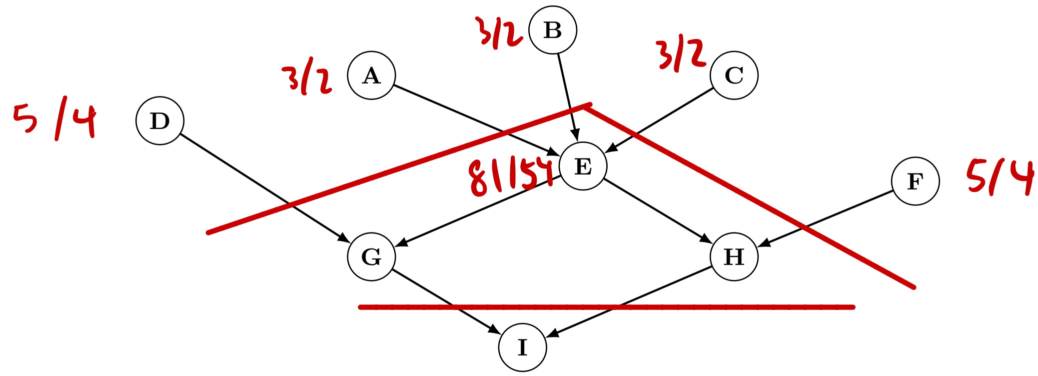
01.112 Machine Learning, Fall 2017
Homework 5

Due Monday 4 Dec 2017, 11.59pm

Sample Solutions

In this homework, we would like to look at the Bayesian Networks. You are given a Bayesian network as below. All nodes can take 2 different values: $\{1, 2\}$.



1. (10 pts) Without knowing the actual value of any node, are node **A** and **F** independent of each other? What if we know the value of node **C** and **I**?

   Answer Without knowing the actual value of any node, node A and F are independent of each other. This is because there does not exist any path from A to F that is open. Based on the Bayes' ball algorithm, A and F are independent of each other.

   If we know the value of node C and I, then the two variables A and F become dependent. This is because there exist a path connecting A and F that is open: $A-E-H-I-H-F$ or $A-E-G-I-H-F$. Based on the Bayes' ball algorithm, A and F are dependent.

2. (10 pts) What is the *effective* number of parameters needed to for this Bayesian network? What would be the number of parameters for the same network if node **D** and **F** can take 5 different values: $\{1, 2, 3, 4, 5\}$, and all other nodes can only take 3 different values: $\{1, 2, 3\}$?

   Answer The number of parameters correspond to the number of entries in the probability table of each node in the Bayesian network. Assume the number of values for node k to take is r_k. For a node i with parents pa_i, the number of rows is $\prod_{j \in pa_i} r_j$. The number of columns is r_i. However the values in the last column can be uniquely determined from the other columns since the values of each row sum to 1. This means for the node i there are $(r_i - 1)\prod_{j \in pa_i} r_j$ free/independet/effective parameters involved.

   Therefore in the initial Bayesian network, the number of free parameters is:

   $$1(A)+1(B)+1(C)+1(D)+2\times2\times2\times1(E)+1(F)+2\times2\times1(G)+2\times2\times1(H)+2\times2\times1(I) = 25$$

If node D and F can take 5 different values: 1, 2, 3, 4, 5, and all other nodes can only take 3 different values: 1, 2, 3, the number of free parameters is:

parameters from D   parameters from F

$$2(A)+2(B)+2(C)+4(D)+3\times3\times3\times2(E)+4(F)+3\times5\times2(G)+3\times5\times2(H)+3\times3\times2(I) = 146$$

parameters from E

3. (10 pts) If we have the following probability tables for the nodes. Compute the following probability. Clearly write down all the necessary steps.

$$P(\mathbf{E}=2|\mathbf{C}=1)$$

A 1	**A** 2
0.2	0.8

B 1	**B** 2
0.7	0.3

C 1	**C** 2
0.2	0.8

D 1	**D** 2
0.5	0.5

A	**B**	**C**	**E** 1	**E** 2
1	1	1	0.1	0.9
1	1	2	0.3	0.7
1	2	1	0.5	0.5
1	2	2	0.1	0.9
2	1	1	0.9	0.1
2	1	2	0.4	0.6
2	2	1	0.5	0.5
2	2	2	0.4	0.6

F 1	**F** 2
0.3	0.7

D	**E**	**G** 1	**G** 2
1	1	0.1	0.9
1	2	0.6	0.4
2	1	0.6	0.4
2	2	0.5	0.5

E	**F**	**H** 1	**H** 2
1	1	0.1	0.9
1	2	0.4	0.6
2	1	0.5	0.5
2	2	0.5	0.5

G	**H**	**I** 1	**I** 2
1	1	0.1	0.9
1	2	0.9	0.1
2	1	0.7	0.3
2	2	0.9	0.1

Answer One standard approach is to start by computing the following marginal probability:

$$P(C,E) = \sum_{A,B,D,F,G,H,I} P(A)P(B)P(C)P(D)P(E|A,B,C)P(F)P(G|D,E)P(H|E,F)P(I|G,H)$$

Simplify the above expression, and next compute $P(C=1,E=1)$ and $P(C=1,E=2)$ respectively, and then compute $P(C=1) = P(C=1,E=1) + P(C=1,E=2)$. The conditional probability $P(E=2|C=1) = P(C=1,E=2)/P(C=1)$.

Here we describe an alternative approach based on some observations about the independence properties of the graph.

