CS 475 Machine Learning: Homework 4 Analytical (70 points)

Assigned: Monday, Nov. 1st, 2021

Due: Monday, Nov. 15th, 2021, 11:59 pm US/Eastern

Partner 1: Harrison Khoo (hkhoo2), Partner 2: Dimitri Lezcano (dlezcan1)

Instructions

We have provided this LaTeX document for turning in this homework. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box. You can change the size of the box if you need more space.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

Do not add text outside of the answer boxes. You are allowed to make boxes larger if needed.

We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

MRFs

Question 1.

Consider the graphical model shown in Figure 1. In this model, $\mathbf{x}$ is a sequence of observations for which we want to output a prediction $\mathbf{y}$, which itself is a sequence, where the size of $\mathbf{y}$ is the same as $\mathbf{x}$. Assume that the potential functions have a log-linear form: $\psi(Z) = \exp\{\sum_i \theta_i f_i(Z)\}$, where Z is the set of nodes that are arguments to the potential function (i.e. some combination of nodes in $\mathbf{x}$ and $\mathbf{y}$,) θ are the parameters of the potential functions and f_i is a feature function.

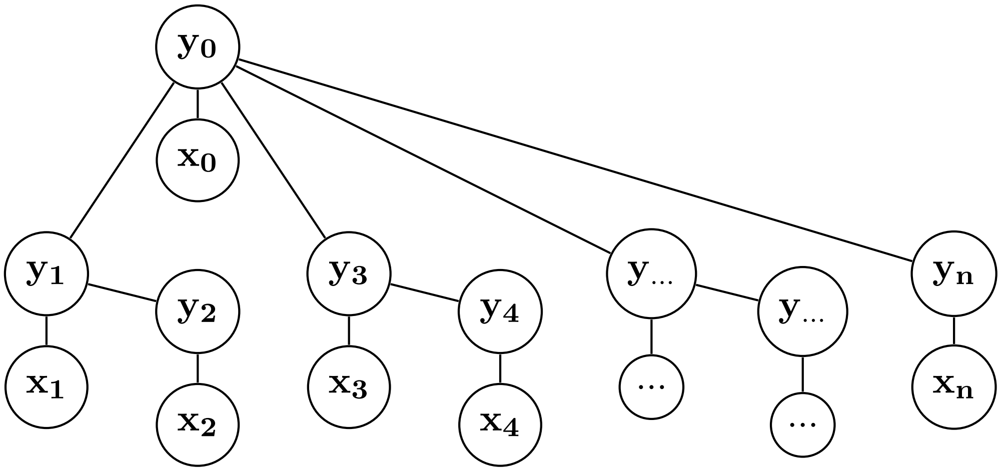


Figure 1: Tree structure model

(a) Write the log likelihood for this model of a single instance $\mathbf{x}$: $\log p(\mathbf{y}, \mathbf{x})$.

(b) Write the conditional log likelihood for this model of a single instance $\mathbf{x}$: $\log p(\mathbf{y}|\mathbf{x})$.

(c) Assume that each variable y_i can take one of k possible states, and variable x_i can take one of k' possible states, where k' is very large. Describe the computational challenges of modeling $\log p(\mathbf{y}, \mathbf{x})$ vs $\log p(\mathbf{y}|\mathbf{x})$.

(a) Cliques: $(y_0, y_j), \forall 0 < j \leq n$, (y_j, x_j) and $(y_j, y_{j+1}), \forall j$

