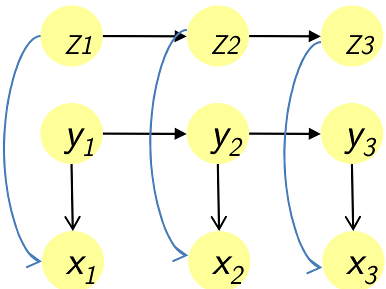
CS 726: Advanced Machine Learning, Fall 2021, Mid-Semester exam

February 27, 2021.
6:00 – 8:00 pm

Write all your answers and upload on SAFE. Do not spend time/space giving irrelevant details or details not asked for. Use the marks as a guideline for the amount of time you should spend on a question. You are only allowed to refer your notes, and my lecture slides. No other material.

---

1. Consider the parallel chain HMM given below.



   (a) Is $y_i \perp\!\!\!\perp z_{i+1}$? Give a brief justification. No marks without proper justification. ..1
   Yes, all trails between them involve a V-node.

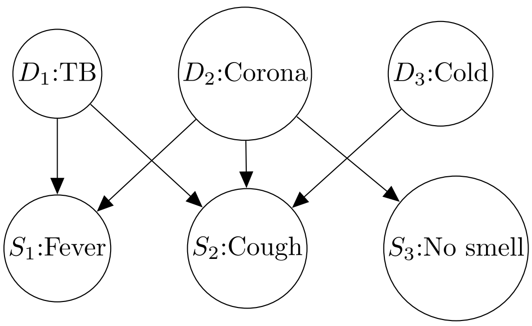
   (b) If $y_1 \perp\!\!\!\perp z_3 | x_1, W$, then list a minimal set other than y_1, z_3 that must be present in W. Your answer may be an empty set too. ..1 Either z_1 or z_2.

   (c) List a CI that holds in this BN but will be lost when it is converted to an undirected graphical model. ..1 Many, e.g. $y_i \perp\!\!\!\perp z_i$

   (d) Using this parallel HMM to answer CI queries, let us say you try to create a fresh BN using the insertion order $y_1, y_2, y_3, z_1, x_1, x_2, x_3, z_2, z_3$ on the algorithm for BN construction discussed in class. Draw the BN you will get using this algorithm. ..3
   y_1, y_2, y_3, x_1 are in same order. z_1 will have y_1, x_1 as parents. x_2 will have y_2, z_1 as parents. x_3 will have y_3, z_1 as parents. ..

   (e) List two variable orderings for which the BN that you will get will be the same as the original graph. ..1 $y_3, y_2, y_1, z_1, z_2, z_3, x_1, x_2, x_3$
   $y_3, y_2, y_1, z_3, z_2, z_1, x_2, x_1, x_3$

2. Consider a special QMR Bayesian network comprising of disease nodes $D_1, \dots, D_n$ and symptom nodes $S_1, \dots S_k$ An example appears below.



Assume all variables are binary. Let probability of a disease D_i occurring be denoted as $P(D_i = 1) = \beta_i$ and α_{ij} denote the probability of a patient showing symptom S_j if he only has disease D_i. Also, assume the probability of a symptom not occurring conditioned on its parent nodes is $\Pr(S_j = 0|Pa(S_j)) = \prod_{i \in Pa(S_j)} (1 - \alpha_{ij})^{D_i}$. For example, if someone has D_1, D_2 but not D_3 then the probability of his not showing symptom S_2 is $(1 - \alpha_{12})(1 - \alpha_{22})$

(a) In a study over 1000 random Indians, 20 had Cold, 2 had TB, and 1 had Corona. Doctors have estimated that Corona causes each of these symptoms with probability 0.2, Flu causes cough with probability 0.7, and TB causes fever with probability 0.8, and cough with probability 0.3. Write exact numerical values for the following potentials as a table:

