Homework assignment #2 Due: 10/23

1. Download data set ‘glass.txt’ for discretization. Feature selection will be stopped when the goodness of an attribute subset cannot be improved. Name your program file by your student ID and upload it before 17:00 of the due date.

(a) (30 points) Perform equal-width with ten bins, equal-frequency with ten bins, and the entropy-based discretization (refer to page 123, reference 73) on the continuous attributes. Clearly specify the intervals of a discretized attribute.

(b) (20 points) For each discretization method, perform forward and backward selection to find an attribute subset for classification. Clearly specify the chosen attribute and the goodness of the attribute subset in each iteration.

(c) (10 points) Discuss the results obtained in (b).

2. (10 points) Suppose that the class value for each instance in a data set is either 0 or 1. The score of continuous attribute Xj is defined as , where Dk is the set of training instances with class value k, and I(xij–xmj) = 1 if xij > xmj, or 0 otherwise. Explain why and how the score can be used to rank continuous attributes.

3. (15 points) Use data set ‘lenses.txt’ to grow a decision tree by the gain ratio.