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z}\psi(x_n, y_n) \prod_{j=2}^{n} \psi(y_0, y_j) \prod_{j=0}^{n-1} \psi(x_j, y_j)\psi(y_j, y_{j+1})$$

$$\log p(\mathbf{y}, \mathbf{x}) = -\log Z + \sum_i \theta_i \Bigg(f_i(x_0, y_0) + f_i(x_1, y_1) + f_i(y_0, y_1)$$

$$+ \sum_{j=2}^{n} f_i(x_j, y_j) + f_i(y_0, y_j) + f_i(y_{j-1}, y_j) \Bigg)$$

(b) Since there are no cliques entirely in $\mathbf{x}$, we have that

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})}\psi(x_n, y_n)\prod_{j=2}^{n}\psi(y_0, y_j)\prod_{j=0}^{n-1}\psi(x_j, y_j)\psi(y_j, y_{j+1})$$

$$\log p(\mathbf{y} \mid \mathbf{x}) = -\log Z(\mathbf{x}) + \sum_i \theta_i\Bigg(f_i(x_0, y_0) + f_i(x_1, y_1) + f_i(y_0, y_1)$$
$$+ \sum_{j=2}^{n} f_i(x_j, y_j) + f_i(y_0, y_j) + f_i(y_{j-1}, y_j)\Bigg)$$

(c) The main difference between $\log p(\mathbf{y}, \mathbf{x})$ and $\log p(\mathbf{y} \mid \mathbf{x})$ is the Z term. Here, we have that k' is very large, so computing Z for $\log p(\mathbf{y}, \mathbf{x})$ would perform an operation proportional to an exponential of k'. However, in considering $Z(\mathbf{x})$ in $\log p(\mathbf{y} \mid \mathbf{x})$, Z is parameterized by $\mathbf{x}$, and we only have to sum the cliques over $\mathbf{y}$, therefore we never have to perform any computations over the very large number of parameters k', which is much more favorable.

Question 2.

(a) Suppose you wanted to compute $S = \sum_{x_1=1}^{100} \cdots \sum_{x_8=1}^{100} h(x)$ where

$$h(x) = \exp(x_1x_2 + x_4x_5 + x_7x_8) \prod_{i=2,5,7} (x_i + x_3 + x_6)^i.$$

It looks like the sum has $100^8 = 10^{16}$ terms, so it seems we must evaluate h 10^{16} times. Explain (precisely) how you can compute S with at most 10^7 evaluations of h or something simpler than h.

(b) Draw the MRF associated with this distribution.

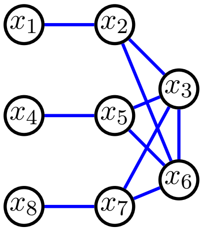
(a)

$$
\begin{aligned}
S &= \sum_{x_1,\ldots,x_8=8}^{100} e^{x_1x_2}e^{x_4x_4}e^{x_7x_8}(x_2+x_3+x_6)^2(x_5+x_3+x_6)^5(x_7+x_3+x_6)^7 \\
&= \sum_{x_3,x_6=1}^{100}\left(\sum_{x_1,x_2=1}^{100} e^{x_1x_2}(x_2+x_3+x_6)^2\right)\left(\sum_{x_4,x_5=1}^{100} e^{x_4x_5}(x_5+x_3+x_6)^5\right) \\
&\quad \left(\sum_{x_7,x_8=1}^{100} e^{x_7x_8}(x_7+x_3+x_6)^7\right)
\end{aligned}
$$

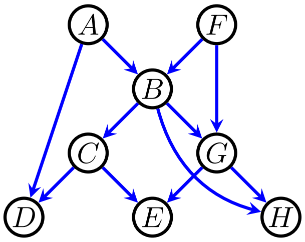
The number of terms are 100^2 for the x_3, x_6 terms multiplied by 3×100^2 for each pairing $(x_1, x_2), (x_4, x_5), (x_7, x_8)$. Together, this means there are $3 * 10^8$ terms. We're aware that this is greater than 10^7 terms listed in the problem statement, but aren't sure how to further reduce the number of x_3, x_6 terms.

Edit: Say we store all the terms in a look-up table. Then each of the exponential terms would evaluate $100^2 = 10^4$ evaluations of h and the terms with x_3, x_6 would result in $100^3 = 10^6$ terms. Since we have 3 sets of these, that would result in a maximum number of evaluations of h of $3(10^4 + 10^6) < 10^7$.

(b) MRF:



DAGs, Clique Trees and Message Passing.



In a statistical DAG model for the graph shown, let $\mathbf{V} = \{A, B, C, D, E, F, G, H\}$.