i. $P(D_1)$ ..1

$D_1 =$	0	1
	0.998	0.002

ii. $P(S_3|Pa(S_3))$ ..2

$D_3 =$	0	1
$P(S_3) = 0$	1	0.3
$P(S_3) = 1$	0	0.7

iii. $P(S_1|Pa(S_1))$ ..2

$D_1, D_2 =$	0,0	0,1	1,0	1,1
$P(S_3) = 0$	1	0.2	0.8	0.16

(b) What is the marginal probability $P(S_1 = 1)$ in the above example graphical model? Express only in terms of α_{ij}s and β_is. ..2 $\sum_{D_1,D_2} \Pr(S_i = 0|D_1, D_2)P(D_1)P(D_2) = (1 - \alpha_{21})(1 - \alpha_{11})\beta_1\beta_2 + (1 - \alpha_{11})\beta_1(1 - \beta_2) + (1 - \alpha_{21})(1 - \beta_1)\beta_2 + (1 - \beta_1)(1 - \beta_2)$

3. Consider the reduction of 2-SAT to Bayesian network inference discussed in class. Consider an instance of 2-SAT on the following set of clauses over three Boolean variables.

- $C_1 = x_1 \vee x_3$
- $C_2 = \bar{x_1} \vee \bar{x_2}$
- $C_3 = x_2 \vee \bar{x_3}$

where we are interested in the truth of $S = C_1 \wedge C_2 \wedge C_3$

(a) For this 2-SAT problem, draw the corresponding Bayesian network (BN) over nodes $x_1, x_2, x_3, C_1, C_2, C_3, S$. Do not introduce any other nodes. For node C_2 show the potentials as a conditional probability table. ..2

(b) Moralize the above BN ..1

(c) Triangulate the above graph by adding extra edges as required. ..2

(d) Identify the maximal cliques in the graph ..2

4. Consider the problem of training the parameters of a simple HMM of length two where the state and observation variables are binary. Thus, we have two state variables y_1 and y_2 and two output variables x_1 and x_2 and all four variables can take one of two possible values. The parameters of the HMM are $\Pr(y_1)$ $\Pr(y_2|y_1)$ and $\Pr(x_1|y_1)$ and $\Pr(x_2|y_2)$. Assume $\Pr(x_1|y_1) = \Pr(x_2|y_2) = \Pr(x_t|y_t)$ We use the EM algorithm for training the parameters.

Let the initial values at $t = 0$ be

$$\Pr{}^t(y_1 = 0) = \theta_0^t = 0.5$$
$$\Pr{}^t(y_2 = 0|y_1 = 0) = \theta_1^t = 0.7, \quad \Pr{}^t(y_2 = 0|y_1 = 1) = \theta_2^t = 0.2$$
$$\Pr{}^t(x_t = 0|y_t = 0) = \theta_3^t = 0.1, \quad \Pr{}^t(x_t = 0|y_t = 1) = \theta_4^t = 0.8.$$

Let a dataset D consist of four instances:

- In the first one all four variables are observed and take values $\mathbf{x}^1 = [0, 1], \mathbf{y}^1 = [0, 1]$,
- In the second instance, three variables are observed and take values $\mathbf{x}^2 = [1, 0], y_1^2 = 1$,
- In the third instance, observed variables are $x_1^3 = 0, y_2^3 = 1$.
- In the fourth instance, observed variables are $x_1^4 = 1, x_2^4 = 1$.

(a) List all the variational q variables that need to be estimated in the E-step of EM. You need to simplify them as much as possible. For example, if two variable v, w are independent for a given instance i, write these as $q_i(v) = P(v|..), q_i(w) = P(w|...)$ instead of $q_i(v, w) = P(v, w|..)$. ..2 None for first, only $q_2(y_2)$ for second, $q_3(y_1), q_3(x_2)$ for third and $q_4(y_1, y_2)$ for the last.

(b) Estimate the values of $\Pr(y_2|\mathbf{x}^2, y_1^2, \theta^t)$ ..2

For the E-step

The node potentials at y_2 is $\psi(y_2) = \Pr(x_2^2 = 0|y_2)\Pr(y_2|y_1^2 = 1) = [0.1 * 0.2 + 0.8 * 0.8]$ Normalizing this we get the required probabilities.

