CS 188 Spring 2013 Introduction to Artificial Intelligence Midterm II

- You have approximately 1 hour and 50 minutes.
- The exam is closed book, closed notes except a one-page crib sheet.
- Please use non-programmable calculators only.
- Mark your answers ON THE EXAM ITSELF. If you are not sure of your answer you may wish to provide a *brief* explanation. All short answer sections can be successfully answered in a few sentences AT MOST.

First name	
Last name	
SID	
EdX username	

First and last name of student to your left	
First and last name of student to your right	

For staff use only:

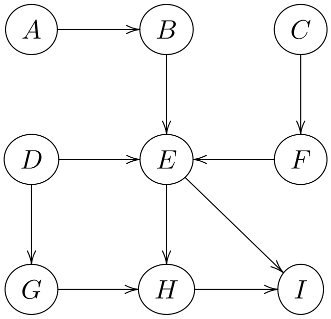
Q1.	Bayes' Nets Representation	/17
Q2.	Bayes' Net Reasoning	/12
Q3.	Variable Elimination	/21
Q4.	Bayes' Net Sampling	/14
Q5.	Probability, Bayes' Nets and Decision Networks	/28
Q6.	Perceptron	/8
	Total	/100

THIS PAGE IS INTENTIONALLY LEFT BLANK

Q1. [17 pts] Bayes' Nets Representation

(a) [8 pts] **Graph Structure: Conditional Independence**

Consider the Bayes' net given below.



Remember that $X \perp\!\!\!\perp Y$ reads as "X is independent of Y given nothing", and $X \perp\!\!\!\perp Y|\{Z, W\}$ reads as "X is independent of Y given Z and W."

For each expression, fill in the corresponding circle to indicate whether it is True or False.

(i) ○True ●False It is guaranteed that $A \perp\!\!\!\perp B$

An active path: $A \rightarrow B$.

(ii) ●True ○False It is guaranteed that $A \perp\!\!\!\perp C$

No active paths.

(iii) ○True ●False It is guaranteed that $A \perp\!\!\!\perp D \mid E$

An active path: $A \rightarrow B \rightarrow E$ (observed) $\leftarrow D$.

(iv) ○True ●False It is guaranteed that $A \perp\!\!\!\perp I \mid E$

An active path: $A \rightarrow B \rightarrow E$(observed) $\leftarrow D \rightarrow G \rightarrow H \rightarrow I$.

(v) ○True ●False It is guaranteed that $B \perp\!\!\!\perp C \mid I$

An active path: $B \rightarrow E$ (descendent I observed) $\leftarrow F \leftarrow C$.

(vi) ○True ●False It is guaranteed that $F \perp\!\!\!\perp A \mid H$

An active path: $F \rightarrow E$ (descendent H observed) $\leftarrow B \leftarrow A$.

(vii) ●True ○False It is guaranteed that $D \perp\!\!\!\perp I \mid \{E, G\}$

No active paths.

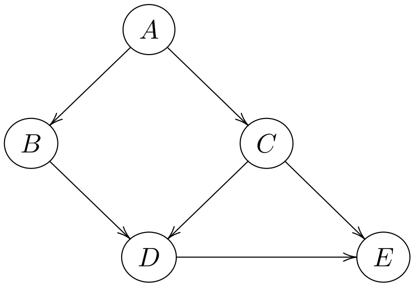
(viii) ○True ●False It is guaranteed that $C \perp\!\!\!\perp H \mid G$

An active path: $C \rightarrow F \rightarrow E \rightarrow H$.

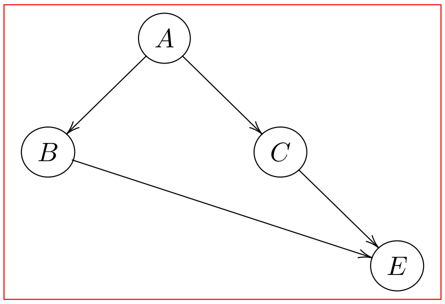
(b) Marginalization and Conditioning

Consider a Bayes' net over the random variables A, B, C, D, E with the structure shown below, with full joint distribution $P(A, B, C, D, E)$.

The following three questions describe different, unrelated situations (your answers to one question should not influence your answer to other questions).