(a) Answer (and explain your answer) the following d-separation queries:

$$A \perp\!\!\!\perp F \mid D$$
$$A \perp\!\!\!\perp G \mid B, C$$
$$G \perp\!\!\!\perp A \mid B, H, D, E, F$$
$$F \perp\!\!\!\perp D \mid A, B$$
$$C \perp\!\!\!\perp H \mid B$$

(b) Write down the local Markov property of this model.

(c) Consider a new graph where we reverse the direction of the edge $B \to G$ to point the other way: $B \leftarrow G$ (and leave the other edges the same). Does the new graph represent the same model as the old?

Hint: write down the local Markov property for the new graph, and see if all statements in it are implied by d-separation in the original graph. In general, if local Markov of $\mathcal{G}_1$ is implied by global Markov of $\mathcal{G}_2$, and local Markov of $\mathcal{G}_2$ is implied by global Markov of $\mathcal{G}_1$, then $\mathcal{G}_1$ and $\mathcal{G}_2$ represent the same model. Otherwise they do not.

(d) A moralized graph $\mathcal{G}^a$ is obtained from a DAG $\mathcal{G}$ by connecting all non-adjacent variables V_i and V_j such that $V_i \to V_k \leftarrow V_j$ is in the graph (for some V_k), and replacing all directed edges by undirected edges. What is the moralized graph for the DAG in this problem?

(e) Write down the MRF factorization of the moralized graph $\mathcal{G}^a$.

(f) Is this graph chordal? If not, add a set of edges to make it chordal. If you added edges, write the factorization of the new graph.

(g) Create a clique tree from the triangulated graph (either $\mathcal{G}^a$ or the graph obtained from $\mathcal{G}^a$ by adding new edge(s)).

(h) Pick a root $\mathbf{R}$ of the clique tree, and calculate both incoming messages $\phi^{\mathbf{S}_i \to \mathbf{S}_j}$ from each $\mathbf{S}_i$ towards its neighbor $\mathbf{S}_j$ closer to the root, and outgoing messages $\phi^{\mathbf{S}_k \leftarrow \mathbf{S}_i}$ from $\mathbf{S}_i$ to each neighbor $\mathbf{S}_k$ further than $\mathbf{S}_i$ from the root, in terms of clique potentials and other messages.

(i) By substituting in the clique factors in each message, show that in this example, for each leaf node $\mathbf{S}_i$ with a neighbor node $\mathbf{S}_j$,

$$p(\mathbf{S}_i) = \frac{\phi^{\mathbf{S}_i \leftarrow \mathbf{S}_j}_{\mathbf{S}_j \setminus \mathbf{S}_i} \phi_{\mathbf{S}_i}}{\sum_{\mathbf{S}_i} \phi^{\mathbf{S}_i \leftarrow \mathbf{S}_j}_{\mathbf{S}_j \setminus \mathbf{S}_i} \phi_{\mathbf{S}_i}} = \frac{\sum_{\mathbf{V} \setminus \mathbf{S}_i} \prod_{C \in \mathcal{C}(\mathcal{G})} \phi_C}{\sum_{\mathbf{V}} \prod_{C \in \mathcal{C}(\mathcal{G})} \phi_C}$$

for each non-leaf note $\mathbf{S}_i$ with a neighbor $\mathbf{S}_j$ closer to the root, and neighbors $\mathbf{S}_1, \dots, \mathbf{S}_m$ further from the root that

$$p(\mathbf{S}_i) = \frac{\phi^{\mathbf{S}_i \leftarrow \mathbf{S}_j}_{\mathbf{S}_j \setminus \mathbf{S}_i} \left(\prod_{k=1}^{m} \phi^{\mathbf{S}_k \to \mathbf{S}_i}_{\mathbf{S}_k \setminus \mathbf{S}_i} \right) \phi_{\mathbf{S}_i}}{\sum_{\mathbf{S}_i} \phi^{\mathbf{S}_i \leftarrow \mathbf{S}_j}_{\mathbf{S}_j \setminus \mathbf{S}_i} \left(\prod_{k=1}^{m} \phi^{\mathbf{S}_k \to \mathbf{S}_i}_{\mathbf{S}_k \setminus \mathbf{S}_i} \right) \phi_{\mathbf{S}_i}} = \frac{\sum_{\mathbf{V} \setminus \mathbf{S}_i} \prod_{C \in \mathcal{C}(\mathcal{G})} \phi_C}{\sum_{\mathbf{V}} \prod_{C \in \mathcal{C}(\mathcal{G})} \phi_C}$$

