

# CSCI 4380/6380 Data Mining

## Assignment Number 2: Due 9/24/2015 (in class)

1. **[25 points]** Consider the following training set of samples for data mining:

Example	A1	A2	A3	A4	label
1	1	2	2	2	0
2	1	1	1	1	0
3	2	3	2	1	1
4	1	3	3	3	0
5	3	1	2	1	1
6	1	1	1	2	0

The attributes **A1** through **A4** are integers with values in the range [1,2,3] each.

- (a) Give a minimal size (measured by the total number of nodes) decision tree that can correctly classify all the training examples.
  - (b) How would the tree given in Part (a) above classify the following examples: (1,2,2,3) and (3,2,1,1)?
  - (c) Give three association rules consistent with this training set and specify the support and confidence for each rule.
2. **[25 points]** **Short answers please**
- (a) How can a decision tree be converted to a set of rules?
  - (b) Can over-fitting happen even in the absence of noise? Briefly explain.
  - (c) Why are there usually many more association rules that can be inferred from a data-set than there are classification rules?
  - (d) Give one advantage to using the information Gain Ratio measure over the information Gain measure for constructing decision trees.
  - (e) Give one advantage for using two-fold cross-validation over ten-fold cross-validation.
3. **[50 points]** Do exercise 17.2 on page 566 of the text book.