IE 5300 Data Mining and Analytics
Homework 3
Classification Using Bayesian Decision Theory and Linear Discriminant Analysis

1. For the Iris dataset, perform the 3-class classification using Naïve Bayes classification. Use N-fold cross-validation, report the prediction accuracy and confusion matrix. (Hint: use Matlab commands **NaiveBayes.fit**, **predict** and **confusionmat**, use #fold N=5)

2. Make a Bayesian classification function using the Bayesian decision rule: choose a class that maximizes the posterior probability based on the Bayes formula

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

Assume features of the three classes follow multivariate normal distribution, implement the equivalent decision rule: predict the class that maximize the discriminant function $g_j(x) = p(x|\omega_j)P(\omega_j)$, where $p(x|\omega_j)$ is the probability density of a feature vector *x* given it belongs to class j. In this function, use Matlab function **mvnpdf** to calculate the likelihood probability density $p(x|\omega_j)$ directly. For your function, also make an option for feature dimension equal to 1. For one-dimension feature, you can use Matlab function **normpdf** calculate the likelihood probability density $p(x|\omega_j)$.

For the Iris dataset, perform the 3-class classification using the function, and report prediction accuracy and confusion matrix based on 5-fold cross-validation.

3. Make a Bayesian classification function using the derived discriminant function given features follow multivariate normal distribution. The decision rule is to select the class that maximize the following discriminant function $g_i(x)$:

$$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$$
$$where: W_i = -\frac{1}{2}\sum_i^{-1}$$
$$w_i = \sum_i^{-1}\mu_i$$
$$w_{i0} = -\frac{1}{2}\mu_i^t\sum_i^{-1}\mu_i - \frac{1}{2}\ln|\sum_i| + \ln P(\omega_i)$$

Perform 3-class classification using the function, and report prediction accuracy and confusion matrix based on 5-fold cross-validation.

4. Make a function to achieve binary classification using Fisher Linear Discriminant. Your function should first obtain the optimal projection direction *w*, and then perform classification for the projected features in the projected one-dimensional feature space. For the classification part, you can use two options for classification decision:
a) use the decision boundary = $(\tilde{m}_1 + \tilde{m}_2)/2$, where $\tilde{m}_1$ and $\tilde{m}_2$ are the means of the projected values of class 1 and class 2, respectively.
b) use the Bayesian classification function you made in problem 2 to make classification in the projected one-dimensional feature space.

Select iris class versicolor and virginica for the binary classification with 5-fold cross validation, report sensitivity and specificity. Here Define: *sensitivity = the prediction accuracy for the class labeled as 1, specificity = the prediction accuracy for the class labeled as -1.*