CS 726: Advanced Machine Learning, Fall 2019, Mid-Semester exam

February 28, 2019.
4:00 – 6:00 pm

Roll: _____

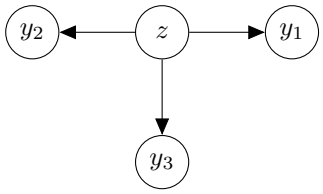
Name: _____

Mode: Credit/Audit/Sit-through _____

Write all your answers in the space provided. 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 allowed to write elsewhere only under special circumstances like total cancellation of a previously written answer. Use the last sheet of this booklet in such cases. You are only allowed to refer your notes, no one else's notes or textbook.

---

1. In this question we will draw a Bayesian network corresponding to a crowd-labeling scenario. Let $\mathbf{x}$ be an instance and z be its true binary label, that is, either -1 or +1. We do not observe the true z. But we are given a set of n noisy labelers that output labels $y_1, \ldots, y_n$. Each y_j could be +1, -1, or 0. We express their dependency as a Bayesian network. An example with $n = 3$ is shown below.



We write the potentials as

$$P(z = 1) = \rho, \quad P(z = -1) = 1 - \rho$$

and

$$P_j(y_j|z) = \begin{cases} \frac{e^{\theta_j}}{Z_j} & \text{if } z = y_j, \\ \frac{e^{-\theta_j}}{Z_j} & \text{if } z \neq y_j \ \& \ y_j \neq 0, \\ \frac{1}{Z_j} & \text{if } \ y_j = 0 \end{cases} \tag{1}$$

where $\rho, \theta_1, \ldots, \theta_n$ are parameters of the network, and $Z_1, \ldots, Z_n$ are normalizers.

(a) What is value Z_j in terms of θ_j so that $P_j(y_j|z)$ is a valid distribution for $z = 1$ and $z = -1$. ..1 $Z_j = 1 + e^{\theta_j} + e^{-\theta_j}$

Consider the three samples below with $n = 2$.

y_1	y_2	z	$\Pr(y_1, y_2, z)$
0	1	1	
-1	0	-1	
-1	1	-1	

(b) First, let us assume that z is visible. Write the expression for probability for the first two samples samples in the last column above in terms of ρ, θ_1, θ_2 ..2

y_1	y_2	z	$\Pr(y_1, y_2, z)$
0	1	1	$\rho \frac{1}{1+e^{\theta_1}+e^{-\theta_1}} \frac{e^{\theta_2}}{1+e^{\theta_2}+e^{-\theta_2}}$
-1	0	-1	$(1-\rho) \frac{e^{\theta_1}}{1+e^{\theta_1}+e^{-\theta_1}} \frac{1}{1+e^{\theta_2}+e^{-\theta_2}}$
-1	1	-1	$(1-\rho) \frac{e^{\theta_1}}{1+e^{\theta_1}+e^{-\theta_1}} \frac{e^{-\theta_2}}{1+e^{\theta_2}+e^{-\theta_2}}$

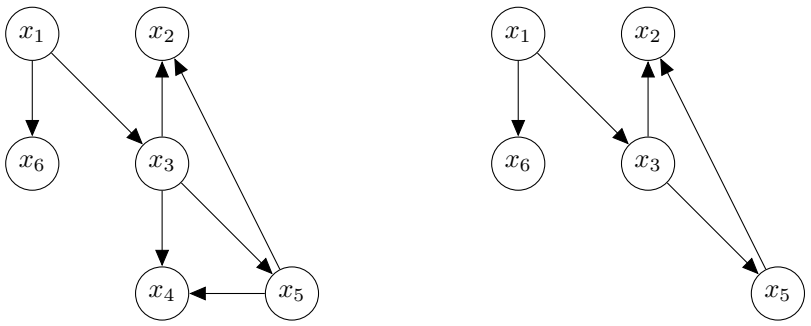
(c) Now, we will apply the EM algorithm to estimate the values of the various parameters where only y_j values are observed but the z values are hidden. Let the data sample for y_1, y_2 be the same as above except that the z column is missing. Let the values of parameters at iteration t of EM be $\rho^t, \theta_1^t, \theta_2^t$ be 1/3, 1, and 0. Using these compute the E step for the first two examples. [You should substitute the given values, and not just write the expression in terms of θ^t and ρ^t. The answer can be left in fractional form] ..3

