IE 6318 Data Mining and Analytics
Homework 4
Classification Using Bayesian Decision Theory and LDA

1. Complete the function for Bayesian Classification, which makes classification based on the posterior probability:

$$P(\omega_j \mid x) = \frac{P(x \mid \omega_j)P(\omega_j)}{P(x)}$$

In the provided sample codes, we have provided a Bayesian Classification function with two model options:

1) Assume the four features are independent and follow normal distributions. This function option selects **Naïve Bayes to** calculate $P(x \mid \omega_j)$ and predict class labels for testing data.

2) Assume the four features follow multivariate normal distribution. In this function option, for each testing sample, it first calculates the likelihood probability $P(x \mid \omega_j)$ using multivariate normal distribution, and the prior probability $P(\omega_j)$; then the Bayesian decision model selects the class that maximize $P(x \mid \omega_j)P(\omega_j)$

In this problem, **complete the third model option for the Bayesian classification function**. In the lecture slides, we showed the discriminant function that derived based on the assumption that the feature vector ***x*** follows multivariate normal distribution (In slides, the general case with the covariance matrix of each class **Σ**

$$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$$

$$where: \mathrm{W_i} = -\frac{1}{2}\Sigma_i^{-1}$$

$$\mathrm{w_i} = \Sigma_i^{-1}\mu_i$$

$$\mathrm{w}_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln\left|\Sigma_i\right| + \ln P(\omega_i)$$

Perform classification for the IRIS dataset using the Bayesian classification function made above. For each classification option, perform 5-fold cross-validation, and report the prediction accuracy and confusion matrix for each fold, and also report the overall prediction accuracy and confusion matrix for 5 folds. Which model option generates the best classification performance?

2. Make a binary classification function based on Fisher Linear Discriminant Analysis (LDA). From the lecture, we introduced the optimal projection direction *w* is:

$$w = S_W^{-1}(m_1 - m_2).$$

One can perform classification on the one-dimensional space for the projected data samples $w^t x$*. Make a binary LDA classification function* using the derived Bayesian Decision Boundary in lecture 4:

$$if \ \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Then take action α_1 (decide ω_1)

Otherwise take action α_2 (decide ω_2)

Perform LDA 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. 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) $\lambda_{11} = 0$, $\lambda_{22} = 0$, $\lambda_{12} = 5$, $\lambda_{21} = 1$; (II) $\lambda_{11} = 0$, $\lambda_{22} = 0$, $\lambda_{12} = 1$, $\lambda_{21} = 1$. Here $\lambda_{\mathbf{ij}}$ is the penalty cost to classify a class **j** sample as class **i**.

For each penalty cost setting, report the classification accuracy, sensitivity, and specificity using 5-fold cross-validation. Report the results of each fold and the overall performance for 5 folds. (Sensitivity = the accuracy to detect cancer patients, Specificity = the accuracy to detect healthy subject. Check Lecture slides for more information of sensitivity and specificity.)