, results, and report. The distribution of points for the assignment grade are:

3/30 points: all and only the numerical values that are produced by the code and that are presented in the results

5/30 points: quality of the code in the modified "starter" functions, and any other changes in the submission file that was used to generate values and plots for the report

22/30 points: quality of the report, especially including the figures and discussion; clarity may be assessed, in part, by the written introduction, assessment of findings, and the discussion of results

What to turn in:

- You will submit your answers electronically as two files. The code will be tested by one or more graders. The PDF report will be read by one or more graders and will be checked, using electronic methods, to ensure that it meets professional standards for originality.
- The code must be in a single `MATLAB` file, `a4_xxxxxxxx.m`. This file will contain all of the code needed to verify that the values and tables in the report can be reproduced. The functions must produce the values for your table and the figures.
- Your function must take no arguments, return no values, and require no user input or action such as using the "enter" key. Running this function should produce, on the console, every value that is in the report; the functions should also produce every plot that is in your report. The functions should produce no other values or figures. The graders will compare your computed values to the values in the report and may deduct marks from the report for differences between any reported value/plot and the corresponding computed value/plot.
- The report must be in a single PDF file (`a4_xxxxxxxx.pdf`). The PDF file must include a description of how you tested your code. You can also include notes, comments on the problems, and assumptions you have made, as appropriate.
- You may assume that the file `dmrisk.csv` is in the current directory when a grader tests your code.
- The assignment must be submitted using the Queen's "onQ" software.

Grading Considerations:

- The quality of your report will be considered. You need, at minimum, to conform to the "student version" of the report style in the onQ website; you may wish to consider the "grader version" that we will use for assessing your report.
- The quality of your MATLAB code will be considered. Your code should be appropriately indented, sufficiently commented, and otherwise be appropriate software.
- The output of your code will be considered.
- Your code can use functions provided by MATLAB, but the code that you submit *must* be your original work.
- Code that causes MATLAB to produce an error or warning will result in a failing grade.

Policies:

- You must complete these questions individually.
- Although you are allowed to discuss the questions with other students, you must write your own answers and MATLAB code.
- The syllabus standards apply to this assignment.
- Lateness policy applies starting the minute after the submission deadline, at a rate of 20% off the assignment value per calendar day. *Please note: the time in the onQ system is beyond your control, so submitting within an hour of the deadline is inherently a risky process for which you assume full responsibility.*