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 the same prior distribution, $P(A=0)=P(C=0)=P(E=0)=0.7$. 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 $\theta_{C=j,E=k}=0.4\,j+0.2\,k$.

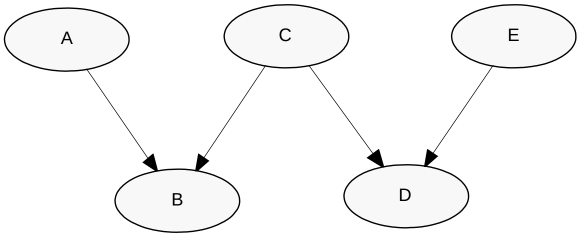

Figure 1: Bayesian network for Problem 1.

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

$$P(A,B,C,D,E)=P(A)P(C)P(E)P(B|A,C)P(D|C,E)$$

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

Node	Markov blanket
A	B, C
C	A, B, D, E
E	C, D

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

$$\begin{aligned} P(A=1,B=1,C=1,D=1,E=1) &= P(A=1)P(C=1)P(E=1) \\ &\quad \times P(B=1|A=1,C=1)P(D=1|C=1,E=1) \\ &= 0.3^3 \times 0.6 \times 0.6. \end{aligned}$$

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

   $P(A=1|E=0)=P(A=1)=0.3$ *because A is independent of E when the other nodes are not instantiated.*

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

   Nodes A and B can be eliminated without affecting the rest of the network. Then

$$\begin{aligned} &P(E=1|D=1,C=1) \\ &= \frac{P(E=1,D=1,C=1)}{P(D=1,C=1)} = \frac{P(E=1,D=1,C=1)}{\sum_E P(E,D=1,C=1)} \\ &= \frac{P(D=1|C=1,E=1)P(E=1)P(C=1)}{P(D=1|C=1,E=1)P(E=1)P(C=1)+P(D=1|C=1,E=0)P(E=0)P(C=1)} \\ &= \frac{0.3^2 0.6}{0.3^2 0.6 + 0.3\;0.7\;0.4} = \frac{0.18}{0.18+0.28} = \frac{9}{25} \end{aligned}$$

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.

Ordinarily, when F is instantiated C would become conditionally dependent on A (converging arcs on B and F an instantiated child node of B). However, because F is completely uninformative (0.5 conditional probabilities for all possible combinations of B and F) it effectively has no influence on the rest of the network. Hence,

$$P(C = 0|F = 1, A = 0) = P(C = 0) = 0.7$$

.

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.

   One can describe the problem using a Naive Bayes model with a root node dedicated to the state of the mix and ten leaf nodes corresponding to the two sets of sensor measurements.

   Prior mix probabilities are uninformative, $P(normal) = P(abnormal) = 0.5$. *Hence,*

$$\begin{aligned} P(normal \mid measurements) &\sim P(measurements \mid normal) P(normal) \\ &= P(Sensor_1 \mid normal) P(Sensor_2 \mid normal) P(normal) \\ &= P(N, N, N, L, L \mid normal) P(N, N, H, L, L \mid normal) P(normal) \\ &= P(N \mid normal)^5 P(L \mid normal)^4 P(H \mid normal)\ P(normal) \\ &= 0.8^5 0.1^4 0.1\ 0.5. \end{aligned}$$

   Similarly,

$$P(abnormal \mid measurements) = 0.1^5\ 0.4^4\ 0.5\ 0.5.$$

   Since

$$\frac{P(normal \mid measurements)}{P(abnormal \mid measurements)} = \frac{0.8^5 0.1^5 0.5}{0.1^5 0.4^4 0.5^2} = \frac{2^4\ 8}{5} > 1$$

   the best guess is the mix is in normal *condition.*

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.

   This situation can be modeled as a hidden Markov model shown in Figure 2. One would make the decisions by finding the sequence of decisions that maximize the posterior probability of decisions, given the measurements

$$\arg \max_{mix_0, \ldots, mix_4} P(mix_0, \ldots, mix_4 \mid measurements).$$

   This can, of course, be done using the Viterbi algorithm.

