IE 6318 Data Mining and Analytics
Homework 2

1. For the IRIS dataset, prepare a training dataset and a testing dataset for classification model training and testing. For each class, take the first 40 samples into the training dataset, and the remaining 10 samples into the testing dataset.

2. Make a KNN classifier for the 3-class classification problem using the distance function you made in HW1 for Minkowski Distance. The KNN function performs classification based on the majority voting of K-nearest neighbors. Implement the KNN classifiers to the IRIS dataset using K = 3, 5, 7 for K-nearest neighbors, and r = 1, 2, 5 for the distance order of Minkowski Distance. For each parameter setting of K and r, perform the classification experiment using the training and testing dataset you made in problem 1.
   1) For each KNN parameter setting, report classification accuracy and the confusion matrix.
   2) Calculate and report the accuracy of each class for each parameter setting.
   3) Assume we use the average accuracy of each class as the overall model performance measure, find the best parameter setting that generates the highest average accuracy for each class.

3. Design a simple decision tree using the attributes of petal width and petal length to classify two types of iris flowers: Versicolor and Virginica, as shown below. Assume the binary decision boundary on Petal Length is 4.8, and the decision boundary on Petal Width is 1.7. Make this simple decision tree into a decision tree classification function. The function should have two inputs: petal width and petal length, and one output: classification outcome Versicolor or Virginica. Implement the decision tree function to classify the 100 iris samples of Versicolor and Virginica, and report the classification accuracy, sensitivity, and specificity. Here define sensitivity = accuracy for class Versicolor, and specificity = accuracy of class Virginica.