MATH 569 HOMEWORK 3

Due: March 1^{st}, Wednesday, 11:59pm

How to submit: via Blackboard

If you have multiple files, upload a zipped file

Problem 1

Theory Q

Consider using LDA for a binary classification problem. Suppose we have p-dimensional predictors $x \in \mathbb{R}^p$, and the response variable $y \in \{-1, +1\}$. The numbers of data points in class 1 and class 2 are N_1 and N_2, respectively.

1) If $N_1 = N_2$, show the LDA classification rule.

2) If $N_1 \neq N_2$, show the LDA classification rule.

3) Use linear regression, describe the condition under which the linear regression results in the same classification rule as in LDA.

4) If we use a different coding scheme, e.g., $y \in \{0, 1\}$, how does this affect your conclusion in 3)?

Problem 2

1) Solve the generalized eigenvalue problem:

Theory Q

$$\max_a \quad a^T B a$$

subject to:

$$a^T W a = 1$$

2) When B is the between-class variance, and W is the within-class variance of data X, show that the optimal solution a^* for the generalized eigenvalue problem defined in 1) is the same as the v_l defined on page 114 with $l = 1$.

Problem 3 Consider a binary classification problem with two-dimensional features, i.e., $K = 2, p = 2$. Class 1 has multivariate Gaussian distribution $\mathcal{N}(\mu_1, \Sigma)$, and class 2 has multivariate Gaussian distribution $\mathcal{N}(\mu_2, \Sigma)$. Let $\mu_1 = (1, 2)^T$ and $\mu_2 = (1, -2)^T$. $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ is the common covariance matrix.

Show the decision boundary in each of the following cases. First show the equation of the decision boundary, then draw the decision boundary as well as the data points (by sampling from the distributions) in a plot. Show the plots in a similar fashion to Figure 4.9.

1) Use LDA without dimension reduction.

2) Use reduced rank LDA by projecting data to the direction of greatest centroid spread.

3) Use reduced rank LDA by projecting data to the discriminant direction.

Problem 4

1) Consider the vowel training data, which has 11 classes with $X \in \mathbb{R}^{10}$. Follow the procedure on page 114 to find the first and second discriminant variables (also called canonical coordinates), then draw the scatter plot and mark the centroids in the two dimensional subspace spanned by the first two canonical variates (see Figure 4.11).

2) Draw the decision boundary as in Figure 4.11, and report the classification accuracy.

Problem 5

Use linear regression and logistic regression to classify the vowel data in Problem 4, and report classification accuracy for each of them.