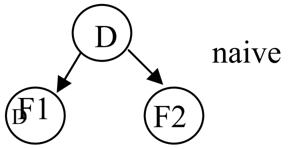
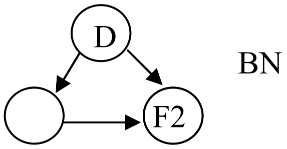
Midterm HST951 solutions

Section 1: Bayesian systems (100 points)

1. (25 points) *What is the difference between simple naïve Bayes systems and Bayesian networks?*

A simple naïve Bayes system makes assumptions of conditional independence. For example, in medical diagnostic systems, it assumes that findings are conditionally independent given diseases. Bayesian networks (which include naïve Bayes systems) may either make those assumptions or not.
A simple way to picture this is:



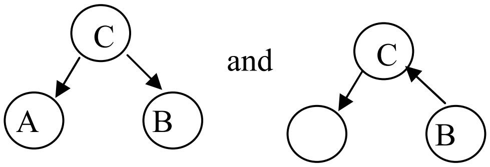


2. (25 points) *Why there is a need for a "leak" in some Bayesian diagnostic systems?*

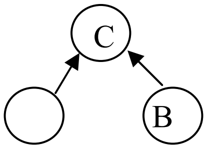
Because not all causes of symptoms can be accounted for, so there is a need to represent the probabilities related to "other" causes.

3. (25 points) *Given three nodes A, B, and C, draw all possible Bayesian networks that model the following relations:*
(I was expecting at least these)

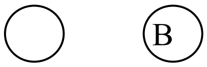
a. (8 points) A is independent of B



b. (8 points) A is dependent of B, given C



c. (8 points) A is independent of B



4. (25 points) *What is the mathematical definition of conditional probability?*

P(A|B) = P(AB)/P(B)

Notice that the definition of conditional independence is something else. I gave some credits for people who misunderstood this question.

Section 2: Evaluation/NN/LR (100 points)

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. (13 points) *Which system has the highest number of correct classifications? State your assumptions.*

A if a good threshold (between 0.01 and 0.1) is used.

2. (13 points) *Which is better calibrated, A or B? (use the median to form 2 groups for the HL test)*

B better calibrated according to HL. Using two groups for the HL test, we come up with expected sums for A: 0.04 and 0.6, with observed sums for the same groups: 0 and 4. For B: 1.12 and 2.9; 1 and 3 (or 0 and 4, depending on whether you used the pair 0.6,0 or 0.6,1 in the first group).
Without further calculation, it is clear that the differences between observed and expected are higher for A than for B (chi-square would be higher, p-value would be lower).

FOR A		FOR B	
0.01	0	0.01	0
0.01	0	0.01	0
0.01	0	0.5	0
0.01	0	0.6	0
0.04	0	1.12	0

0.1	1	0.6	1
0.1	1	0.7	1
0.2	1	0.8	1
0.2	1	0.8	1
0.6	4	2.9	4

3. (13 points) *What is the c-index for A? and for B? What is the area under the ROC for A? and for B?*

16/16 concordant pairs for A
c-index for A is 1 = area under ROC for A

15 concordant, 1 tie for B
c-index for B is (15 + ½)/16 = 0.97 = area under ROC for B

4. (13 points) *Which is better at discrimination, A or B?*

A (c-index is higher).

5. (13 points) *Is this a linearly separable problem?*

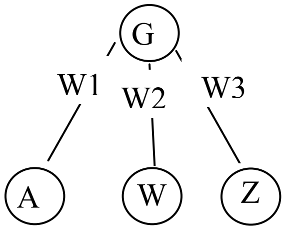
Yes. There is a 2-D plane that can perfectly separate the 3-D data. If any 2 of 3 inputs are "1", then the gold-standard value is "1".

6. (13 points) *What kind of classification models would work well for this problem?*

Almost all of them would work, as the problem is linearly separable.

7. (13 points) *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"?*

t = 1.5
W1 = 1
W2 = 1
W3 = 1



8. (13 points) *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?*

Since the coefficient for X is given and 0, we need to fit the values for Z and W in a logistic model like:

$P = 1/(1 + \exp-(a+Z+W))$

Remember that P in]0,1[for this equation. I will call a value closer to 0 than to 1 as "~0", and a value closer to 1 as "~1" below.

Let's consider the four possible combinations of Z and W values:

Scenario 1: $Z = 0$, $W = 0$
This should result in $P \sim 0$ (or some P lower than scenarios 2 and 3 below), so
$P1 = 1/(1 + \exp-(a))$

Scenario 2:
$Z=0$, $W=1$ or $Z=1$, $W=0$
These should result in $P \sim 0$ if $X=0$, and $P\sim1$ if $X=1$. So P2 value is ½, between P1 in Scenario 1 and P3 in Scenario 3), so

$P2 = 1/(1 + \exp-(a+1))$

Scenario 3:
$Z=1$, $W=1$
This should result in $P\sim1$, so

$P3 = 1/(1 + \exp-(a+2))$

So an intercept "a" such that $P1 < (P2 = 0.5) < P3$ would be reasonable.

