Final

CS440, Fall 2003

This test is closed book, closed notes, no calculators. You have 3:00 hours to answer the questions. If you think a problem is ambiguously stated, state your assumptions and solve the problem under those assumptions. You can use both sides of the test book to write your answers.

Name:	
ID:	

Problem	Score	Max. score
1		21
2		21
3		28
4		16
5		14
Total		100

1 Bayesian networks

Consider the Bayesian network shown in Figure 1. Assume all random variables are binary. Nodes A, B, and C have

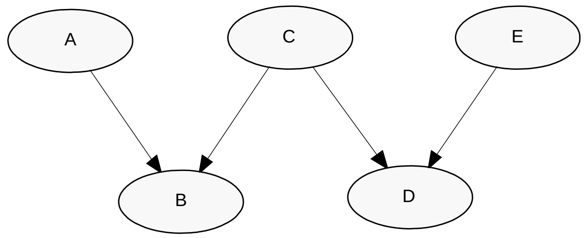

Figure 1: Bayesian network for Problem 1.

the same prior distribution, $P(A = 0) = P(B = 0) = P(C = 0) = 0.7$. Node B has a conditional Bernoulli (binomial) distribution $P(B = i|A, C) = \theta^i_{A,C}(1 - \theta_{A,C})^{1-i}$ with parameters $\theta_{A=j,C=k} = 0.2\ j + 0.4\ k$. Node D also has a conditional Bernoulli distribution but with parameters $\theta_{C=j,E=k} = 0.4\ j + 0.2\ k$.

Note: In the following problems you do not need to compute

1. [3 pts] Write the expression for the joint distribution defined by this network.

2. [3 pts] List the nodes that belong to Markov blankets of nodes A, C and E.

3. [3 pts] What is $P(A = 1, B = 1, C = 1, D = 1, E = 1)$? Show your work.

4. [4 pts] What is $P(A = 1|E = 0)$? Show your work.

5. [3 pts] What is $P(E = 1|D = 1, C = 1)$? Show your work.

6. [5 pts] A child node, F, is added to node B. It has a conditional distribution defined by $P(F = 1|B = 1) = 0.5$ and $P(F = 0|B = 0) = 0.5$. Compute $P(C = 0|F = 1, A = 0)$. Show your work.

2 Dynamic models and statistical learning

John works in a wine cellar where he needs to implement a system for monitoring the levels of sugar in the wine. He purchased two sensors that return three discrete measurement corresponding to low, normal, and high levels of sugar and can be used to detect whether the grape mix is in normal or abnormal condition. However, the sensors are not perfect. The specification lists the following sensor characteristics:

$P(sensor = l \mid grape\_condition)$	$normal$	$abnormal$
low	0.1	0.4
$normal$	0.8	0.1
$high$	0.1	0.5

1. [5 pts] John took a pair of measurements with the two sensors, at five different times. They were $Sensor_1 = \{N, N, N, L, L\}$ and $Sensor_2 = \{N, N, H, L, L\}$. He knew nothing about what condition the mix was in before the measurements were taken. What is his best guess about the state of the mix during the measurements if he assumes that all of the measurements were taken independently? Show the work that justifies your answer.

2. [6 pts] John's boss told him he should not really make a global decision like that. Rather, he should decide the condition of the mix for each pair of measurements with the two sensors (i.e., John would have to make five decisions), after all the measurements were taken. But the boss also realized that the condition of the mix does not change abruptly after each pair of measurements is taken. Since he did not know any better he told John to assume the following sets of probabilities that relate the mix state at two consecutive times:

$$P(\text{normal at } t \mid \text{normal at } t-1) = 0.5, \quad P(\text{normal at } t \mid \text{abnormal at } t-1) = 0.5.$$

What are the five decisions that John would make under these assumptions? Show your work.

3. [5 pts] After seeing the results of John’s work, his boss told him he should come up with better estimates of the transition probabilities. How could John do that?

4. [5 pts] How would John use those new estimates to make better future decisions?

3 Decision making

Consider three ways of computing the final score on an exam that consists of N questions. One way is to average scores of all N questions. Another one is to drop the lowest of N scores and then compute the average of scores of the other $N-1$ questions. Finally, one can assume that one of the questions will be counted towards extra credit and the score will be computed by adding all the question scores and dividing the sum by $N-1$.

All problems are equally hard and it would take you an equal amount of time to solve each of them. The probability of correctly solving each of the N problems is p.

1. [4 pts] Assume the problems are independent and the probability of solving each individual problem does not depend on how many other problems you can solve. What is the probability of correctly solving k out of N problems? How many such events k are there? Write the expression for this probability in terms of p, N and k.

2. [4 pts] Assume the total score for the set of N problems is T. Each problem will be given the following score (reward, utility):

   - Grading scheme 1 (GS1): T/N if you solve it correctly and 0 if you do not.
   - Grading scheme 2 (GS2): $T/(N-1)$ if you solve it correctly, 0 if you do not.
   - Grading scheme 3 (GS3): $T/(N-1)$ if you solve it correctly, 0 if you do not.

   What are the scores (utilities) of the three ways of grading? Explain your work.

3. [6 pts] What are your expected scores under these three grading schemes? Show your work. You may want to use the fact that the expected value of the binomial distribution of L trials with trial probability θ is θL.

4. [3 pts] Which grading scheme one should choose? Justify your answer.

5. [6 pts] What are the maximum differences between the expected scores of grading schemes GS2-GS1 and GS3-GS2 and when do they occur? Write your results in terms of N, p.

6. [5 pts] Analyze how the maximal differences depend on the student's ability to solve the problems (probability p) and the number of problems N.

4 Miscellaneous questions

1. [5 pts] On your way out of the hit feature To Build a Decision Tree, you are surprised to find out the movie theater is giving away prizes. You watch the people ahead of you choose their prize either from behind Door #1 or Door #2. Of those who chose Door #1, half received $6, 1% got a new bike worth $1000, and the rest got a worthless movie poster. Everyone who chose Door #2 got $13.

   Assuming you want to maximize the likely dollar value of your prize, what door should you choose? Why?

2. Consider the joint probability distribution given by the table below

A	B	C	P(A, B, C)
False	False	False	0.05
False	False	True	0.16
False	True	False	0.03
False	True	True	0.25
True	False	False	0.15
True	False	True	0.02
True	True	False	0.11
True	True	True	0.23

- [3 pts] What is $P(A = True)$? Show your work.

- [3 pts] What is $P(B = False|A = True)$? Show your work.

3. [5 pts] Consider this formulation of the N-input, 1-output perceptron learning problem. Assume we want to devise a learning rule that does the following:

   (a) Activation function $g(in)$ is the (hard) threshold function.

   (b) Assume (incorrectly) $g'(in) = 1$, for all in.

   Derive a gradient learning rule that minimizes the sum of square errors $E(w) = 1/2 \sum_{k=1}^{K} (y_k - g(in_k))^2$.

   Discuss how this learning rule updates the network weights and how it is different from the general gradient learning rule with the sigmoid activation function (assume that after we train the network weights using the rule you just derived, we add back the threshold function so that the network output once again becomes 0 or 1.)

5 Important concepts

Briefly describe the following concepts.

1. [2 pts] Classification margin.

2. [2 pts] Utility of money and expected monetary value.

3. [2 pts] Consistent and inconsistent hypotheses.

4. [2 pts] Information gain.

5. [2 pts] Reinforcement learning.

6. [2 pts] Value of perfect information.

7. [2 pts] Completed dataset.