   However, the fact that the boss gave John essentially no information about how the mix condition changes in time means that, effectively, all the temporal arcs between the nodes (e.g, $mix_0 \rightarrow mix_1$*) can be dropped. Decisions can now be made independently for the five pairs of measurements.*

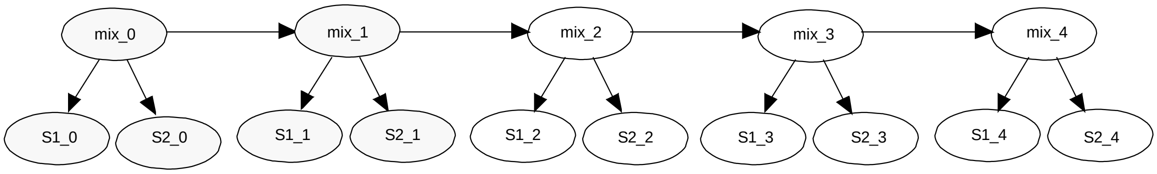

Figure 2: Problem 2.

The decisions for cases where two measurements were N, N are, e.g.,

$$\begin{aligned} P(mix_0 = normal|measurements) &\sim P(N, N|mix_0 = normal)P(mix_0 = normal) \\ &= P(N|mix_0 = normal)P(N|mix_0 = normal)P(mix_0 = normal) \\ &= 0.8^2 0.5 \\ P(mix_0 = abnormal|measurements) &\sim 0.1^2 0.5 < 0.8^2 0.5 \end{aligned}$$

Thus, $mix_0 = mix_1 = normal$.

For mix_2 the decision is made using

$$\begin{aligned} P(mix_2 = normal|measurements) &\sim 0.8\ 0.1\ 0.5 \\ P(mix_2 = abnormal|measurements) &\sim 0.1\ 0.5\ 0.5, \end{aligned}$$

hence, $mix_2 = normal$ is more likely that $mix_2 = abnormal$.

Finally, for the two cases when the measurements are all L, the decision is made by comparing, e.g.,

$$\begin{aligned} P(mix_4 = normal|measurements) &\sim 0.1^2 0.5 \\ P(mix_4 = abnormal|measurements) &\sim 0.4^2 0.5, \end{aligned}$$

and leads to $mix_3 = mix_4 = abnormal$.

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?

Having computed the five decisions

$$mix_0, \ldots, mix_4 = \{normal, normal, normal, abnormal, abnormal\}$$

John can estimate the transition parameters using the maximum likelihood method. Essentially, he should count how frequently one of the four possible transitions occurs in the sequence of those five decisions. That leads to the following transition probability table

$P(mix_t \vert mix_{t-1})$	*normal*	*abnormal*
normal	*2/3*	*0*
abnormal	*1/3*	*1*

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

He could make decisions using the Viterbi algorithm with these new transition probabilities. Then, the decisions across different time instances would depend on each other.

However, one problem would be the 0 probability of the transitions from $abnormal$ to $normal$ which would preclude any return from abnormal to normal state. One way to avid this would be to give some small non-zero probability to this transition, even though it was not seen in the five original samples.

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.

   There are $N+1$ possible events: solve no problems correctly, solve one problem correctly, ..., solve all N problems correctly. The probability of correctly solving k particular problems (and not solving the other $N-k$ problems) is
   $$P(k|particular\_set\_of\_problems) = p^k(1-p)^{N-k}.$$
   There are $\binom{N}{k}$ ways to choose k problems, hence the probability of correctly solving any k problems is
   $$P(k) = \binom{N}{k} p^k (1-p)^{N-k}.$$
   This is a binomial distribution of N trials.

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.

   We need to define scores for each of $N+1$ possible events that can occur in the three grading schemes. The table below shows these scores.

