IITM-CS6730 : Probabilistic Graphical Models Release Date: Feb 9, 2020

Assignment 2 Due Date : Feb 18, 23:59

Roll No: PGM20XYZ Name: Bayesia
Collaborators (if any):
References (if any):

- Use LaTeX to write-up your solutions (in the solution blocks of the source LaTeX file of this assignment), and submit the resulting single pdf file at GradeScope by the due date. (Note: As always, **no late submissions** will be allowed, other than one-day late submission with 10% penalty! Within GradeScope, indicate the page number where your solution to each question starts, else we won't be able to grade it!).
- Collaboration is encouraged, but all write-ups must be done individually and independently, and mention your collaborator(s) if any. Same rules apply for codes written for any programming assignments (i.e., write your own code; we will run plagiarism checks on codes).
- If you have referred a book or any other online material for obtaining a solution, please cite the source. Again don't copy the source *as is* - you may use the source to understand the solution, but write-up the solution in your own words.

1. (4 points) [APPLYING D-SEP PRECISELY!] Let's clarify some implications of d-sep sketched in class. Let X, Y be two **non-adjacent** r.v.s in a BN G, and let X, Y have at least two common ancestors and at least two common descendants. Note that X and Y can either be in an ancestor-descendant relation (denoted *adr*), or not (denoted *non-adr*). Answer if each statement below is True/False.
   (a) (1 point) For adr case, $(X \perp Y \mid \text{all common ancs.})$ holds (i.e., this CI is in $I(G)$ for any G).
   (b) (1 point) For adr case, $(X \perp Y \mid Pa_X \cup Pa_Y)$ doesn't hold in some G.
   (c) (1 point) For non-adr case, whether $(X \perp Y \mid \text{a common anc. } A)$ holds (or not) can depend on which common anc. is chosen as A.
   (d) (1 point) For non-adr case, whether $(X \perp Y \mid \text{a common desc. } D)$ holds (or not) can depend on which common desc. is chosen as D.

2. (5 points) [ISING MEETS BOLTZMANN] Consider a MN H over n nodes (such as a grid graph, with $i \sim j$ indicating edges (X_i, X_j) in H). Ising model defines a joint distbn. over ± 1-valued r.v.s X as $P(x) = \frac{1}{Z'} \exp\left\{-\sum_{i\sim j} a_{ij} x_i x_j - \sum_i b_i x_i\right\}$, whereas Boltzmann machine defines a joint distbn. over $0/1$-**valued** r.v.s. using the same form as $P(x) = \frac{1}{Z} \exp\left\{-\sum_{i\sim j} c_{ij} x_i x_j - \sum_i d_i x_i\right\}$, with the result being a single non-zero contribution of each edge (X_i, X_j)'s potential when $X_i = X_j = 1$).
   (a) (2 points) Show how any Ising model can be reformulated as a Boltzmann machine (with -1 mapped to 0) by showing how the node/edge potentials and partition function of both models are related.

(b) (2 points) Derive the conditional distribution $P(X_i \mid \mathrm{Nbrs}_{X_i})$ for the Boltzmann machine, and express it as a "neuron activation model" (i.e., in terms of the function $\mathrm{sigmoid}(z) = e^z/(1+e^z)$).

(c) (1 point) Is Ising model in the exponential family of distributions? Explain using the defn. of exponential family seen in the last assignment.

3. (7 points) [(OVER)-PARAMETERIZATIONS]

(a) (2 points) Consider a complete graph MN over 3 binary r.v.s parameterized either as a single clique potential or as a product of 3 pairwise potentials. The former requires 8 parameters to write down the clique potential table, whereas the latter requires 12 parameters to write down the pairwise potentials. Can you explain how to resolve the apparent contradiction between these two different parameter counts?

(b) (2 points) Consider a MN comprising among all other factors, two specific factors $\phi_1(A,B)$ and $\phi_2(B,C)$. Will changing these factor parameterizations as below change the joint distribution? Justify your answer.

$$\begin{aligned} \phi_1(A, B=b) &:= \phi_1(A, B=b)\,\lambda_b \\ \phi_2(B=b, C) &:= \phi_2(B=b, C)\,(1/\lambda_b), \end{aligned}$$

where λ_1 and λ_2 are any positive constants.