$P(z = 1|0, 1) = \frac{(\rho)e^{\theta_2}}{(1-\rho)e^{-\theta_2}+\rho e^{\theta_2}} = 1/3$

$P(z = 1| - 1, 0) = \frac{(1/3)e^{-1}}{2/3e^{1}+1/3e^{-1}} = \frac{e^{-1}}{2e^{1}+e^{-1}}$

(d) Let $q_i^t(z)$ denote value of the EM variables for example i that you computed above. In terms of these, write the expression for the expected log-likelihood for the first example in the table and simplify. ..3 $q_1^t(1) \log P(0, 1, 1) + (1 - q_1^t(1)) \log P(0, 1, 1) = q_1^t(1) \log \rho + q_1^t(1)\theta_2 + (1 - q_1^t(1)) \log(1 - \rho) - \theta_2(1 - q_1^t(1)) - \log((1 - \rho)e^{-\theta_2} + \rho e^{\theta_2})$

2. Let L be a BN with n variables. Let R be obtained from L by removing a node w that has no children. For example, if we call the graph on the left side below as L, and node w as x_4 (Note x_4 has no children), we will get the rightside graph R where node x_4 and the edges incident on it are dropped.



Let X, Y, Z denote disjoint subsets of variables in R. We use dsep(X, Y, Z, R) to denote that Z d-separates X and Y in R. For example, dsep$(\{x_1, x_6\}, \{x_2, x_5\}, \{x_3\}, R)$ is true.

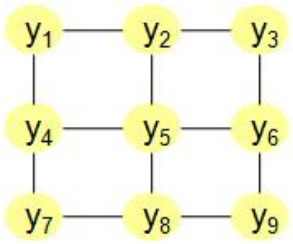
(a) Justify briefly that dsep(X, Y, Z, R) implies dsep(X, Y, Z, L). ..2

w can only appear as a V-node in any path between X and Y and it is not a part of Z.

(b) Let P denote pa$(w) - Z$ where pa(w) denotes parents of w. Justify briefly that dsep(w, Y, Z, L) implies dsep(P, Y, Z, R). ..3

Assume by contradiction that there is path p between P and Y that is not blocked by Z in G. p extended with w will also be unblocked in L since no node in P can be a v-node on the w to Y path. Thus, dsep(w, Y, Z, L) will be violated.

3. Suppose we are given an arbitrary undirected graph $G = (V, E)$ where $V = \{v_1, \ldots, v_n\}$ denotes the set of n nodes, and $E = \{(h_1, t_1), \ldots, (h_m, t_m)\}$ denotes its m edges with $h_i, t_i \in V$. Our goal is to find the maximum b-matching of G. A b-matching is defined as the largest subset $M \subseteq E$ of edges of G such that each vertex v of G is adjacent to at most b edges in M. For example, for the grid graph below with $b = 2$, one maximum b-matching is the eight edges on the boundary of the grid.



We will solve this problem by creating a new undirected graphical model H with appropriately defined potentials so that the maximum b-matching problem in G can be obtained from the highest scoring labeling (MAP) in H. The number of nodes in H should be polynomial in the number of nodes and edges in G. Specify each of the steps below to provide your reduction. ..5

(a) Variables $x_1, \ldots, x_K$ in H along with the set of values each node is allowed to take. What is K in terms of number of nodes n and edges m in G? For each edge e in G create a binary variable x_e in H.

