비주얼 컴퓨팅 Project #1

In 1935 E. Anderson collected the measurements of the petal and sepal leaves of three different species of the Iris of the family Iridaceae: "*Iris Setosa*", "*Iris versicolor*", "*Iris virginica*". One year later R.A. Fisher used the data to illustrate the concept of Linear Discriminant Functions. Four different features were investigated: sepal length, sepal width, petal length and petal width. For each of the three flower classes 50 samples were collected, totaling 150 samples. The features are all continuous. Since then this data set has turned into a classic in pattern recognition.

The different classes in this example are species of flowers. The number of classes C=3 with W={*Setosa, Versicolor, Virginica*}. The sensor in this case is a ruler, and 4 measurements (features) are made. The classification task is to discriminate the flowers by the 4 attributes.

1) For the purpose of this homework assignment, the Iris data set is divided into a Training set (Iris_train.dat) and a Testing set (Iris_test.dat). The training set has 40 samples for each class, and the testing set has 10 samples for each class.

a. Assume that each class has a normal distribution, and use the training data to estimate the mean and covariance of each class.
b. Determine the decision boundaries
c. Classify the testing data set and construct the **confusion matrix**.

The confusion matrix is a CxC matrix (3x3 in this case), where c_{ij} = Number of samples classified as class i, and their true class is j.

2) In this problem we will use the first 2 features of the Iris data.

a. Plot the training data samples (each class should be assigned a different symbol).
b. Estimate the mean and covariance of each class.
c. Plot the means and the contours for which the Mahalanobis distance = 2. (same plot as in (a))
d. Determine the decision boundaries and plot them (same plot as in (a))
e. Add the test data set to the plot. (correctly classified points, misclassified points, and points from different classes should be assigned different symbols).
f. Classify the test data set (using discriminant functions and using the graph) and construct the confusion matrix. Compare the results with the previous problem.