(c) (3 points) Consider the following factorisation of a distribution P over 4 r.v.s that take values in $\{1, 2, \ldots, a\}$

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = \frac{1}{Z}\psi_1(x_1,x_2)\,\psi_2(x_2,x_3)\,\psi_3(x_3,x_4)\,\psi_4(x_1,x_4)\,\psi_5(x_2,x_4)$$

Let $\mathcal{H}$ be the Markov network for the factorisation above. If Q is a distribution that factorizes according to $\mathcal{H}$, what can be said about the form of the distribution Q? How many parameters would be required to represent such a distribution Q? Compare this with the number of parameters required to represent the distribution P. (You may assume a is large enough that $a - 1 \approx a$ and a^k is negligible when compared to a^{k+1}.)

4. (9 points) [I-EQUIVALENCE AND POSITIVE FRIENDS]

(a) (2 points) How many graphs are I-equivalent to the simple directed chain $X_1 \to X_2 \cdots \to X_n$?

(b) (2 points) Show a BN G (if any) that is I-equivalent to the Diamond Graph MN H (Student Misconception Example with $I(H) = \{(A \perp C \mid B, D), (B \perp D \mid A, C)\}$), when extra variables are allowed in the BN. That is, you are allowed a few extra variables W in the BN G that are always unobserved s.t. the set of dsep-implied CIs in G over the X variables alone (which holds in the marginal distbn. $P(X) = \sum_w P(X, W)$) is identical to $I(H)$.

(c) (2 points) Let $\mathscr{P}$ be a positive distribution, then prove that the Markov Blanket $MB_{\mathcal{P}}(X)$ is unique. [Hint: What happens if two different sets U_1 and U_2 are each minimal $MB_P(X)$?]

(d) (3 points) If a DAG G has no immoralities, then prove that G is chordal (i.e, the underlying undirected graph is chordal). Does such DAGs G always have a perfect MN I-map?

(e) (3 points) [BONUS] Solve Exercise 4.1 from [KF] book related to the non-positive distribution example in Page 116. It will show how Hammersley-Clifford theorem doesn't hold for this distribution.

5. (8 points) [GAUSS EYES MARKOV] A multivariate normal distribution can be viewed either as a BN or as a MN as pointed in our Friday reading. Here we will motivate the latter. Let $\boldsymbol{\Sigma}$ be the covariance matrix of a set of p variables X. Consider the partial covariance matrix $\Sigma_{a|b} := \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$ between the two subsets of variables $X_a = (X_1, X_2)$ consisting of the first two, and X_b the rest. This is the covariance matrix between these two variables, after linear adjustment for all the rest. In the Gaussian distribution, this is the covariance matrix of the conditional distribution of $X_a|X_b$, which is also multivariate Gaussian. Define $\Theta = \Sigma^{-1}$.

   (a) (2 points) Show that $\Sigma_{a|b} = (\Theta_{aa})^{-1}$.

   (b) (3 points) Show that if any off-diagonal element of Θ is zero, then the partial covariance between the corresponding variables is zero. Does this suggest a (minimal I-map) MN H for this distribution, with an edge between any variable pair exhibiting non-zero partial covariance? Why or why not?

   (c) (1 point) Can any multivariate joint distbn. in the exponential family be expressed as a MN distbn. P? If so, specify the (local) factors in this Markov Network MN and their scopes (using terms in the exponential family defn. given in the last assignment).

   (d) (2 points) Express a multivariate normal distribution as an exponential family in terms of its natural parameters: $\Theta = \Sigma^{-1}$ and $\eta := \Sigma^{-1}\mu$, and thereby show that the MN for this exponential family is identical to the MN derived in part (b) above.

6. (7 points) [MINIMAL BN I-MAPS] Solve Exercise 3.11 from [KF] book (first two parts and the additional part below).

   (a) (3 points) Give the minimal I-map for the marginal distribution over all r.v.s except *Alarm* in the Burglary Alarm Network.

   (b) (3 points) Give an algorithm that outputs a minimal I-map G' as in part (a), but when removing (marginalizing out) a r.v. in a general DAG G.
   (Hint: Recall that the minimal BN I-map is not always unique and depends on the node ordering; employ a node ordering that simplifies your algorithm by adding as few edges as possible to G to obtain G').

   (c) (1 point) What happens if your algorithm in part (b) is applied to remove the class variable Y in the Naive Bayes model $P(Y, X) = P(Y)\prod_{i=1}^{p} P(X_i|Y)$?