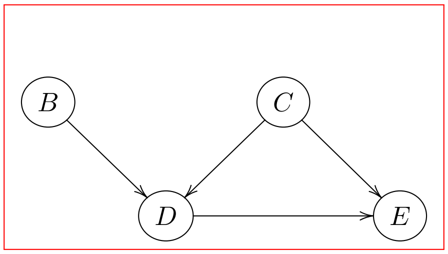


(i) [3 pts] Consider the marginal distribution $P(A, B, C, E) = \sum_d P(A, B, C, d, E)$, where D was eliminated. On the diagram below, draw the minimal number of arrows that results in a Bayes' net structure that is able to represent this marginal distribution. If no arrows are needed write "No arrows needed."



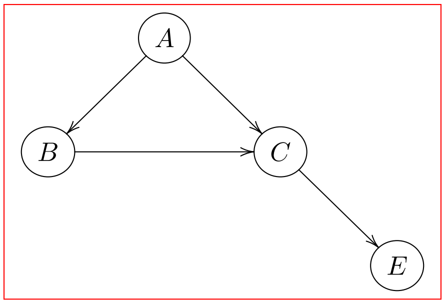
Multiple solutions exist — each solution has exactly the same set of conditional independence assumptions as the graph shown. These other solutions have the same set of edges, but the directionality could be different in such a way that every triple has the same active/inactive properties as above. Concretely, could change directionality of $A - C$, of $A - B$, but not of both $A - C$ and $A - B$ at the same time.

(ii) [3 pts] Assume we are given an observation: $A = a$. On the diagram below, draw the minimal number of arrows that results in a Bayes' net structure that is able to represent the conditional distribution $P(B, C, D, E \mid A = a)$. If no arrows are needed write "No arrows needed."



Only one solution exists for this question. The solution needs to have all edges in the original graph, and then additional edges as needed to ensure the same nodes are connected by active paths (for when A is observed). A observed doesn't activate any paths, in fact all paths through A are inactive with A observed, so no additional edges needed.

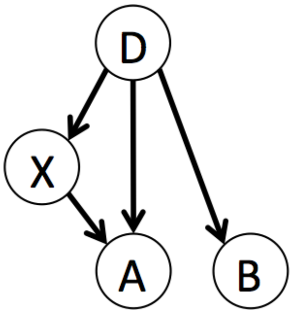
(iii) [3 pts] Assume we are given an observations: $D = d$. On the diagram below, draw the minimal number of arrows that results in a Bayes' net structure that is able to represent the conditional distribution $P(A, B, C, E \mid D = d)$. If no arrows are needed write "No arrows needed."



Multiple solutions exist. The most natural choice is the one shown on the left. Other solutions have the same set of conditional independence assumptions as the graph shown on the left.

Q2. [12 pts] Bayes' Net Reasoning

$P(A\|D,X)$			
$+d$	$+x$	$+a$	0.9
$+d$	$+x$	$-a$	0.1
$+d$	$-x$	$+a$	0.8
$+d$	$-x$	$-a$	0.2
$-d$	$+x$	$+a$	0.6
$-d$	$+x$	$-a$	0.4
$-d$	$-x$	$+a$	0.1
$-d$	$-x$	$-a$	0.9



$P(D)$	
$+d$	0.1
$-d$	0.9

$P(X\|D)$		
$+d$	$+x$	0.7
$+d$	$-x$	0.3
$-d$	$+x$	0.8
$-d$	$-x$	0.2

$P(B\|D)$		
$+d$	$+b$	0.7
$+d$	$-b$	0.3
$-d$	$+b$	0.5
$-d$	$-b$	0.5

(a) [3 pts] What is the probability of having disease D and getting a positive result on test A?

$P(+d,+a) =$

$\sum_x P(+d,x,+a) = \sum_x P(+a|+d,x)P(x|+d)P(+d) = P(+d)\sum_x P(+a|+d,x)P(x|+d) = (0.1)((0.9)(0.7)+(0.8)(0.3)) = 0.087$

(b) [3 pts] What is the probability of *not* having disease D and getting a positive result on test A?

