CSCI 4380/6380 Data Mining
First Midterm Exam - Fall 2017
Open Notes

NAME:

Problem(1):

Problem(2):

Problem(3):

Total:

1. [**10 points**] Consider the following training set of samples for data mining:

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

All attributes are binary and **label** is the target attribute.

(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: (1,0,1,1) and (1,1,0,1)?

(c) Give two association rules consistent with this training set, one with confidence equal to 100% and one with confidence less than 100%. Each of the two rules should include at least three attributes.

(d) **For 6900 students only** How would the Naive Bayes method, using the Laplace estimator with $\mu = 1$ and equal prior probabilities, classify the following example: (1,0,1,1)? Would the result be different if you use Naive Bayes without the Laplace estimator? Briefly explain.

2. [**10 points**] **Short answers please**

   (a) Why is the re-substitution error usually a bad predictor of future performance?

   (b) Give **two** advantages to using ball trees over kD-trees for nearest neighbor learning.

   (c) Give **two** major differences between aggregating the input and aggregating the output for multi-instance learning problems.

3. **[10 points] Short answers please**

   (a) Give one advantage to using the 1R classifier over the decision tree classifier.

   (b) Give one advantage to using the decision tree classifier over the 1R classifier.

   (c) How does the Nearest Neighbor classifier deal with missing values?

   (d) The K-nearest Neighbor rule assigns a new instance the most common classification among its K closest instances from the training set. Give one advantage and one disadvantage to using the K-nearest neighbor rule over the (single) Nearest Neighbor rule for classification.