COL 776 - Practice Questions

November 16, 2016

Notes:

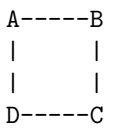
1. Suppose you are the system administrator for a server room running safety critical applications. It is important for the temperature in the room to be below a certain threshold for the servers to function properly and you have installed a system to monintor the temperature on a continuous basis. When the temperature rises above the desired threshold, the system sends you a text message. The temperature in the room could rise above the desired threshold under the following two circumstances a) Someone leaves the door open for a long time b) the AC in the server room stops working. There is a security alarm in the room which goes off when the door is kept open for a long time. We will model this problem using a Bayesian network. Use the following (binary) variables:
   T: Temperature in the server room rises above the desired threshold
   O: Someone leaves the door open for a long time
   A: Air conditioner stops working
   M: A text message is received on the phone
   S: The security alarm goes off.

   (a) Draw a Bayesian network which precisely encodes the conditional independences implied by the statements above. Also, write the expression for the joint probability distribution.

   (b) For each of the following conditional independence statements, state whether they are true or false based on the network structure you drew in the part above. Justify your answers briefly.

      i. O and A are independent
      ii. O and A are independent given M
      iii. S and M are independent
      iv. S and M are independent given O

2. Consider the following Markov network structure:



   Let the only clique potentials be those defined over edges in the network. Let the potentials be given as below (this is inspired by the misconception example covered in the book).
   $P(A, B, C, D) = \frac{1}{Z}\phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(D, A)$

A	B	phi1(A,B)
0	0	100
0	1	1
1	0	1
1	1	100

B	C	phi2(B,C)
0	0	1
0	1	100
1	0	100
1	1	1

C	D	phi3(C,D)
0	0	1
0	1	100
1	0	100
1	1	1

D	A	phi4(D,A)
0	0	100
0	1	1
1	0	1
1	1	100

   Calculate the probability of $P(A = 1|B = 1, C = 1)$. You can produce the final answer as a simplified fraction.

3. Suppose you like to go out for a movie whenever a) there is a new movie in the town or b) you just finished the exams. For both of these cases (and their combinations thereof), you have certain probability of going out to watch a movie.

   - Design an appropriate Bayesian network (along with the CPDs) to model this problem. Use the variables G (going out for a movie), O (exams are over) and N (new movie in the town). Come up with reasonable numbers for your conditional distributions. Is $G \perp O \mid N$?
   - Suppose now that you necessarily go out for a movie whenever there is a new movie in the town. Is it now the case that $G \perp O \mid N = 1$. Argue. These kind of independences are called context specific independences.

4. Consider two events α and β such that $P(\alpha) = p_\alpha$ and $P(\beta) = p_\beta$.

   - Suppose you do not have any additional knowledge about how α and β are related. What are the minimum and maximum possible values for $P(\alpha \cup \beta)$ and $P(\alpha \cap \beta)$?
   - What can you say if you know that α and β are independent?

5. Consider a distribution specified by a Markov network.

   - Show that multiplying all the entries in any one of the factor tables by a constant $k > 0$ does not change the original distribution.
   - Now, suppose we multiply some of the entries of a factor table by a constant $k_1 > 0$ and the remaining entries by a constant $k_2 > 0$ such that $k_1 \neq k_2$. Assume that all the other factors remain the same as before. Show that this necessarily results in a distribution which is different from the original distribution.

6. Consider a distribution $P(X_1, \ldots, X_n)$ $(n > 2)$ which satisfies conditional independencies (CIs) of the form: $X_k \perp X_l | X_1, X_2, Z \quad \forall k > 2, \forall l > 2, k \neq l, \forall Z \subseteq \mathbf{X} - \{X_1, X_2, X_k, X_l\}$, and no other CI. Here $\mathbf{X} = \{X_1, X_2, \ldots X_n\}$ denotes the set of all the variables.

   (a) Draw a Markov network H to represent $P(\mathbf{X})$ as faithfully as possible. In other words, H should satisfy the maximum number of independences in P while ensuring $I(H) \subseteq I(P)$. Justify your construction.

   (b) Draw the junction tree (with minimum possible size of the largest clique node) corresponding to the network above.

7. For the distribution defined in the previous question

   (a) Draw a Bayesian netwrok B to represent $P(\mathbf{X})$ as faithfully as possible (use the same definition of faithful as above). Justify your construction.

   (b) What can you say about the relationship between $I(B)$ and $I(H)$? Justify your answer.