and finally for the root node $\mathbf{S}_i$ with neighbors $\mathbf{S}_1, \dots, \mathbf{S}_m$ that

$$p(\mathbf{S}_i) = \frac{\left(\prod_{k=1}^{m} \phi^{\mathbf{S}_k \to \mathbf{S}_i}_{\mathbf{S}_k \setminus \mathbf{S}_i} \right) \phi_{\mathbf{S}_i}}{\sum_{\mathbf{S}_i} \left(\prod_{k=1}^{m} \phi^{\mathbf{S}_k \to \mathbf{S}_i}_{\mathbf{S}_k \setminus \mathbf{S}_i} \right) \phi_{\mathbf{S}_i}} = \frac{\sum_{\mathbf{V} \setminus \mathbf{S}_i} \prod_{C \in \mathcal{C}(\mathcal{G})} \phi_C}{\sum_{\mathbf{V}} \prod_{C \in \mathcal{C}(\mathcal{G})} \phi_C}$$

Here $\mathbf{V}$ is all variables in the graph, and $\mathcal{C}(\mathcal{G})$ is the set of maximal cliques in the graph.

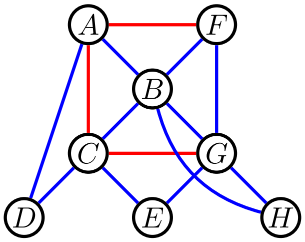
(a) Below is a table to explain the following independences.

Independence	**Answer**	**Explanation**
$A \perp\!\!\!\perp F \mid D$	False	A and F collide on B and D is a descendent of B therefore, A and F are not d-separated.
$A \perp\!\!\!\perp G \mid B, C$	False	A and F collide on B $\therefore$ $A \to B \leftarrow F$ is open. Furthermore, $F \to G$ is open, so $A \to B \leftarrow F \to G$ is open.
$G \perp\!\!\!\perp A \mid B, H, D, E, F$	False	The only way out of A is through D which collides with C so we have $A \to D \leftarrow C$. We also have that C and G collide with E, therefore we have $A \to D \leftarrow C \to E \leftarrow G$ is open.
$F \perp\!\!\!\perp D \mid A, B$	True	All chains out of F are blocked by B, and A and F don't collide $\because$ A is conditioned on, so the only open paths out of F are through G. There are no open paths from G to D since $C \to E \leftarrow G$ is closed.
$C \perp\!\!\!\perp H \mid B$	True	$C \leftarrow B \to G$ is closed $\because$ B is conditioned on and $C \to E \leftarrow G$ is closed because E is not conditioned on, therefore all paths from C to G are blocked.

(b) $(C \perp\!\!\!\perp A, F, G, H \mid B)$,$(H \perp\!\!\!\perp A, C, D, E, F \mid B, G)$,$(G \perp\!\!\!\perp A, C, D \mid B, F)$, $(D \perp\!\!\!\perp B, D, E, F, G, H \mid A, C)$, $(E \perp\!\!\!\perp A, B, D, F, H \mid C, G)$

(c) New local Markov property: $(C \perp\!\!\!\perp A, F, G, H \mid B)$, $(H \perp\!\!\!\perp A, C, D, E, F \mid B, G)$,$(\mathbf{G} \perp\!\!\!\perp \mathbf{A}, \mathbf{C}, \mathbf{D} \mid \mathbf{F})$, $(D \perp\!\!\!\perp B, D, E, F, G, H \mid A, C)$, $(E \perp\!\!\!\perp A, B, D, F, H \mid C, G)$. In the new graph, we have that $G \perp\!\!\!\perp_d A \mid F$, however in the previous graph, $A \to B \to G$ is open, so $G \not\perp\!\!\!\perp_d A \mid F$ in the old graph. So they represent different distributions.

