Intelligent Data Analysis: CS6052-001

Homework #1

Due Date: Sept. 12th, 2016, by 10PM.

Consider the Glass Dataset attached with this assignment. The first column is the serial number for instances, the 2nd through the 10th columns contain 9 attribute values for each instance, and the 11th column is the class label for the instance. Using this dataset perform the following tasks and show your work for each task. If you do computations using a short script or program in any toolbox/environment, include the script or program and its output.

All your work and answers must be submitted on the Blackboard in the form of a single pdf file. Please do not submit your answers in any other format. It will be graded only if everything you have submitted exists in the form of a single pdf file.

1. Take the third attribute (column #4 of the original data file). Divide the value range into four intervals of equal width. What are the ranges for each interval?
2. Find the Gini index, information gain, and gain ratio for each of the three splitting points. (Four intervals have three splitting points at their boundaries).
3. Split the value range into four intervals of equal frequency. What are the ranges for each interval?
4. Find the GINI index, information gain, and gain ratio for each of the three splitting points.
5. Consider only the first three attributes. Perform equal width partitions for each of these three attributes. From among the nine splitting points select the best attribute for the decision tree using GINI index as the metric. Now repeat the process for the dataset at each branch of the resulting decision node. This will give you a two level decision tree consisting of three question/test nodes.
6. What is the class purity for each of the four resulting leaf nodes of the decision tree constructed in #5 above?