8. Construct an example of a Bayesian network where the corresponding moralized Markov network graph is not chordal.

9. Let X, Y and Z be binary valued random variables. Let z^1 denote the event $Z = 1$. Does $(X \perp Y \mid z^1)$ imply $(X \perp Y \mid Z)$? Prove or provide a counter example.

10. Consider the Bayesian network given in Figure 1 defined over the variables Difficulty(D), Intelligence(I), Grade(G), SAT(S) and Letter (L). Recall that the table associated with each variable node in the network represents its conditional distribution given the values of its parent nodes. Note that the variable Grade in the example considered can take 3 possible values. Calculate the probability $P(I = 0 | G = 3)$. You can use the Bayesian network independence property that a node is indepedent of its non-descendants given its parents. Feel free to use a calculator to get the final answer.

11. Let $\mathcal{G}$ denote a Bayesian network structure. Let $\mathcal{G}$ be a perfect I-map for a set of independencies $\mathcal{I}$, i.e. $I(\mathcal{G}) = \mathcal{I}$. Let $\mathcal{G}'$ be a graph obtained by removing an edge in $\mathcal{G}$. Show that $\mathcal{G}'$ can not be an I-map for $\mathcal{I}$.

12. Consider a Bayesian Network graph $\mathcal{G}$ constructed from a Markov Network graph $\mathcal{H}$ using the procedure discussed in the class such that $\mathcal{G}$ is an I-map for $\mathcal{H}$ i.e., $I(\mathcal{G}) \subseteq I(\mathcal{H})$. If $\mathcal{G}$ is also a perfect I-Map for $\mathcal{H}$ then show that the underlying undirected graph for $\mathcal{G}$ is same as $\mathcal{H}$.

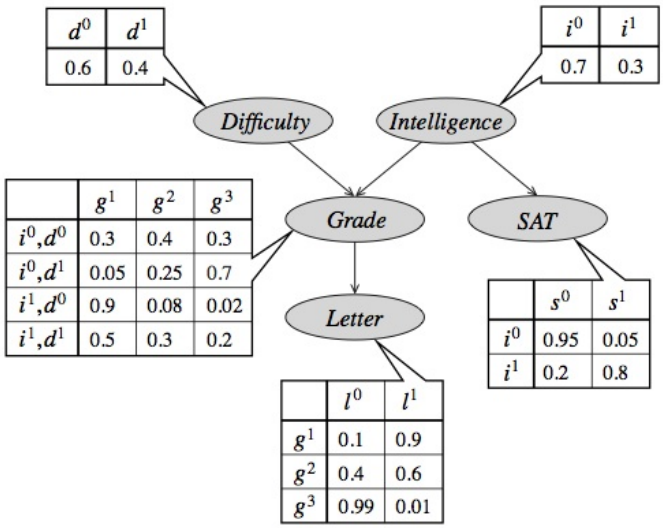


Figure 1: Source: Probabilistic Graphical Models. Daphne Koller and Nir Friedman (2009)

13. Consider the star network as discussed in the class. Recall that the network consists of a hub node X_0 which is connected to rest of the nodes in the network i.e. $X_1, \ldots X_n$. There are no other edges in the network. Assume that all the X_i's are binary. Consider the Markov network represented by the star network with potentials defined over singletons and edges in the graph. Let the singleton potentials be uniform i.e. $\phi^s(X_i = 0) = \phi^s(X_i = 1) = 1, \forall i$. Let the edge potentials be given as $\phi_i^e(X_0, X_i) = 3$ if $X_0 = X_i$ and 1 otherwise, $\forall i \in \{1, \ldots n\}$

    - Construct a clique tree for a star network with $n = 4$, i.e., the variables in the network are $X_0, X_1, \ldots X_4$. You should choose a clique tree which minimizes the size of the largest clique node in the tree.
    - Perform message passing over the clique tree as constructed above to calculate $P(X_0 = 1)$.

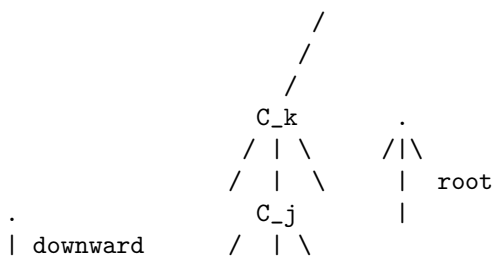
14. Show that the complexity of (marginal) inference, i.e., the complexity of calculating the marginal probability for each of the hidden nodes, in a Hidden Markov Model is linear in the number of hidden nodes. You can assume that all the hidden nodes as well as the observed nodes are binary valued.