# Problems solved	*Total score* *under GS1 (TS1)*	*Total score* *under GS2 (TS2)*	*Total score* *under GS3 (TS3)*
0	*0*	*0*	*0*
1	$\frac{T}{N}$	$\frac{T}{N-1}$	$\frac{T}{N-1}$
2	$2\frac{T}{N}$	$2\frac{T}{N-1}$	$2\frac{T}{N-1}$
.	.	.	
k	$k\frac{T}{N}$	$k\frac{T}{N-1}$	$k\frac{T}{N-1}$
.	.	.	
N-1	$(N-1)\frac{T}{N}$	$(N-1)\frac{T}{N-1}$	$(N-1)\frac{T}{N-1}$
N	$N\frac{T}{N}$	$(N-1)\frac{T}{N-1}$	$N\frac{T}{N-1}$

   Essentially, we multiply the individual problem score by the number of correctly solved problems, except in the case $k=N$ for GS2 where we drop the score of one of the problems. (We drop one in all other cases as well but this has no effect on the total score because the score of the dropped problem is 0.)

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.

The expected scores (utilities) of the three grading schemes are be computed as

$$\begin{aligned} EU_{GS1} &= \sum_{k=0}^{N} P(k)TS1(k) \\ EU_{GS2} &= \sum_{k=0}^{N} P(k)TS2(k) \\ EU_{GS3} &= \sum_{k=0}^{N} P(k)TS3(k). \end{aligned}$$

To compute these expected scores we need to deal with the quantity

$$\sum_{k=0}^{N} kP(k).$$

This quantity can be interpreted as the average number of problems one can correctly solve. It can be shown that this value is (see the hint)

$$\sum_{k=0}^{N} kP(k) = \sum_{k=0}^{N} k\binom{N}{k}p^k(1-p)^{N-k} = pN.$$

For example, for GS1 the expected score is

$$EU_{GS1} = \sum_{k=0}^{N} p(k)TS1(k) = \sum_{k=0}^{N} P(k)k\frac{T}{N} = \frac{T}{N}\sum_{k=0}^{N} P(k)k = \frac{T}{N}pN = pT.$$

For GS1 and GS3, the expected scores can be interpreted as the average number of problems one can solve correctly multiplied by the score of one problem. For GS2, the expected score is the same as that of GS3 except that it is reduced by the non-zero score of that dropped problem multiplied by the probability that we had to drop one problem with non-zero score, which is $P(N) = p^N$. Hence, the expected scores for the three grading schemes are:

$$\begin{aligned} EU_{GS1} &= Tp \\ EU_{GS2} &= Tp\left[\frac{N}{N-1} - \frac{p^{N-1}}{N-1}\right] \\ EU_{GS3} &= Tp\frac{N}{N-1}. \end{aligned}$$

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

From the solution of Problem 3 it is obvious that $EU_{GS1} \leq EU_{GS2} \leq EU_{GS3}$. One should choose the grading scheme with the highest expected score (maximum expected utility), which is EU_{GS3}. It gives the highest extra score to those who can solve the problems with the highest probability (i.e., good students.)

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.

The difference between EU_{GS2} and EU_{GS1} is

$$\Delta_{21} = EU_{GS2} - EU_{GS1} = Tp\frac{1-p^{N-1}}{N-1}.$$

To find where the maximum occurs (p^*) we can take the first derivative with respect to p and set it to 0:

$$\left.\frac{\partial \Delta_{21}}{\partial p}\right|_{p^*} = T\left(\frac{1}{N-1} - \frac{N}{N-1}{p^*}^{N-1}\right) = 0 \Rightarrow p^* = \frac{1}{N^{\frac{1}{N-1}}}$$

This leads to the maximal value of the difference

$$\Delta^*_{21} = \frac{Tp^*}{N} = \frac{T}{N^{\frac{N}{N-1}}}.$$

To find the maximal difference of expected scores EU_{GS3} and EU_{GS2} we note that

$$\Delta_{32} = EU_{GS3} - EU_{GS2} = T\frac{p^N}{N-1} > 0$$

and is monotonically increasing in p. Hence, the maximal difference occurs at $p = 1$ and is

$$\Delta^*_{32} = \frac{T}{N-1}.$$

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.

