University of TEHRAN

Probabilistic Graphical Models in Bioinformatics

Final Exam

June 19, 2019

Time limit for problems 1-4: 8:00-9:30.

Time limit for problems 5-6: 9:45-12:00.

Number of pages: 9

Total point: 100

A two-sided cheat sheet is allowed.

You can access to the slides and notes for half an hour in the second section of the exam (10:30-11:00).

Question: Have you had your cheat sheet signed by your instructor or TA? Have you written your initials on every page of the exam sheet? (1 point)

a) Yes b) No

Please leave the below table empty.

PROBLEM	MAXIMUM POINTS	OBTAINED POINTS
1	10	
2	17	
3	12	
4	17	
5	29	
6	15	
TOTAL	100	

1. Short Questions (10 points)

Indicate true or false for the following statements.

i. A Bayesian network is necessarily a tree. (True/False)

ii. BIC scores of two I-equivalent Bayesian networks are the same. (True/False)

iii. The correctness of the variable elimination depends on the order. (True/False)

iv. In every Bayesian network there is at least a node without any parent. (True/False)

v. Forward algorithm is used to perform MAP-inference on hidden Markov models. (True/False)

vi. Complete-data likelihood is often a unimodal function. (True/False)

vii. Marginal likelihood avoids overfitting because it is not sensitive to a particular choice of parameters. (True/False)

viii. Missing values in the data might introduce new dependencies between parameters of a Bayesian network. (True/False)

ix. An EM iteration is guaranteed to increase the observed data likelihood function. (True/False)

x. Consider a joint distribution over the parameter θ and the data $x[1], \dots, x[M]$. Bayesian prediction refers to as predicting next observation X[M+1] given the data and parameter, i.e., $P(X[M+1] \mid x[1], \dots, x[M], \theta)$. (True/False)

2. Factorization and Independence in Bayesian networks (17 points)

In this exercise, if needed assume all CPDs are positive.

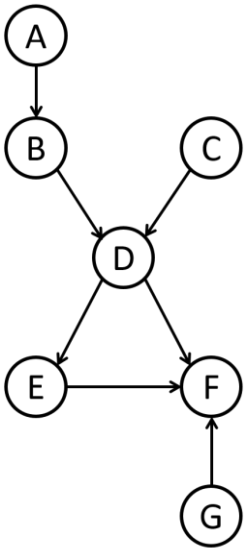


Figure 1

i. (1 point) Write the factorized form of the joint distribution of the Bayesian network in Figure 1 over all variables.

ii. (6 points) Assess the following independence statements. If a statement is true, give the blocking node otherwise give an active trail.

a. $A \perp C$

b. $B \perp C \mid E$

c. $B \perp G \mid E, F$

iii. (3 points) Apply the chain rule for $P(C, G \mid B, D)$ and simplify the result as much as you can.

iv. (1 point) Write the Markov blanket of node D.

v. (2 points) Find all I-equivalent networks for the network in Figure 1.

vi. (3 points) Consider the skeleton of the Bayesian network in Figure 1. For the obtained Markov network, write the factorized form of the joint distribution over all the variables, assuming the model is parameterized by one factor over each node and one factor over each edge in the graph.

vii. (1 point) Write the Markov blanket of node D for the obtained Markov network.

3. Inference questions (12 points)

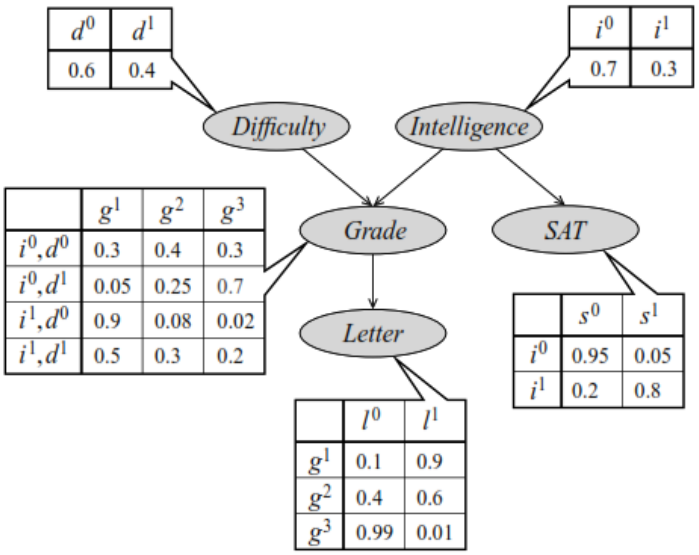
Compute the following probabilities for the below Bayesian network. You need to give a simplified solution.