(b) Potentials of H. For every vertex v in G, create a clique potential ψ_v in H over all edges incident on v. Let $\mathbf{x}_v$ denote the subset of edges incident on vertex v. $\psi_v(\mathbf{x}_v)$ is 0 when number of ones $\mathbf{x}_v$ is $\geq b$. For all other cases it is exponent of the number of ones in $\mathbf{x}_v$.

(c) Edges of H. Two nodes x_i and x_j will be connected if there is a vertex v of G where edge i and j are co-incident.

4. Consider the problem of training CRFs via the loglikelihood method which gives rise to an objective of the form:

$$\max \sum_{i=1}^{N} \sum_{c} \theta.\mathbf{f}(\mathbf{x}^i, \mathbf{y}_c^i, c) - \log Z(\mathbf{x}^i) \tag{2}$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_c \exp(\theta.\mathbf{f}(\mathbf{x}, \mathbf{y}_c, c))$ The presence of the $Z(\mathbf{x})$ requires us to run a message passing step to compute the objective and the gradient. We will avoid that by approximating as $Z(\mathbf{x}) \approx Z_A(\mathbf{x}) = \prod_c \sum_{\mathbf{y}_c} \exp(\theta.\mathbf{f}(\mathbf{x}, \mathbf{y}_c, c))$.

For example, on a chain graph of three variables y_1, y_2, y_3,

$$Z(\mathbf{x}) = \sum_{y_1, y_2, y_3} \exp(\theta.\mathbf{f}(\mathbf{x}, (y_1, y_2), (1, 2))) \exp(\theta.\mathbf{f}(\mathbf{x}, (y_2, y_3), (2, 3)))$$

whereas

$$Z_A(\mathbf{x}) = (\sum_{y_1, y_2} \exp(\theta.\mathbf{f}(\mathbf{x}, (y_1, y_2), (1, 2))))(\sum_{y_2, y_3} \exp(\theta.\mathbf{f}(\mathbf{x}, (y_2, y_3), (2, 3))))$$

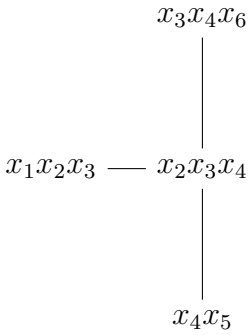
(a) Give as simple an example as you can where $Z(\mathbf{x}) \neq Z_A(\mathbf{x})$ using only three variables and setting appropriate values for θ and $\mathbf{f}$. ..2 let θ be zero. The $Z(\mathbf{x}) = 8$ where $Z_A(\mathbf{x}) = 4 * 4 = 16$

(b) Show an example of an undirected graphical model where the time required for $Z(\mathbf{x})$ is exponentially more than the time required for $Z_A(\mathbf{x})$. ..3 If graph is complete but potentials are only edges.

(c) Write the gradient of $\log Z_A(\mathbf{x})$ wrt θ ..3

5. Suppose a distribution $P(x_1, \ldots, x_n)$ is specified only in terms of a Junction Tree JT along with its clique potentials.

(a) For the junction tree below, draw alongside the triangulated graph H representing an undirected graphical model that could have generated this JT. ..2



(b) For the example above, draw the graph structure of a Bayesian network G to represent P (you need to assign directions to all edges, just a skeleton will not do.) ..3

(c) Let's say we next wish to compute $\Pr(x_6, x_5|x_4 = 1)$. Assume message passing has been performed on the entire junction tree, and you can use the precomputed messages along with any extra computation to answer the query. Write the expression for this computation in terms of clique potentials $\psi_c(\mathbf{x}_c)$, and messages $m_{ci\to cj}(\mathbf{x}_{S_{ij}})$ ..3
$\Pr(x_6, x_5|x_4 = 1) = \Pr(x_5|x_4 = 1)\Pr(x_6|x_4 = 1)$ since x_4 separates x_5 from the rest of the tree. Therefore, we do not need to recompute any messages. Just compute the conditional probability in each clique node separetely and multiply.

Total: 35