**IE 6318 Data Mining and Analytics**

**Homework 3**

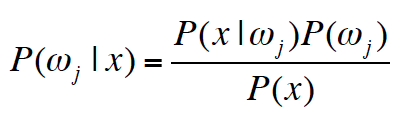
Classification Using Bayesian Decision Theory

Still do the classification task for the Iris flower data set: three types of Iris flowers: setosa, versicolor, and virginica. Four attributes: sepal length, sepal width, petal length, petal width (unit: cm).



Iris Setosa Iris Versicolor Iris virginica

1. Make a program to perform the 3-class classification for the IRIS dataset using the Bayesian decision rule: choose a class that maximizes the posterior probability. Based on Bayes formula

.

the decision rule is equivalent to select the class that maximize. Make a program to run **5-fold cross validation** and report the prediction accuracy and confusion matrix for two conditions:

1) Assume the four features are independent and follow normal distributions, use **Naïve Bayes to** calculate. (Hint: if you use Matlab, you can use the functions **NaiveBayes.fit**, **predict and confusionmat**)

2) Assume you do not know if the four features are correlated and calculate based on multivariate normal distribution. (Hint: use Matlab function **mvnpdf** to calculate the likelihood). (***Note:*** *If you use Python or R, use the corresponding functions to calculate Naïve Bayes and probability of multivariate normal distribution.)*

2. Make a function for multi-class classification using the discriminant function that derived based on the assumption that observation feature vector ***x*** follows multivariate normal distribution. The general case the covariance matrix of each class **Σi = arbitrary**. The classification decision is to determine the class that maximizes the discriminant function :



1. Perform the 3-class classification for the IRIS dataset using the Bayesian classification function made above. Use 5-fold cross-validation, report the prediction accuracy and confusion matrix.
2. Perform the binary classification for the Breast Cancer dataset using the Bayesian classification function. Still do 5-fold cross-validation, report the prediction accuracy and confusion matrix. The data can be downloaded from the UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>.