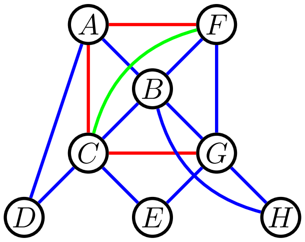
(d) The moralized graph is shown below. New edges are shown in red.



(e) The MRF factorization is

$$p(A,B,C,D,E,F,G,H) = \frac{1}{Z}\phi_{ABF}\phi_{ABC}\phi_{ACD}\phi_{CEG}\phi_{BCG}\phi_{BGH}\phi_{BFG}$$

(f) No, it is not since (A,F,C,G) is a 4-cycle that is not triangulated. Added edges are in green to moralize the graph.

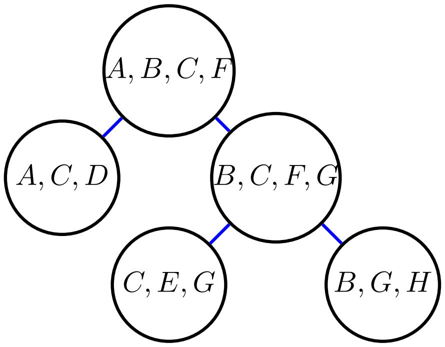


Maximal Cliques: $(A,B,C,F),(B,C,F,G),(A,C,D),(C,E,G),(B,G,H)$

The new factorization is:

$$p(A,B,C,D,E,F,G,H) = \frac{1}{Z}\phi_{ABCF}\phi_{BCFG}\phi_{ACD}\phi_{CEG}\phi_{BGH}$$

(g) The clique tree for the moralized graph is show below



(h) Visually, (A, B, C, F) appears like a fine root to me. The inbound messages are:

$$
\begin{aligned}
\phi_{AC}^{ACD\to ABCF} &= \sum_{D} \phi_{ACD}, \quad \phi_{CG}^{CEG\to BCFG} = \sum_{E} \phi_{CEG} \\
\phi_{BG}^{BGH\to BCFG} &= \sum_{H} \phi_{BGH} \\
\phi_{BCF}^{BCFG\to ABCF} &= \sum_{G} \phi_{BCFG}\phi_{CG}^{CEG\to BCFG}\phi_{BG}^{BGH\to BCFG}
\end{aligned}
$$

Outbound messages are:

$$
\begin{aligned}
\phi_{AC}^{ACD\leftarrow ABCF} &= \sum_{BF} \phi_{ABCF}\phi_{BCF}^{BCFG\to ABCF} \\
\phi_{BCF}^{BCFG\leftarrow ABCF} &= \sum_{A} \phi_{ABCF}\phi_{AC}^{ACD\to ABCF} \\
\phi_{CG}^{CEG\leftarrow BCFG} &= \sum_{BF} \phi_{BCFG}\phi_{BCF}^{BCFG\leftarrow ABCF}\phi_{BG}^{BGH\to BCFG} \\
\phi_{BG}^{BGH\leftarrow BCFG} &= \sum_{CF} \phi_{BCFG}\phi_{BCF}^{BCFG\leftarrow ABCF}\phi_{CG}^{CEG\to BCFG}
\end{aligned}
$$