$P(-d,+a) = \sum_x P(-d,x,+a) = \sum_x P(+a|-d,x)P(x|-d)P(-d) = P(-d)\sum_x P(+a|-d,x)P(x|-d) = (0.9)((0.6)(0.8)+(0.1)(0.2)) = 0.45$

(c) [3 pts] What is the probability of having disease D given a positive result on test A?

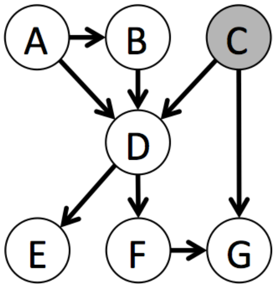
$P(+d|+a) = \frac{P(+a,+d)}{P(+a)} = \frac{P(+a,+d)}{\sum_d P(+a,d)} = \frac{0.087}{0.087+0.45} \approx 0.162$

(d) [3 pts] What is the probability of having disease D given a positive result on test B?

$P(+d|+b) = \frac{P(+b|+d)P(+d)}{P(+b)} = \frac{P(+b|+d)P(+d)}{\sum_d P(+b|d)P(d)} = \frac{(0.7)(0.1)}{(0.7)(0.1)+(0.5)(0.9)} \approx 0.135$

Q3. [21 pts] Variable Elimination

(a) [9 pts] For the Bayes' net below, we are given the query $P(A, E \mid +c)$. All variables have binary domains. Assume we run variable elimination to compute the answer to this query, with the following variable elimination ordering: B, D, G, F.



Complete the following description of the factors generated in this process:

After inserting evidence, we have the following factors to start out with:

$$P(A), P(B|A), P(+c), P(D|A, B, +c), P(E|D), P(F|D), P(G| + c, F)$$

When eliminating B we generate a new factor f_1 as follows:

$$f_1(A, +c, D) = \sum_b P(b|A)P(D|A, b, +c)$$

This leaves us with the factors:

$$P(A), P(+c), P(E|D), P(F|D), P(G| + c, F), f_1(A, +c, D)$$

When eliminating D we generate a new factor f_2 as follows:

$$f_2(A, +c, E, F) = \sum_d P(E|d)P(F|d)f_1(A, +c, d)$$

This leaves us with the factors:

$$P(A), P(+c), P(G| + c, F), f_2(A, +c, E, F)$$

When eliminating G we generate a new factor f_3 as follows:

$$f_3(+c, F) = \sum_g P(g| + c, F)$$

This leaves us with the factors:

$$P(A), P(+c), f_2(A, +c, E, F), f_3(+c, F)$$

Let's make sure to account for error propagation in our grading of this one.

When eliminating F we generate a new factor f_4 as follows:

$$f_4(A, +c, E) = \sum_f f_2(A, +c, E, f) f_3(+c, f)$$

This leaves us with the factors:

$$P(A), P(+c), f_4(A, +c, E)$$

(b) [2 pts] Write a formula to compute $P(A, E \mid +c)$ from the remaining factors.

$P(A, E \mid +c) = \frac{P(A)P(+c)f_4(A,+c,E)}{\sum_{a,e} P(a)P(+c)f_4(a,+c,e)}$ or alternatively: $P(A, E \mid +c) \propto P(A)P(+c)f_4(A, +c, E)$ and include statement that says renormalization is needed to obtain $P(A, E \mid +c)$.

(c) [2 pts] Among f_1, f_2, f_3, f_4, which is the largest factor generated, and how large is it? Assume all variables have binary domains and measure the size of each factor by the number of rows in the table that would represent the factor.

$f_2(A, +c, E, F)$ is the largest factor generated. It has 3 non-instantiated variables, hence $2^3 = 8$ entries.

(d) [8 pts] Find a variable elimination ordering for the same query, i.e., for $P(A, E \mid +c)$, for which the maximum size factor generated along the way is smallest. Hint: the maximum size factor generated in your solution should have only 2 variables, for a size of $2^2 = 4$ table. Fill in the variable elimination ordering and the factors generated into the table below.