15. Show that the clique tree generated by VE computation satisfies the running intersection property.

16. Recall the asynchronous two way message passing algorithm to compute the bi-directional messages in a clique tree. Show that the messages computed by this algorithm are equivalent to the ones computed by the following procedure:

    - Designate (any) node in the clique tree as root C_r.
    - Calculate the upward messages going towards the root (as described in class).
    - Pass the appropriate messages downward from the root all the way to the leaves.

    For the last point above, derive the exact expression for a downward message $\delta_{j \to i}$, where $C_j - C_i$ is an edge and C_j is the immediate upward neighbor of C_i. You should clearly explain the dependence of $\delta_{j \to i}$ on the downward message coming down from the side of the root to C_j, i.e., $\delta_{k \to j}$ where C_k is the immediate upward neighbor of C_j.



```
  | messages     /   |  \
 \|/               C_i
                  /  | \
                 /   |  \
```

17. Consider the student network as given in Figure 4. Use variable elimination (with elimination ordering L, I) to calculate the prbability of $P(S = 0|G = 2, D = 1)$. Simplify your expression as much as you can (use of calculators is not allowed).

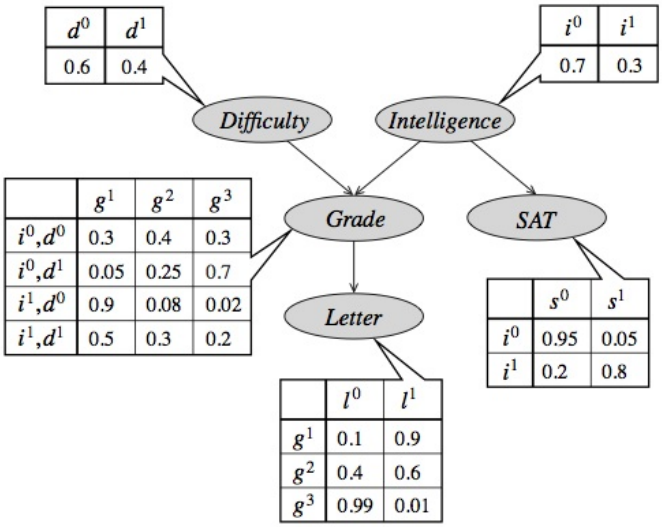


Figure 2: Source: Probabilistic Graphical Models. Daphne Koller and Nir Friedman

18. Consider the Bayesian network shown in Figure 3. Write the expression for the joint distribution specified by this network. Suppose you have a training set composed of the examples given in Table 1, with "?" indicating a missing value. Show the first iteration of the EM algorithm (initial parameters, E-step, M-step), assuming the parameters are initliazed ignoring missing values.

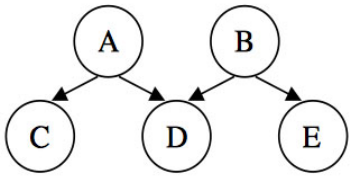


Figure 3: Source: Course Taught by Pedro Domingos.

A	B	C	D	E
0	0	1	1	1
0	0	0	?	1
1	1	0	0	?
1	1	0	0	0
0	1	?	0	1
0	1	1	1	1
1	0	1	0	0
?	1	1	0	1

Table 1: Data with Missing Values

19. Consider the student network as shown in Figure 4. We would like to use loopy BP to calculate the single node marginals in this network. Moralize the graph and construct the corresponding Bethe cluster graph. Recall that Bethe cluster graph is a bi-partite graph with nodes on one side and factors on the other. All the original potentials should be assigned to the factor side of the network. Do not forget to create the factors for potentials involving single variables. Let $\delta^{(t)}_{x \to f}$ denote the message from node x to factor f at time t and similarly, $\delta^{(t)}_{f \to x}$ denote the message from a factor f to a node to x at time t. Let all the messages be initialized to 1 at $t = 0$. At a given time t, $\delta^{(t)}_{x \to f}$ messages are sent first followed by the messages $\delta^{(t)}_{f \to x}$. Assume that we are given the evidence $G = 2, I = 1$. Show the message computation in the network performed by loopy BP at times $t = 1$ and $t = 2$ (you will need to show the messages in both the directions). Also, compute the node marginals at the end of each iteration (i.e. $t = 1$ and $t = 2$). Note: You need to think about how to incorporate the evidence $G = 2, I = 1$ in the message computation step. You should not change the structure of the original cluster graph.