Note that given E, the variables A,B,C and D,F,G,H,I are conditionally independent. Mathematically, this means:

$$P(D,F,G,H,I|A,B,C,E) = P(D,F,G,H,I|E)$$

Now, mathematically, we always have the following:

$$P(A,B,C,D,E,F,G,H,I) = P(A,B,C,E)P(D,F,G,H,I|A,B,C,E)$$

Based on the earlier equation, we have:

$$P(A,B,C,D,E,F,G,H,I) = P(A,B,C,E)P(D,F,G,H,I|E)$$

This yeids:

$$\begin{aligned}
P(C,E) &= \sum_{ABDFGHI} P(A,B,C,D,E,F,G,H,I) \\
&= \sum_{ABDFGHI} P(A,B,C,E)P(D,F,G,H,I|E) \\
&= \sum_{AB} P(A,B,C,E) \sum_{DFGHI} P(D,F,G,H,I|E) \\
&= \sum_{AB} P(A,B,C,E) \\
&= \sum_{AB} P(A)P(B)P(C)P(E|A,B,C) \\
&= P(C) \sum_{AB} P(A)P(B)P(E|A,B,C)
\end{aligned}$$

$$P(E|C) = \frac{P(C,E)}{P(C)} = \sum_{AB} P(A)P(B)P(E|A,B,C)$$

$$\begin{aligned}
P(E=2|C=1) = P(A=1)P(B=1)P(E=2|A=1,B=1,C=1) \\
+P(A=1)P(B=2)P(E=2|A=1,B=2,C=1) \\
+P(A=2)P(B=1)P(E=2|A=2,B=1,C=1) \\
+P(A=2)P(B=2)P(E=2|A=2,B=2,C=1) \\
= 0.2\times 0.7\times 0.9 + 0.2\times 0.3\times 0.5 + 0.8\times 0.7\times 0.1 + 0.8\times 0.3\times 0.5 \\
= 0.126 + 0.03 + 0.056 + 0.12 = 0.332
\end{aligned}$$

4. (10 pts) Now, assume we do not have any knowledge about the probability table for the nodes in the network, but we have the following 12 observations. Find a way to estimate the probability table associated with the nodes **A** and **H**.

A	B	C	D	E	F	G	H	I
1	1	2	2	2	1	1	1	1
1	2	1	1	2	1	1	1	2
2	2	2	1	2	2	1	2	1
1	1	2	1	2	1	1	2	2
1	2	1	1	1	1	2	1	1
2	2	1	2	1	2	2	1	2
2	1	2	2	1	2	2	2	1
2	2	2	1	2	1	2	2	2
1	1	1	1	2	2	1	1	1
1	1	1	1	2	1	1	1	2
1	2	1	2	2	1	2	1	2
2	2	1	2	1	2	2	1	1

		I	
G	H	1	2
1	1	0.1	0.9
1	2	0.9	0.1
2	1	0.7	0.3
2	2	0.9	0.1

Answer We can use the maximum likelihood estimation to find the optimal model parameters.

$$\theta_A(1) = \frac{\text{Count}(A=1)}{\text{Count}(A)} = 7/12$$

$$\theta_A(2) = \frac{\text{Count}(A=2)}{\text{Count}(A)} = 5/12$$

$$\theta_H(1) = \frac{\text{Count}(E=1,F=1,H=1)}{\text{Count}(E=1,F=1)} = 1$$

1/1

$$\theta_H(2) = \frac{\text{Count}(E=1,F=1,H=2)}{\text{Count}(E=1,F=1)} = 0$$

0/1

$$\theta_H(1) = \frac{\text{Count}(E=1,F=2,H=1)}{\text{Count}(E=1,F=2)} = 2/3$$

$$\theta_H(2) = \frac{\text{Count}(E=1,F=2,H=2)}{\text{Count}(E=1,F=2)} = 1/3$$

$$\theta_H(1) = \frac{\text{Count}(E=2,F=1,H=1)}{\text{Count}(E=2,F=1)} = 2/3$$

4/6

$$\theta_H(2) = \frac{\text{Count}(E=2,F=1,H=2)}{\text{Count}(E=2,F=1)} = 1/3$$

2/6

$$\theta_H(1) = \frac{\text{Count}(E=2,F=2,H=1)}{\text{Count}(E=2,F=2)} = 1/2$$

$$\theta_H(2) = \frac{\text{Count}(E=2,F=2,H=2)}{\text{Count}(E=2,F=2)} = 1/2 \tag{1}$$

The resulting probability tables for A and H are:

	A	
	1	2
	7/12	5/12

E	F	H = 1	H = 2
1	1	1	0
1	2	2/3	1/3
2	1	2/3	1/3
2	2	1/2	1/2

5. (20 pts) Based on the above observations, you would like to find a good Bayesian network structure to model the data. You started with the initial structure shown on the previous page, and decided to delete the edge between **H** and **I**. Is the resulting new structure (after deleting the single edge between **H** and **I** from the original graph) better than the original structure in terms of BIC score? Clearly explain the reason. (Hint: Try to find a short answer.)

   Answer Deletion of the edge between **H** and **I** will only change the probability table of the node **I**. Now let's see what happens to the probability table of node **I**.

   Before deletion:

G	H	I = 1	I = 2
1	1	$1/2$	$1/2$
1	2	$1/2$	$1/2$
2	1	$1/2$	$1/2$
2	2	$1/2$	$1/2$

   After deletion:

G	I = 1	I = 2
1	$1/2$	$1/2$
2	$1/2$	$1/2$

   This means deleting this edge does not affect the Bayesian network's log-likelihood (when the model parameters are estimated using MLE from the data). However, the BIC scores for the two networks are different now. Specifically the new Bayesian network needs 2 less free parameters. Therefore the resulting new structure has a better (higher) BIC score.