Homework #1
CS6052: Intelligent Data Analysis: Fall 2017
Due Date: September 18th, 2017

Consider the dataset of printed letters and their features as given at the UCI Machine Learning repository site (http://archive.ics.uci.edu/ml/datasets/Letter+Recognition). This dataset has 20000 instances and 16 attributes. Perform the following tasks and submit the results/answers for each of the following tasks. With each answer state the toolbox / program that you use for getting the answer. Also state the commands, function calls, and parameter values used for obtaining each answer.

1. Normalize the columns for their values to be in uniform ranges. Describe the process you followed to do the normalization.
2. Split the dataset into three randomly selected parts: 12000 instances for training, 4000 for validation, and 4000 for testing. Describe how you made these partitions.
3. Use Matlab (or other toolboxes) to generate decision trees with the following conditions on the leaf nodes of the resulting decision trees: Each leaf node must have at least (all the following cases) 250, 225, 200, 175, 150, 125, 100, 75, 50, 25, 10, and 5 instances of the dataset. Determine the accuracy for each of these decision trees on the training data and the validation data.
    a. Plot the accuracy values of the trees for all the above cases, and for the training and the validation datasets. How do you interpret the plots?
    b. Which decision tree is the best from among all the above decision trees and why?
    c. Show the selected (best) decision tree in tabular (list of rules) or tree structure.
    d. Use the selected decision tree to compute the Confusion Matrix for the test data set.
    e. Compute Precision, Recall, and F metrics for any randomly selected three of the 26 classes.
4. Consider the Pima Indians Diabetes dataset from the UCI Machine Learning repository. (https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes) This dataset has 768 Instances and 8 attributes. Find the decision tree that has at least 15 instances at each leaf node by doing a 10-fold cross-validation while learning the decision tree.
    a. Show the decision tree in the table/list/tree form.
    b. Find the best possible decision tree by adopting the appropriate parameter values. Show this decision tree. Show the parameters that yield this tree and also list its precision, recall, and accuracy values.