For example, a= -1 would work, since it would produce:
$P1 = 1/(1+e)$
$P2 = ½$
$P3 = 1/(1+e^{-1})$

Section 3: General (100 points)

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:

(17 points) *Which of the algorithms do you use, and why?*

Use LR as baseline. Then try other algorithms that handle non-linearly separable problems RS, CART, NN, SVM with non-linear kernel.

(17 points) *What are the parameters you need to tune to optimize the algorithm's performance?*

Depends on the algorithm. Examples of correct answers would be:
Select variables in any algorithm
Modify learning rate in NN
Choose non-linear kernel
Etc.

(17 points) *How do you tune them?*
By having a hold-out set taken out of the training set and monitoring the changes in performance as you tune the parameters.

(17 points) *How do you know that your model generalizes well?*
Test in previously unseen cases. Or at least run cross-validation or bootstrap if set is small.

(17 points) *How do you measure the performance of your model?*
Calibration (e.g., calibration plot, HL) and discrimination (e.g., c-index) indices. Also interpretability, whether it is sensible, etc.

(17 points) *Can you tell which variables are more important for the classification?*
In certain models like LR you can do that based on the odds ratios or standardized coefficients.

Section 4: Fuzzy/Rough (100 points)

1) (4 points*) Is a membership function a characteristic function?*

   No, a characteristic function is a membership function.

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

   a. (1 point) *A is a subset of A*

      True

   b. (1 point) *A∩B is a subset of A*

      True

   c. (1 point) *∅ is a subset of A*

      True

   d. (1 point) *A is a subset of A∩U*

      True

3) *Given is the collection C={(a,1), (b,1), (a,1), (c,2), (b,2)}.*

   a. (2 points) *Is C a function from {a,b,c} to {1,2}?*
      No. Both (b,1) and (b,2) are in C, hence the relation is not single valued.

   b. (2 points) *Is C a relation from {a,b,c} to {1,2}?*
      Yes. C is a subset of {a,b,c} × {1,2}

   c. (2 points) *Is C a binary relation?*
      No, {a,b,c} ≠ {1,2}.

   d. (6 points) *Find a sub-collection C' of C of maximal size such that C' is a partial function from {a,b,c} to {1,2}.*

      A function is a set. The sets of maximal size of elements of C that are partial functions from {a,b,c} to {1,2} are {(a,1), (b,1), (c,2)} and {(a,1), (c,2), (b,2)}.

4) *Given the following data table T. Let the column X represent the characteristic function for the set X.*

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

a. (10 points) *Find the upper and lower approximations of X using attributes A and B.*

The equivalence classes induced by {A, B} are {1, 4, 8, 10}, {2, 5, 7, 9}, and {3, 6}. The underlined elements are in X.

The upper approximation X^U of X using {A, B} is the union of all equivalence classes that contain at least one element from X. Hence, X^U = {1,2,4,5,7,8,9,10}.
The lower approximation X_L of X using {A, B} is the union of all equivalence classes that are subsets of X. Hence, $X_L = \emptyset$.

b. (5 points) *What is the resulting boundary region?*

The boundary region is $X^U - X_L$ = {1,2,4,5,7,8,9,10}.

c. (10 points) *How many equivalence classes does each of the following sets of attributes induce on the set of rows of T?*

i. *{A}*
2

ii. *{B}*
2

iii. *{A, B}*
3

d. (30 points) *Find the membership of elements 1,3, and 7 in X using attributes {A, B}.*

$m_X(1) = |\{\underline{1}, \underline{4}, 8, \underline{10}\} \cap X| / |\{\underline{1}, \underline{4}, 8, \underline{10}\}| = ¾$
$m_X(3) = |\{3, 6\} \cap X| / |\{3, 6\}| = 0$
$m_X(7) = |\{\underline{2}, 5, 7, \underline{9}\} \cap X| / |\{\underline{2}, 5, 7, \underline{9}\}| = ½$

e. (25 points) *List all subsets of {A, B} that, when used as columns, preserve the approximations you have found.*

The approximations we want to preserve are $X^U = \{1,2,4,5,7,8,9,10\}$ and $X_L = \varnothing$. We now investigate all subsets of {A,B} in turn.

{A} induces the following equivalence classes: {1,2,4,5,7,8,9,10} and {3,6}. Resulting in the same upper and lower approximations.

{B} induces the the following equivalence classes: {1, 4 ,8, 10} and {2, 3, 5, 6, 7, 9}. The upper approximation becomes {1,2,3,4,5,6,7,8,9,10} and the lower becomes $\varnothing$.

If we define two elements different if there exists an attribute (in the set of interest) in which they differ, $\varnothing$ induces one equivalence class {1,2,3,4,5,6,7,8,9,10}, resulting in the same approximations as {B}.

Hence the answer is {A, B} and {A}.

Note that even though {A} preserves the approximations, the memberships differ, as we changed the partition of our elements into equivalence classes.