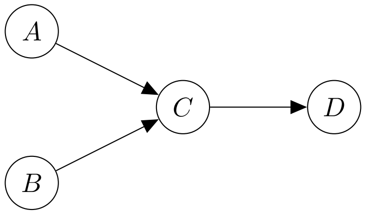
Variable Eliminated	Factor Generated
B	$f_1(A, +c, D)$
G	$f_2(+c, F)$
F	$f_3(+c, D)$
D	$f_4(A, +c, E)$

For example, in the naive ordering we used earlier, the first row in this table would have had the following two entries: B, $f_1(A, +c, D)$.

Note: multiple orderings are possible. An ordering is good if it eliminates all non-query variables (B, D, F, G) and its largest factor has only two variables.

Q4. [14 pts] Bayes' Net Sampling

Assume you are given the following Bayes' net and the corresponding distributions over the variables in the Bayes' net.



$P(A)$	
+a	0.1
-a	0.9

$P(B)$	
+b	.7
-b	.3

$P(C\|A,B)$			
+c	+a	+b	.25
-c	+a	+b	.75
+c	-a	+b	.6
-c	-a	+b	.4
+c	+a	-b	.5
-c	+a	-b	.5
+c	-a	-b	.2
-c	-a	-b	.8

$P(D\|C)$		
+d	+c	.5
-d	+c	.5
+d	-c	.8
-d	-c	.2

(a) [2 pts] Assume we receive evidence that $A = +a$. If we were to draw samples using rejection sampling, on expectation what percentage of the samples will be **rejected**?

Since $P(+a) = \frac{1}{10}$, we would expect that only 10% of the samples could be saved. Therefore, expected 90% of the samples will be rejected.

(b) [6 pts] Next, assume we observed both $A = +a$ and $D = +d$. What are the weights for the following samples under likelihood weighting sampling?

Sample	Weight
$(+a, -b, +c, +d)$	$P(+a) \cdot P(+d\| + c) = 0.1 * 0.5 = 0.05$
$(+a, -b, -c, +d)$	$P(+a) \cdot P(+d\| - c) = 0.1 * 0.8 = 0.08$
$(+a, +b, -c, +d)$	$P(+a) \cdot P(+d\| - c) = 0.1 * 0.8 = 0.08$

(c) [2 pts] Given the samples in the previous question, estimate $P(-b| + a, +d)$.

$$P(-b| + a, +d) = \frac{P(+a) \cdot P(+d| + c) + P(+a) \cdot P(+d| - c)}{P(+a) \cdot P(+d| + c) + 2 \cdot P(+a) \cdot P(+d| - c)} = \frac{0.05 + 0.08}{0.05 + 2 \cdot 0.08} = \frac{13}{21}$$

(d) [4 pts] Assume we need to (approximately) answer two different inference queries for this graph: $P(C| + a)$ and $P(C| + d)$. You are required to answer one query using likelihood weighting and one query using Gibbs sampling. In each case you can only collect a relatively small amount of samples, so for maximal accuracy you need to make sure you cleverly assign algorithm to query based on how well the algorithm fits the query. Which query would you answer with each algorithm?

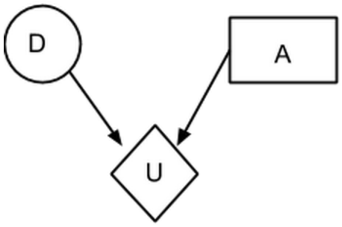
Algorithm	Query
Likelihood Weighting	$P(C\| + a)$

Algorithm	Query
Gibbs Sampling	$P(C\| + d)$

Justify your answer:
You should use Gibbs sampling to find the query answer $P(C| + d)$. This is because likelihood weighting only takes upstream evidence into account when sampling. Therefore, Gibbs, which utilizes both upstream and downstream evidence, is more suited to the query $P(C| + d)$ which has downstream evidence.

