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

Assigned: Friday, November 01, 2024

Due: Friday, November 15, 2024, 11:59 pm US/Eastern

TREVOR BLACK (TBLACK20)

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 change any formatting in this document, or we may be unable to grade your work. This includes, but is not limited to, the font sizes, and the spacing of text. Additionally, do not add text outside of the answer boxes. Entering your answers are the only changes allowed.

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.

Markov Random Fields

Question 1. [15 pts] 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.

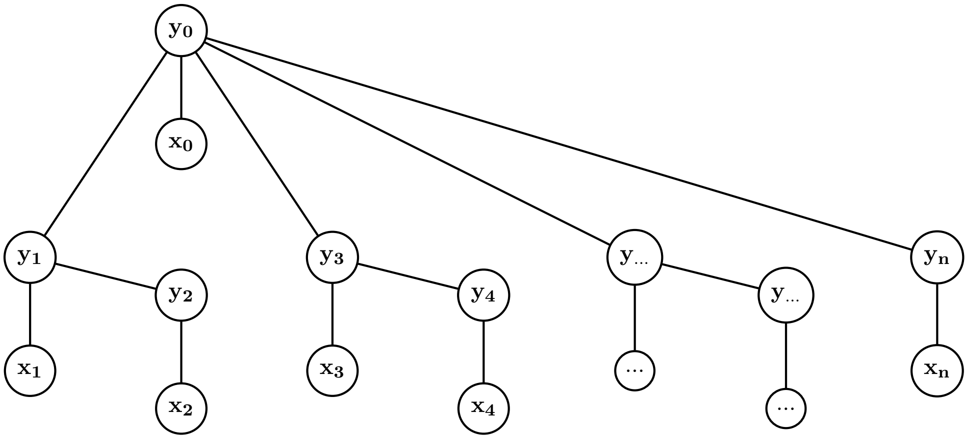


Figure 1: Tree structure model

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

$\prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{Z} \prod_{C \in C(G)} \psi_C(Z_C)$
$p(\mathbf{y}, \mathbf{x}) = \frac{1}{Z} \prod_{C \in C(G)} \psi_C(Z_C) = \frac{1}{Z} \prod_{C \in C(G)} \Sigma_i \exp\{\Sigma_i \theta_i f_i(Z_C)\}$
$\log p(\mathbf{y}, \mathbf{x}) = \Sigma_{C \in C(G)} \Sigma_i \theta_i f_i(Z_C) - \log(Z)$

Normalizing value: $Z = \int \prod_{C \in C(G)} \psi_C(c) d\mathbf{x}$

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

$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{y},\mathbf{x})}{p(\mathbf{x})}, p(\mathbf{x}) = \frac{\psi_X(\mathbf{x})}{\Sigma_{\mathbf{x}} \psi_X(\mathbf{x})} = \Sigma_{\mathbf{y}} p(\mathbf{y}, \mathbf{x})$
$\log p(\mathbf{y}|\mathbf{x}) = \log p(\mathbf{y}, \mathbf{x}) - \log \Sigma_{\mathbf{y}} p(\mathbf{y}, \mathbf{x})$
$= \Sigma_{C \in C(G)} \Sigma_i \theta_i f_i(Z_C) - \log \Sigma_{\mathbf{y}} \exp\{\Sigma_{C \in C(G)} \Sigma_i \theta_i f_i(Z_C)\}$

(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 $\log p(\mathbf{y}|\mathbf{x})$.

$\log p(\mathbf{y}, \mathbf{x})$ requires calculating the value for Z, which involves summing a value based on all combinations of $\mathbf{x}$ and $\mathbf{y}$. This combination space is defined by the sizes of $\mathbf{x}$ and $\mathbf{y}$, call it n, and the number of possible states, given as k and k'. It is $k^n \cdot (k')^n$. For large values of k' as specified in the problem, this value becomes infeasibly large. $\log p(\mathbf{y}|\mathbf{x})$ assumes $\mathbf{x}$ is fixed, so Z only has dependence on k^n, and is completely independent from the value of k'.