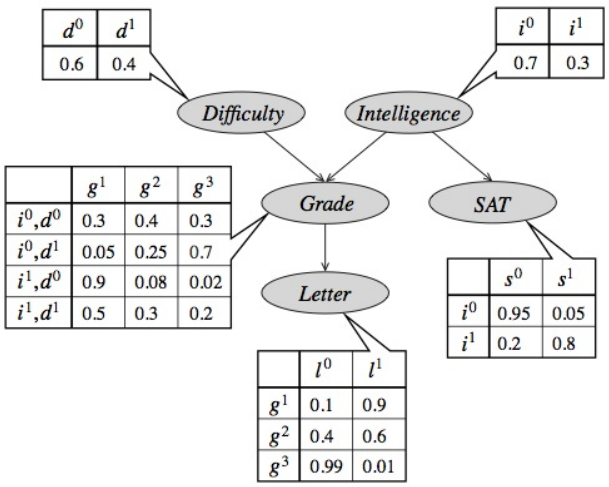


Figure 4: Source: Probabilistic Graphical Models. Daphne Koller and Nir Friedman

20. Show how you can estimate the normalizer Z of an undirected graphical model $P(X) = \frac{1}{Z} \prod_j \phi_j(X_{C_j})$ from a set of m i.i.d. samples $\{x^{(i)}\}_{i=1}^m$ drawn from the distribution $P(X)$. Let n be the number of variables in the network and assume that each variable is binary valued. Hint: Try expressing $1/Z$ as expected value of some function $f(X)$ with respect to the distribution $P(X)$.

21. Consider a Markov network defined over a set of variables represented by X. Let $Y, Z \subseteq X$ denote subsets of variables such that $Y \cap Z = \{\}$, and $Y \cup Z = X$. Consider the following type of inference queries (called the max-marginals) defined over the network.

$$\max_Y \sum_Z P(Y, Z)$$

Do the *max* and $\sum$ operators above commute with each other? In other words, is it the case that $\max_Y \sum_Z P(Y, Z) = \sum_Z \max_Y P(Y, Z)$. Prove formally or give a counter example.

22. (a) Consider a Markov logic network which has the same weight w for all the formulas. What happens in the limit $w \to \infty$? Justify.

    (b) Consider the Friends and Smokers Markov logic network discussed in the class: 1.5 $Smokes(x) \Rightarrow Cancer(x)$; 1.1 $Smokes(x) \wedge Friends(x, y) \Rightarrow Smokes(y)$. Let there be 5 people in the domain $\{Anil, Bunty, Charu, Alka, Mohit\}$. Let the only evidence be that Anil smokes and Anil is friends with Bunty (assume friendship to be symmetric). Rest everything is unknown. What can you say about the (marginal) probabilities of smoking for Charu, Alka and Mohit? Argue. Note that you do not have to calculate the actual probabilities but only reason about how these probabilities may be related to each other.

23. Consider learning a Markov network expressed in the log linear representation: $P(X) = \frac{1}{Z} e^{\sum_k \lambda_k f_k(X_{C_k})}$. Here, each f_k represents a feature defined over the set of variables in the clique C_k and λ_k is the associated

parameter. We would like to learn the MAP (maximum-a-posteriori) parameters for this network using gradient ascent. Let the prior over each parameter be specified using i.i.d. Gaussian distributions with 0 mean and σ^2 variance i.e. $\lambda_k \sim \mathcal{N}(0, \sigma^2)$. Derive the expression for updating the parameter λ_k during each learning iteration. You should calculate the gradient from first principles and not directly use any expressions derived in the class. Note: For $x \sim \mathcal{N}(\mu, \sigma^2)$, $P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

24. Consider using Gibbs sampling for sampling from the distribution $P(X)$ which factorizes over an underlying Markov network. Consider the transition $\mathcal{T}_i$ where variable X_i is sampled given the state $X_{-i} = x_{-i}$ of the remaining variables in the network. Show that T_i satisfies detailed balance with respect to the distribution $P(X)$.

25. Given a distribution $P(\mathcal{X})$ defined over a set of variables $\mathcal{X}$, the problem of MAP inference corresponds to finding an assignment $\mathcal{X} = \mathbf{x}$ which maximizes $P(\mathcal{X} = \mathbf{x})$. The Traveling Sales Person (TSP) problem is a well known NP-hard problem and is defined as follows. Suppose you are given a graph G where nodes represent cities and edge weights denote distances d_{ij} between every pair of cities i, j. The TSP problem asks for the shortest possible tour that visits each city exactly once and returns to the origin city. In other words, it is looking for a *single* cycle in G of smallest weight covering all the nodes in G. Express the TSP problem on a given graph G as a MAP inference problem over some Markov network H. You have to provide all the details of the Markov network and its parameterization: the set of variables, the possible values that each variable can take (they need not be binary valued), the cliques and the corresponding potentials. Each of your potentials should be defined over only a small number of variables in the network.