i. (3 points) $P(G = g^1 \mid D = d^1)$

ii. (3 points) $P(S = s^1 \mid L = l^1,\ I = i^0)$

iii. (3 points) $P(s^1)$

iv. (3 points) $P(G = g^1 \mid s^1)$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

4. Learning tree-structured Bayesian networks (17 points)

Consider learning Bayesian network of five random variables A, B, C, D, E. Assume the data consists of 128 samples. The empirical pairwise mutual information values are given for all pairs as follow:

$\hat{I}(A,B) = 0.32$	$\hat{I}(A,C) = 0.41$	$\hat{I}(A,D) = 0.34$	$\hat{I}(A,E) = 0.39$	$\hat{I}(B,C) = 0.32$
$\hat{I}(B,D) = 0.39$	$\hat{I}(B,E) = 0.37$	$\hat{I}(C,D) = 0.29$	$\hat{I}(C,E) = 0.39$	$\hat{I}(D,E) = 0.35$

i. (5 points) Draw a tree-structured Bayesian network that maximized the likelihood of the observed data. Explain the steps.

ii. (3 points) Compute the BIC score for the tree. If needed suppose the random variables are binary and the summation of the entropies over all random variables is 0.75.

iii. (5 points) Compute the BDe prior for the obtained tree. Assume equivalent sample size is 100 and random variables are binary.

iv. (4 points) Draw a Bayesian network that maximizes the likelihood. Compute its likelihood score.

5. Genome annotation (29 points) *[adjusted from Maji 2012]*

As a computational biologist, you are asked to develop a graphical probabilistic model for annotating each position of an input DNA sequence with Length L to coding(C) or non-coding(N) region. From earlier studies you know the probability of observing nucleotides at coding and non-coding regions are as follow:

Coding region	Non-coding
A: 30%	A: 10%
T: 30%	T: 10%
C: 20%	C: 40%
G: 20%	G: 40%

Traversing on a DNA string, the transition probabilities between two regions are: C-> N : 0.01, N -> C: 0.02. In addition, the first position belongs to the coding or non-coding region with the probability 0.5.

i. Define a general PGM (length L) for this problem. In particular, you need to define
   a. (3 points) random variables and network structure

   b. (3 points) conditional probability tables. Use transition graphs to represent transition probabilities of your model. Write the joint distribution of the model.

ii. (8 points) Compute the probability of observing the DNA string *TG* using variable elimination algorithm (assuming L=2). Draw the model for the special case of L=2. Write down initial, intermediate, and generated factors in each step. Round the numbers to two decimal places (e.g. 0.0971 and 0.003 will round to 0.10 and 0, respectively)

iii. (6 points) Perform the necessary computation to adjust your solution in the previous part for finding the MAP inference. In particular, you need to compute $\max_z P(z|TG)$ as well as $z^* = \arg\max_z P(z|TG)$. Use the traceback procedure for computing z^*.

iv. (4 points) Build the clique tree of the model using the chordal graph approach for the case of L=2.

v. (5 points) Using the clique tree obtained in the previous part, select an appropriate root in your clique tree. First compute the Belief of the root and then compute the probabilities of $P(Z_1 = C \mid TG)$ and $P(Z_2 = \text{N} \mid TG)$.

6. Molecular subtypes of lung cancer (15 points)

In this problem, your task is to develop a probabilistic graphical model for clustering patients with lung cancer in distinct molecular subtypes according to their gene expression patterns. For the sake of simplicity, assume you have binary expression values on only two genes G_1 and G_2 where $G_i = 1$ means the gene i is highly expressed and $G_i = 0$ means the gene is expressed at an expected normal range. In addition, assume we already know there are only two molecular subtypes.

i. (4 points) Suggest a probabilistic graphical model for this task. Define the necessary random variables and your parametrization.

ii. (6 points) Develop an EM algorithm for the estimation of the parameters for *N* observations $\{(g_1[1], g_2[1]), \dots, (g_1[N], g_2[N])\}$. *Hint: start from the ML estimates of parameters and continue from there. No need to prove the ML estimates if you already know from our discussions in the lectures.*

iii. (5 points) Perform the calculations of the E-step and M-step for one iteration of the EM. Choose distinct values for your initial parameters (For example, you can use 0.5 for the mixing parameter; 0.6 and 0.7 for parameters of the first component; and 0.3 and 0.4 for those of the second component). Assume the data consists of

#observations	G_1, G_2
40	$G_1 = 0, G_2 = 0$
20	$G_1 = 0, G_2 = 1$
10	$G_1 = 1, G_2 = 0$
30	$G_1 = 1, G_2 = 1$