IE 5300 Data Mining and Analytics
Homework 2

Classification Using KNN and Decision Tree

1. Make a function for KNN classifier using Minkowski Distance based on two decision options: 1) based on the idea of 'majority vote': the class is determined by the majority class in K nearest neighbors; 2) based on the idea of 'averaged distance' of K-nearest neighbors in each class. In Matlab, the input and output of your function can be as follows:

   Lpred = myknn(Dtrain, Ltrain, Dtest, K, KNNopt, Dorder);

   Where for the function inputs: Dtrain is the training feature matrix, Ltrain is the label vector for the training samples, Dtest is the feature matrix for the testing samples, K is the number of K nearest neighbors to be searched, KNNopt represents the decision making options to be performed for KNN classification (majority voting or averaged distance comparison), Dorder is the order for Minkowski Distance. The output Lpred is the predicted class label vector for the testing samples in Dtest.

2. Prepare N-fold Cross Validation training and testing datasets for the IRIS dataset. Learn and implement the provided two functions to achieve N-fold Cross validation data preparation.

3. Implement the KNN classifier to the IRIS dataset. Perform the classification experiments using 10-fold cross validation with parameter settings: K = 3, 5, 7, 9, KNN options 1 and 2, and Minkowski distance orders of 1, 2, 10. (Using loops to implement different parameter settings and save classification results automatically, do not manually change parameter settings one by one.) Report the classification results table: for each parameter setting, report the mean and std of the 10-fold cross-validation accuracies for each class. Determine which parameter setting generated the highest classification performance?

4. For the IRIS dataset, design a simple decision tree using the discretized attributes of petal width and petal length. Make a scatter plot of the 150 iris samples based on petal width and petal length, you can figure out a simple decision tree roughly based on your observation, draw a flowchart of the decision tree. Make your decision tree into a Matlab function, classify the 150 iris samples (do not need training and testing), and report the classification accuracy and confusion matrix. Calculate the entropy and information gain of each level of decision tree.