Q5. [28 pts] Probability, Bayes' Nets and Decision Networks

It is Monday night, and Bob is finishing up preparing for the CS188 Midterm II that is coming up on Tuesday. Bob has already mastered all the topics except one: Decision Networks. He is contemplating whether to spend the remainder of his evening reviewing that topic (*review*), or just go to sleep (*sleep*). Decision Networks are either going to be on the test ($+d$) or not be on the test ($-d$). His utility of satisfaction is only affected by these two variables as shown below:



D	P(D)
+d	0.5
-d	0.5

D	A	U(D,A)
+d	*review*	1000
-d	*review*	600
+d	*sleep*	0
-d	*sleep*	1500

(a) [5 pts] **Maximum Expected Utility**

Compute the following quantities:

$EU(review) = P(+d)U(+d, review) + P(-d)U(-d, review) = 0.5 * 1000 + 0.5 * 600 = 800$

$EU(sleep) = P(+d)U(+d, sleep) + P(-d)U(-d, sleep) = 0.5 * 0 + 0.5 * 1500 = 750$

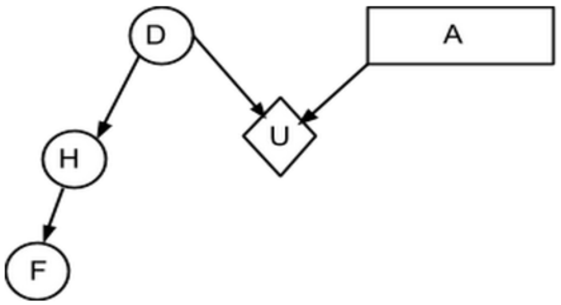
$MEU(\{\}) = max(800, 750) = 800$

Action that achieves $MEU(\{\}) = review$

This result notwithstanding, you should get some sleep.

(b) [11 pts] **The TA is on Facebook**

The TAs happiness (H) is affected by whether decision networks are going to be on the exam. The happiness (H) determines whether the TA posts on Facebook ($+f$) or doesn't post on Facebook ($-f$). The prior on D and utility tables remain unchanged.



F	H	$P(F\|H)$
+f	+h	0.6
-f	+h	0.4
+f	-h	0.2
-f	-h	0.8

D	P(D)
+d	0.5
-d	0.5

H	D	$P(H\|D)$
+h	+d	0.95
-h	+d	0.05
+h	-d	0.25
-h	-d	0.75

D	A	U(D,A)
+d	*review*	1000
-d	*review*	600
+d	*sleep*	0
-d	*sleep*	1500

Decision network.

Tables that define the model are shown above.

H	$P(H)$
+h	0.6
-h	0.4

F	$P(F)$
+f	0.44
-f	0.56

D	F	$P(D\|F)$
+d	+f	0.666
-d	+f	0.334
+d	-f	0.370
-d	-f	0.630

F	D	$P(F\|D)$
+f	+d	0.586
-f	+d	0.414
+f	-d	0.300
-f	-d	0.700

D	H	$P(D\|H)$
+d	+h	0.79
-d	+h	0.21
+d	-h	0.06
-d	-h	0.94

Tables computed from the first set of tables. Some of them might be convenient to answer the questions below.

Compute the following quantities:

$EU(review|+f) = P(+d|+f)U(+d, review) + P(-d|+f)U(-d, review) = 0.666 * 1000 + 0.334 * 600 = 666 + 200.4 = 866.4$

$EU(sleep|+f) = P(+d|+f)U(+d, sleep) + P(-d|+f)U(-d, sleep) = 0.666 * 0 + 0.334 * 1500 = 501$

$MEU(\{+f\}) = \max(866.4, 501) = 866.4$

Optimal Action$(\{+f\}) =$ *review*

$EU(review|-f) = P(+d|-f)U(+d, review) + P(-d|-f)U(-d, review) = 0.370 * 1000 + 0.630 * 600 = 370 + 378 = 748$

$EU(sleep|-f) = P(+d|-f)U(+d, sleep) + P(-d|-f)U(-d, sleep) = 0.370 * 0 + 0.630 * 1500 = 0 + 945 = 945$