(i) The marginal probabilities for each maximal clique are given by:

$$
\begin{aligned}
p(A, B, C, F) &= \frac{\phi_{ABCF}\phi_{AC}^{ACD\to ABCF}\phi_{BCF}^{BCFG\to ABCF}}{\sum_{ABCF}\phi_{ABCF}\phi_{AC}^{ACD\to ABCF}\phi_{BCF}^{BCFG\to ABCF}} \\
p(B, C, F, G) &= \frac{\phi_{BCFG}\phi_{BCF}^{BCFG\leftarrow ABCF}\phi_{CG}^{CEG\to BCFG}\phi_{BG}^{BGH\to BCFG}}{\sum_{BCFG}\phi_{BCFG}\phi_{BCF}^{BCFG\leftarrow ABCF}\phi_{CG}^{CEG\to BCFG}\phi_{BG}^{BGH\to BCFG}} \\
p(A, C, D) &= \frac{\phi_{ACD}\phi_{AC}^{ACD\leftarrow ABCF}}{\sum_{ACD}\phi_{ACD}\phi_{AC}^{ACD\leftarrow ABCF}} \\
p(C, E, G) &= \frac{\phi_{CEG}\phi_{CG}^{CEG\leftarrow BCFG}}{\sum_{CEG}\phi_{CEG}\phi_{CG}^{CEG\leftarrow BCFG}} \\
p(B, G, H) &= \frac{\phi_{BGH}\phi_{BG}^{BGH\leftarrow BCFG}}{\sum_{BGH}\phi_{BGH}\phi_{BG}^{BGH\leftarrow BCFG}}
\end{aligned}
$$

Using $\mathbf{V} = (A, B, C, D, E, F, G, H)$, and $C(\mathcal{G}) = \{\text{cliques from (f)}\}$, we have the following relationships which will resolve the marginal of the MRF factorization:

$$
\begin{aligned}
numerator(p(A, B, C, F)) &= \phi_{ABCF} \sum_{DEGH} \phi_{ACD}\phi_{BCFG}\phi_{CEG}\phi_{BGH} \;\& \\
numerator(p(B, C, F, G)) &= \phi_{BCFG} \sum_{ADEH} \phi_{ABCF}\phi_{ACD}\phi_{CEG}\phi_{BGH} \;\& \\
numerator(p(A, C, D)) &= \phi_{ACD} \sum_{BFGEH} \phi_{ABCF}\phi_{BCFG}\phi_{CEG}\phi_{BGH} \;\& \\
numerator(p(C, E, G)) &= \phi_{CEG} \sum_{BFADH} \phi_{BCFG}\phi_{ABCF}\phi_{ACD}\phi_{BGH} \;\& \\
numerator(p(B, G, H)) &= \phi_{BGH} \sum_{CFADE} \phi_{BCFG}\phi_{ABCF}\phi_{ACD}\phi_{CEG} \implies
\end{aligned}
$$

$$
p(\mathbf{c}) = \frac{\sum_{\mathbf{V}\setminus\mathbf{c}}\prod_{\mathbf{c}'\in C(\mathcal{G})}\phi_{\mathbf{c}'}}{\sum_{\mathbf{V}}\prod_{\mathbf{c}'\in C(\mathcal{G})}\phi_{\mathbf{c}'}}, \ \forall \mathbf{c} \in C(\mathcal{G})
$$

K-Means

(a) Is it possible to initialize the k-means algorithm in such a way that it fails to terminate successfully?

(b) Say our input to k-means is a set of $2k$ points with 2-coordinates arranged in line, e.g. with coordinates:

$$(0,0),(0,1),(0,2),\dots,(0,k),(0,k+1),\dots,(0,2k).$$

Say we initialize k-means with 2 clusters, with initial centroids given by $(0,k)$ and $(0,k+1)$. In many iterations will k-means terminate? What will be the final cluster assignments and centroids?

(a) No. The k means algorithm will always continue to find a local minimum, irrespective of the initializing points. That being said, k means is often performed with several different initialized starting points to find better, more optimal clusters.

(b) In this example, the k means algorithm would terminate after a single iteration, regardless of whether (0,0) is included or not (addressing the related Piazza post). To determine the final cluster assignments and centroids, we assume that we are referring to the scenario with 2k+1 points (including (0,0)). Here, the first cluster will include the points (0,0),(0,1),...,(0,k), and the centroid will be (0, $\frac{k}{2}$). The second cluster will include (0,k+1),...(0,2k), and its centroid would be (0, $\frac{3k+1}{2}$).