MTH 443: Lab Problem Set 9

[1] Consider the "Iris"dataset in "iris.csv". The benchmark Iris dataset was introduced by R. A. Fisher in 1936 as an example for discriminant analysis. The data report four characteristics (sepal width, sepal length, pedal width and pedal length) of three species of Iris flower. The data originally appeared in *Fisher, R. A. (1936). "The Use of Multiple Measurements in Axonomic Problems," Annals of Eugenics 7, 179-188*. Split the available data into 2 parts, keeping randomly chosen 10% records as out of sample test data. Build multiclass classification models under the following setups:

(a) linear discriminant scores and quadratic discriminant scores based classifiers. Use equal misclassification costs, equal prior probabilities with multivariate normality assumption.

(b) Bayes classifier using kernel density based estimation of class conditional densities and estimating prior probabilities from learning sample.

Calculate the misclassification rates of the respective classifiers, separately for the learning data and the test set data.

[2] Consider the wine data in "wine_italy.csv". The dataset gives data that are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different wineries. The analysis determined the quantities of 13 constituents found in each of the three types of wines, these are (1) Alcohol, (2) Malic acid, (3) Ash, (4) Alcalinity of ash, (5) Magnesium, (6) Total phenols, (7) Flavanoids, (8) Nonflavanoid phenols, (9) Proanthocyanins, (10) Color intensity, (11) Hue, (12) OD280/OD315 of diluted wines and (13) Proline.

Split the available data into 2 parts, keeping the randomly chosen 10% records as out of sample test data and apply the following classifier approaches to build multiclass classification models:

(a) linear discriminant score and quadratic discriminant score; with estimated prior probability, equal misclassification cost and multivariate normal assumption.

(b) Bayes classifier using kernel density based estimation of class conditional densities and equal prior probabilities assumption.

Calculate the misclassification rates of the respective classifiers, separately for the learning data and the test set data.