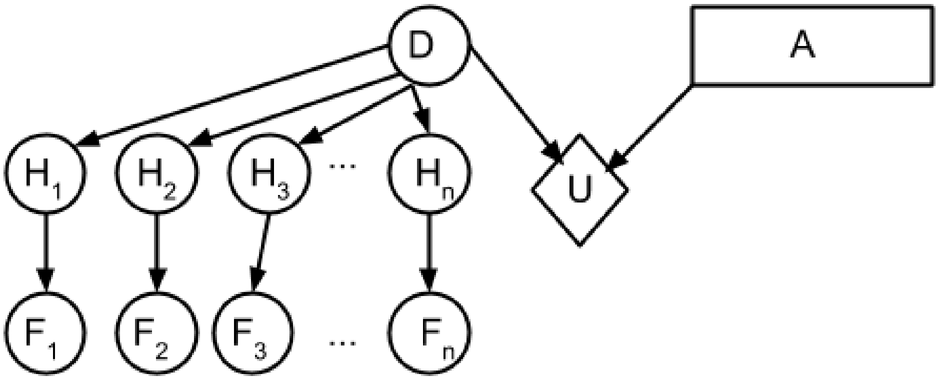
$MEU(\{-f\}) = \max(748, 945) = 945$

Optimal Action$(\{-f\}) =$ *sleep*

$VPI(\{F\}) = P(+f)MEU(\{+f\}) + P(-f)MEU(\{-f\}) - MEU(\{\}) = 0.44 * 866.4 + 0.56 * 945 - 800 = 110.416$

(c) VPI Comparisons

Now consider the case where there are n TAs. Each TA follows the same probabilistic models for happiness (H) and posting on Facebook (F) as in the previous question.



(i) [3 pts] ○True ●False $VPI(H_1|F_1) = 0$

Justify: F_1 is just a noisy version of H_1. Hence finding out H_1 gives us more information about D even when we have already observed F_1. This in turn will allow us to more often make the right decision between *sleep* and *review*.

(ii) [3 pts] ●True ○False $VPI(F_1|H_1) = 0$

Justify:The parent variable of the utility node, D, is conditionally independent of F_1 given H_1.

(iii) [3 pts] ○True ●False $VPI(F_3|F_2, F_1) > VPI(F_2|F_1)$

Justify:The F_i variables give us noisy information about D. The more F_i variables we get to observe, the better chance we end up being able to make the right decision. The more F_i variables we have already observed, however, the less an additiona observation of a new variable F_j will influence the distribution of D.

(iv) [3 pts] ●True ○False $VPI(F_1, F_2, \ldots, F_n) < VPI(H_1, H_2, \ldots, H_n)$

Justify:The F_i variables are noisy versions of the H_i variables, hence observing the H_i variables is more valuable.

Q6. [8 pts] Perceptron

You have decided to become a teacher. The only issue is that you don't want to spend lots of time grading essays, so instead you decide to grade them all with a linear classifier. Your classifier considers the number of 7-letter (f_7) and 8-letter words (f_8) in an essay and then assigns a grade, either A or F, based on those two numbers. You have four graded essays to learn from:

$BIAS$	f_7	f_8	grade
1	2	1	A (+)
1	0	2	F (-)
1	1	2	A (+)
1	1	0	F (-)

(a) [2 pts] You decide to run perceptron and being optimistic about the students essay writing capabilities, you decide to initialize your weight vector as $(1, 0, 0)$. If the score from your classifier is greater than 0, it gives an A, if it is 0 or lower, it gives an F. Fill in the resulting weight vector after having seen the first training example and after having seen the second training example.

	BIAS	f_7	f_8
Initial	1	0	0
After first training example	1	0	0
After second training example	0	0	-2

Use the perceptron update rule.

(b) [2 pts] ● True ○ False The training data is linearly separable with the given features.

Justify: One justification is to draw the points in the 2-D plane, and show that a linear decision boundary separates the classes. Another justification is to provide a weight vector w that classifies all data points correctly, $w = (-2.5, 1, 1)$ is such a weight vector.

(c) [4 pts] For each of the following decision rules, indicate whether there is a weight vector that represents the decision rule. If "Yes" then include such a weight vector.

1. A paper gets an A if and only if it satisfies $(f_7 + f_8 \geq 7)$.

   ● Yes $w =$ $(-6.5, 1, 1)$ ○ No

2. A paper gets an A if and only if it satisfies $(f_7 \geq 5\ AND\ f_8 \geq 4)$.

   ○ Yes $w =$ ● No

3. A paper gets an A if and only if it satisfies $(f_7 \geq 5\ OR\ f_8 \geq 4)$.

   ○ Yes $w =$ ● No

4. A paper gets an A if and only if it has between 4 and 6, inclusive, 7-letter words and between 3 and 5 8-letter words.

   ○ Yes $w =$ ● No