Midterm HST951 - 2002

- This midterm has 4 main sections, each with a number of questions. Including this page, there are 8 numbered pages all in all.
- You have 90 minutes to complete the test.
- Each section has the same weight.
- Some questions may have more than one "right answer".
- Make sure you answer as much as you can without wasting time.
- Use the back of the pages if answers don't fit, and indicate so.
- Remember that the final project is the most important part of this course.
- We will use absolute and not relative performance to determine the grade in this midterm.
- There is no need to remind you about honor code. No group work please.
- Remember to identify yourself on the top of each page.
- Time is monkey…

Good luck!

Student: _____

Section 1: Bayesian systems

1. What is the difference between simple naïve Bayes systems and Bayesian networks?

2. Why there is a need for a "leak" in some Bayesian diagnostic systems?

3. Given three nodes A, B, and C, draw all possible Bayesian networks that model the following relations:

    a. A is independent of B, given C

    b. A is dependent of B, given C

    c. A is independent of B

4. What is the mathematical definition of conditional probability?

Student: _____

Section 2: Evaluation/NN/LR

Examine the predictions of two binary classification systems A and B using variables Z, X, and W, shown below.

Z	X	W	Gold	A	B
0	0	0	0	0.01	0.01
0	0	1	0	0.01	0.6
0	1	0	0	0.01	0.5
0	1	1	1	0.1	0.8
1	0	0	0	0.01	0.01
1	0	1	1	0.1	0.7
1	1	0	1	0.2	0.6
1	1	1	1	0.2	0.8

1. Which system has the highest number of correct classifications? State your assumptions.

2. Which is better calibrated, A or B? (use the median to form 2 groups for the HL test)

3. What is the c-index for A? and for B? What is the area under the ROC for A? and for B?

4. Which is better at discrimination, A or B?

5. Is this a linearly separable problem?

Student: _____

6. What kind of classification models would work well for this problem?

7. Draw a perfect neural network classifier for this problem. Use a step function with a threshold "t" for the output unit. What are the values of the weights and of "t"?

8. What would be a reasonable intercept if this problem were modeled in logistic regression and the coefficients for Z, X, and W were 1, 0, and 1, respectively?

Student: _____

Section 3: General

You are working in a clinical research lab when your boss approaches you with the following problem:

"Our research partners have just sent us a data set collected during a recent study. They wanted to determine which patients respond to a new treatment procedure. Since the treatment is rather expensive, and not all patients respond favorably to it, they could save a lot of money, and improve quality of care, if they just subjected the right patients to the treatment. To determine which patients respond to treatment, they collected 20 pieces of information (variables) about each patient before applying the treatment, and then measured the patient's response to the treatment as a binary outcome (1=respond, 0=did not respond). They want us to help them build a model that allows them to determine whether a patient will respond to treatment based on the 20 measurements alone. Knowing that you covered problems like these in your medical decision support class, I told them I have just the person for the job. Perhaps we could try one of those algorithms I've been hearing so much about lately, logistic regression, rough sets, CART, neural networks or support vector machines?"

Describe how you would go about tackling this problem (assume that you already have all the software you need). In the process, make sure you answer the following questions:

Which of the algorithms do you use, and why?

What are the parameters you need to tune to optimize the algorithm's performance?

How do you tune them?

Student: _____

How do you know that your model generalizes well?

How do you measure the performance of your model?

Can you tell which variables are more important for the classification?

Student: _____

Section 4: Fuzzy/Rough

1) Is a membership function a characteristic function?

2) Let A and B be two subsets of the set U. Indicate whether the following statements are true or false.
   a. A is a subset of A

   b. $A \cap B$ is a subset of A

   c. $\varnothing$ is a subset of A

   d. A is a subset of $A \cap U$

3) Given is the collection C={(a,1), (b,1), (a,1), (c,2), (b,2)}.
   a. Is C a function from {a,b,c} to {1,2}?

   b. Is C a relation from {a,b,c} to {1,2}?

   c. Is C a binary relation?

   d. Find a sub-collection C' of C of maximal size such that C' is a partial function from {a,b,c} to {1,2}.

4) Given the following data table T. Let the column X represent the

ID	A	B	X
1	0	0	1
2	0	1	1
3	1	1	0
4	0	0	1
5	0	1	0
6	1	1	0
7	0	1	0
8	0	0	0
9	0	1	1
10	0	0	1

characteristic function for the set X.

Student: _____

a. Find the upper and lower approximations of X using attributes A and B.

b. What is the resulting boundary region?

c. How many equivalence classes does each of the following sets of attributes induce on the set of rows of T?
   i. {A}
   ii. {B}
   iii. {A, B}

d. Find the membership of elements 1,3, and 7 in X using attributes {A, B}.

e. List all subsets of {A, B} that, when used as columns, preserve the approximations you have found.