(c) For the M-step write the formula for the maximum likelihood estimate of θ_1 in terms of the observations and q-variables from the E-step. ..2

$P(y_2 = 0|y_1 = 0) = (0 + 0 + q_4(y_1 = 0, y_2 = 0))/(1 + q_3(y_1 = 0) + q_4(y_1 = 0))$

5. Consider a real input $\mathbf{x}$ which can be assigned a vector of three binary labels y_1, y_2, y_3. The features for this labeling are as follows:

$$\begin{aligned} f_1((y_1), (1), \mathbf{x}) &= x_1 y_1 \\ f_2((y_2), (2), \mathbf{x}) &= x_1 y_2 \\ f_3((y_3), (3), \mathbf{x}) &= x_1 y_3 \\ f_4((y_1, y_2, y_3), (1, 2, 3), \mathbf{x}) &= (y_1 + y_2 + y_3)^2 \end{aligned} \tag{1}$$

Let $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4]$ denote the corresponding weights of these four features $\mathbf{f} = [f_1, f_2, f_3, f_4]$. Consider two instances in the training data:

$$\begin{aligned} \mathbf{x}^1 &= 0, \quad \mathbf{y}^1 = [0, 0, 0] \\ \mathbf{x}^2 &= -1, \quad \mathbf{y}^2 = [0, 1, 1] \end{aligned} \tag{2}$$

(a) Write the expression for $\Pr(\mathbf{y}^1|\mathbf{x}^1)$ in terms of $\theta_1, \theta_2, \theta_3, \theta_4$ [Do not use $f_k()$s but their defined values and also given values of $\mathbf{x}^1, \mathbf{y}^1$]. Also, express value of the normalizer $Z(\mathbf{x}^1)$ in terms of θ. ..2

Numerator will be 1 since all features evaluate to zero. $Z(\mathbf{x}^1)$ is $1 + 3\exp(\theta_4) + 3\exp(4\theta_4) + \exp(9\theta_4)$

(b) Draw the junction tree and assign the clique potentials at the node of the JT for instance $\mathbf{x}^2$ when $[\theta_1^t, \theta_2^t, \theta_3^t, \theta_4^t] = [1, 2, 4, 0]$. ..2

The JT will have three disconnected nodes since the dependency among the variables is lost when $\theta_4 = 0$. The node potentials will be just $\psi_i(y_i) = \exp(-1 y_i \theta_i)$.

(c) What is the value of the normalizer $Z(\mathbf{x}^2)$ at the above θ? ..2

$(exp(-1 * 1) + 1)(exp(-1 * 2) + 1)(exp(-1 * 4) + 1)$

(d) Compute the gradient of $\log \Pr(\mathbf{y}^2|\mathbf{x}^2, \theta^t)$ wrt θ_4 at $[\theta_1^t, \theta_2^t, \theta_3^t, \theta_4^t] = [0, 0, 0, 1]$ ..3
$f_4(\mathbf{y}^i, \mathbf{x}^i) - E_{\Pr(\mathbf{y}|\mathbf{x}^i,\theta^t)}[f_4(\mathbf{y}^i, \mathbf{x}^i)]$ can be easily computed as. $(0 + 1 + 1)^2 - (3 * 1 * \exp(1)/Z + 3 * 4 * \exp(4)/Z + 1 * 9 * \exp(9)/Z)$ where Z at this theta is $1 + 3\exp(\theta_4) + 3\exp(4\theta_4) + \exp(9\theta_4)$

6. The complexity of computing the value of Z in an undirected graphical model over n variables each of cardinality m and where the graph consists of a single loop is $O(nm^3)$. Now consider the case where half of the adjacent variables, say $x_1, \ldots, x_{\frac{n}{2}}$ have cardinality m, and the other half have cardinality M where $M \gg m$. In this case, design a judicious variable elimination order so as to take less than $O(nM^3)$ time. ..4 By choosing a judicious ordering of variables we can avoid the $O(M^3)$ complexity. For example, if we eliminate variables in the order: $x_1, x_n, x_2, x_{n-1}, x_3, x_{n-2}, \ldots$, etc. We will get a complexity of $O(nm^2M + nmM^2)$

$\boxed{\textbf{Total: 40}}$