Question 2. [10 pts]

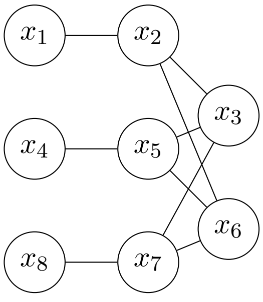
(a) Suppose you wanted to compute $S = \sum_{x_1=1}^{100} \cdots \sum_{x_8=1}^{100} h(x)$ where

$$h(x) = \exp(x_1 x_2 + x_4 x_5 + x_7 x_8) \prod_{i=2,5,7} (x_i + x_3 + x_6)^i.$$

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.

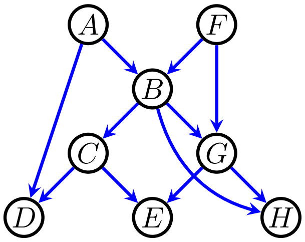
There are two parts to the function h. The pairwise summation in the exponent, as well as the exponential summation in the product. The pairwise summation will recompute the same values many times as $\exp(x_1x_2 + x_4x_5 + x_7x_8)$ can be rewritten as $\exp(x_1x_2)\exp(x_4x_5)\exp(x_7x_8)$. When computing this value for S, we loop through all combinations of the values for x_ix_j. Because all x_i have the same possible values, each term $\exp(x_ix_j)$ is the same combination of values, the values for all combinations have to only be computed once. This is $3 \cdot 100^2$ (3 terms, 100 values, pairwise) $= 3 \cdot 10^4$.
For the product term, values for x_3 and x_6 can be similarly precomputed without having to loop through them for each combination of the other values. There are 100^2 combinations (100 values, pairwise) or 10^4. Additionally, each of these combinations requires the addition of $x_i, i \in 2, 5, 7$ where $x_i \in [0, 100]$. Values can be precomputed for all $x_i \in [0, 100]$. $3 \cdot 100 \cdot 10^4$ (3 sums for x_2, x_5, x_7, 100 values for $x_i, i \in 2, 5, 7$, and 10^4 combinations of x_3 and x_6). This results in $3 \cdot 10^6$.
Combining these, we get $3 \cdot 10^4 + 3 \cdot 10^6$ which is less than 10^7.

(b) Draw the MRF associated with this distribution.



DAGs, Clique Trees and Message Passing.

Question 3. [45 pts]



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:

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

$A \perp\!\!\!\perp F \mid D$ is true as $A \to B \leftarrow F$ still is unblocked.
$A \perp\!\!\!\perp G \mid B, C$ is true as B is observed in the directed path $A \to B \to G$.
$G \perp\!\!\!\perp A \mid B, H, D, E, F$ is true for the same reason as above.
$F \perp\!\!\!\perp D \mid A, B$ is false as there is no way to reach D from F given A, B.
$C \perp\!\!\!\perp H \mid B$ is true as $C \leftarrow B \to H$ is unblocked with B observed.

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

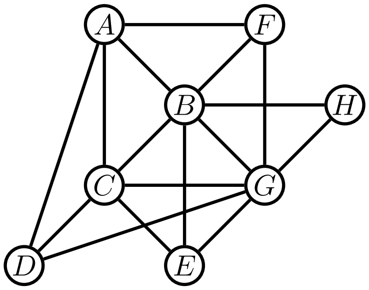
$A \perp\!\!\!\perp F$
$B \perp\!\!\!\perp \emptyset | A, F$
$C \perp\!\!\!\perp A, F, G, H | B$
$D \perp\!\!\!\perp B, E, F, G, H | A, C$
$E \perp\!\!\!\perp A, B, D, F, H | C, G$
$F \perp\!\!\!\perp A$
$G \perp\!\!\!\perp A, D, C, | B, F$
$H \perp\!\!\!\perp A, F, D, C, E | B, G$

(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.

The new graph is not the same as the old one as reversing the direction changes the dependencies of the nodes, thus changing the local Markov property of the model and how the d-separation queries are evaluated.

(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 $V_i \rightarrow V_k \leftarrow V_j$ 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 as few edges as possible to make it chordal. If you added edges, write the factorization of the new graph.

(g) Create a clique tree from the triangulated graph above (either $\mathcal{G}^a$ if it is chordal, or the graph obtained from $\mathcal{G}^a$ by adding new edge(s) to make it chordal).

(h) Pick a root $\mathbf{R}$ of the clique tree, and calculate both incoming messages $\phi^{\mathbf{S}_i \rightarrow \mathbf{S}_j}$ 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 $\mathbf{S}_1, \ldots, \mathbf{S}_m$ further from the root that

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

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

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

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