   The maximal difference between scores GS2 and GS1 occurs at $p^* = \frac{1}{N^{\frac{1}{N-1}}}$ and is proportional to p^*/N. Hence, the more problems there are the smaller the difference. For a large N, $\Delta^*_{21} \to 0$. Also, as N increases p^* moves from 0 to 1. It starts by benefiting not-so-good students (low p) and ends up benefiting better students (higher p) but with smaller and smaller benefits.

   The maximal difference between expected scores GS3 and GS2 occurs for the best students ($p = 1$) but its benefit quickly approaches 0 as the number of problems increases.

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?

   Expected monetary values of prizes for the two doors are $EMV(D1) = 0.01 \times \$1000 + .50 \times \$6 + 0.49 \times 0 = \$13$ *and* $EMV(D2) = 1.00 \times \$13 = \$13$. *Because* $EMV(D1) = EMV(D2)$ *you can choose either of the two doors.*

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.

$$P(A = True) = \sum_B \sum_C P(A, B, C) = 0.15 + 0.02 + 0.11 + 0.23 = 0.51$$

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

$$\begin{aligned} P(B = false|A = True) &= P(A = True, B = False)/P(A = True) \\ &= \frac{\sum_C P(A = True, B = False, C)}{P(A = True)} \\ &= \frac{0.15 + 0.02}{0.51} \\ &= \frac{0.17}{0.51} = 1/3. \end{aligned}$$

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.)

The perceptron with N inputs (including offset) and one output is defined by the following input-output relationship

$$y = g\left(\sum_{i=1}^{N} w_i x_i\right).$$

The gradient learning rule is

$$w_i^{(l)} = w_i^{(l-1)} - \alpha \frac{\partial E(w)}{\partial w_i}.$$

In this case the gradient $\partial E(w)/\partial w_i$ is

$$\begin{aligned}
\frac{\partial E(w)}{\partial w_i} &= -\sum_{k=1}^{K} (y(k) - g(in_k)) \frac{\partial g(in_k)}{\partial in_k} x_{i,k} \\
&= -\sum_{k=1}^{K} (y(k) - g(in_k))\ 1\ x_{i,k} \\
&= -\sum_{k=1}^{K} err(k) x_{i,k}.
\end{aligned}$$

This is the error correction learning rule: the weights are updated when a sample is incorrectly classified. It differs from the sigmoid gradient rule in that it equally weighs the error of all samples.

5 Important concepts

Briefly describe the following concepts.

1. [2 pts] Classification margin.

   Classification margin is the largest minimal distance between two sets of points separable by a classifier such as a support vector machine.

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

   Expected monetary value is the average monetary value of a lottery whose outcomes can be assigned a monetary value. Utility of money is not directly proportional to its monetary value, hence expected utility of money is not the same as its expected monetary value.

3. [2 pts] Consistent and inconsistent hypotheses.

   Consistent hypothesis is the one which agrees with the data. Inconsistent does not fully agree with the data.

4. [2 pts] Information gain.

   Information gain is the difference in information contents of a set of points before and after a split. More precisely, it is a the difference between the entropy of a labeled set of points and the average entropy of a partition of that set of points.

5. [2 pts] Reinforcement learning.

   The task of reinforcement learning is to use observed rewards to learn an optimal policy (the one that maximizes the total reward) for an environment.

6. [2 pts] Value of perfect information.

   The value of perfect information (VPI) of a piece of evidence E is the difference between the average maximum expected utility computed if E were known and the maximum expected utility with E absent. Non-zero information value implies that one should ask for that piece of evidence E. (However, the VPI is often counterbalanced by the cost of getting that evidence E which is not included in the VPI.)

7. [2 pts] Completed dataset.

   A completed dataset is the dataset whose missing or hidden attributes are "filled-in" using a particular methods. In EM, the attributes are filled-in using all possible values of the missing attributes weighted by their posterior probabilities computed under a current probabilistic model.