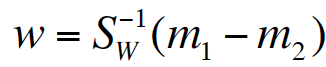
**IE 6318 Data Mining and Analytics**

**Homework 4**

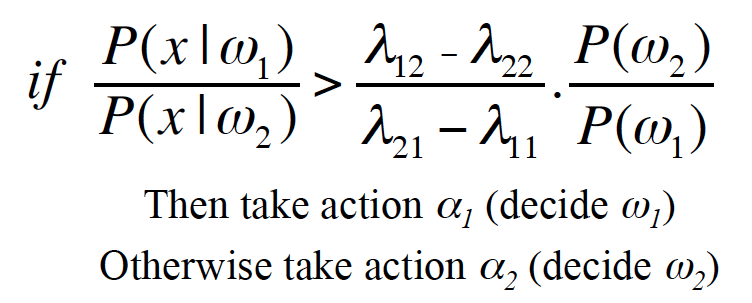
Classification Using Fisher Linear Discriminant and Perform ROC Analysis

1. Make a classification function based on Fisher Linear Discriminant. From the lecture, we introduced the optimal projection direction *w* is:

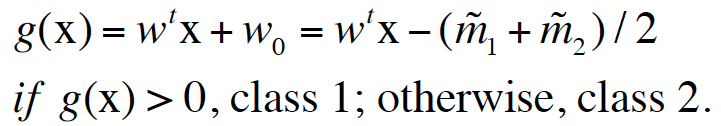
.

One can perform classification on the one-dimensional space for the projected data samples *wtx. Make the function with two classification model choices:*

*1)* Using Bayesian Decision Boundary based on the derived decision making rule in the lecture:



2) Using the middle line of the projected means as the decision boundary for classification, that is



3) For each classification model above, perform the binary classification for the Breast Cancer Dataset, which can be downloaded from UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>. Report the classification accuracy, sensitivity, and specificity using 5-fold cross-validation. (Sensitivity = the accuracy to detect cancer patients, Specificity = the accuracy to detect healthy subject. Check Lecture slides for more information of sensitivity and specificity.)

In the Breast Cancer dataset, it used 1 to represent Healthy Controls, and 2 to represent Patients. For the Bayesian Classification Model, use the following two choices of penalty costs to make classification rule in your program: (I) λ11 = 0, λ22 = 0, λ12 = 2, λ21 = 1; (II) λ11 = 0, λ22 = 0, λ12 = 1, λ21 = 1. Here λ**ij**is the penalty cost to classify a class **j** sample as class **i**.

2 Perform ROC analyses for the three classification models developed in problem 1 (2 Bayesian models and 1 of the mid-line of projected means), and conclude which model is preferred based on the ROC analysis. ROC analysis is an important criterion to evaluate the overall performance of a prediction model. As shown in the following figure, to generate ROC curve, you can shift the decision boundary from the left end to the right end. For each decision boundary, calculate the corresponding sensitivity and specificity to construct one point in the ROC curve. In this implementation, you can shift the decision boundary from -75 to 75 with a step size of 1. For each step, calculate the sensitivity and specificity for the testing dataset. For each of the three models above, plot the ROC curve and calculate the AUC values, determine which model is preferred in this classification task.

***Note:*** *In this problem, you can use the first 38 samples in Class 1 and the first 48 samples in Class 2 to construct training dataset, and the remaining samples for testing dataset.*

