10-708 Probabilistic Graphical Models

Homework 1

Due Feb 10, 7:00 PM

Rules:

1. Homework is due on the due date at 7:00 PM. The homework should be submitted via Gradescope. Solution to each problem should start on a *new page* and marked appropriately on Gradescope. For policy on late submission, please see course website.
2. We recommend that you typeset your homework using appropriate software such as LaTeX. If you are writing, please make sure your homework is cleanly written up and legible. The TAs will not invest undue effort to decrypt bad handwriting.
3. **Code submission:** for programming questions, you must submit the complete source code of your implementation also via Gradescope. Remember to include a small README file and a script that would help us execute your code.
4. **Collaboration:** You are allowed to discuss the homework, but you should write up your own solution and code. Please indicate anyone you collaborated with in your submission.

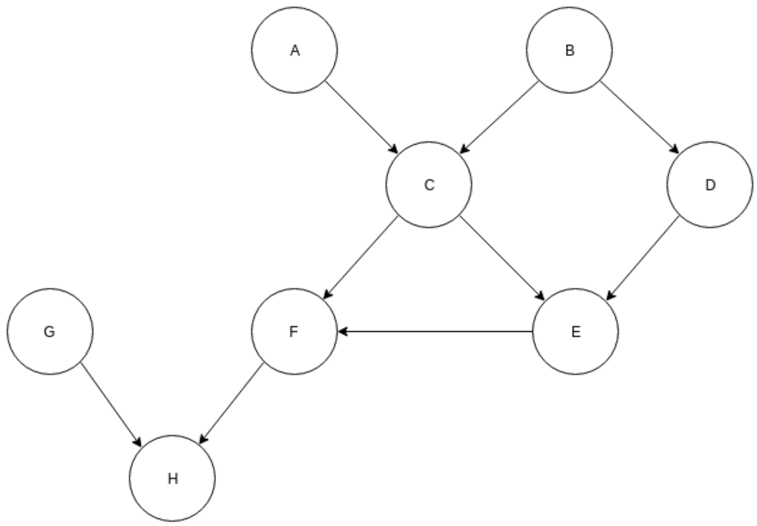

Figure 1: Figure for Problem 1.2

1 Bayesian Network [30 pts] (Haohan)

1.1 I-map (10 pts)

Let $\mathbb{G}$ be a Bayesian Network structure over a set of random variables $\mathbb{X}$, and let $\mathbb{P}$ be a joint distribution over the same space. If $\mathbb{G}$ is an I-map for $\mathbb{P}$, show that $\mathbb{P}$ factorizes according to $\mathbb{G}$.

1.2 D-seperation (10 pts)

According to Figure 1, determine whether the following claims to be True or False and justify your answer.

- $B \perp G \quad | \quad A$
- $C \perp D \quad | \quad F$
- $C \perp D \quad | \quad A$
- $H \perp B \quad | \quad C, F$
- There is at least one node in this BN, that all the other nodes are in its Markov Blanket.

1.3 I-Equivalence (10 pts)

I-equivalence of two graphs means that any distribution P that can be factorized over one of these graphs can be factorized over other. A relevant concept to describe I-equivalence is called **skeleton**. Formally, The **skeleton** of a Bayesian network graph G over X is an undirected graph over X that contains an edge $\{X, Y\}$ for every edge (X, Y) in G. Informally, just remove all the arrows of a Bayesian network, you will get the **skeleton** of it.

For two graphs G_1 and G_2, briefly justify the following arguments:

- G_1 and G_2 has the same skeleton is a necessary, but not sufficient condition for G_1 and G_2 to be I-equivalent.
- G_1 and G_2 has the same skeleton and same v-structures, is a sufficient, but not necessary condition for G_1 and G_2 to be I-equivalent.

2 Independence & Equivalence Testing (Maruan) [35 pts]

2.1 D-separation and Independence (15 pts)

In this exercise, you goal is to implement an algorithm that, for a given Bayesian network (BN) and a list of queries, can automatically test whether a given statement about conditional independence is encoded in the graph[1]. Your algorithm can either make use of the moralized BN, or inspect the active trails between the queried variables (the latter is known as *Bayes-ball* algorithm).

(5 pts) First, prove that D-separation implies conditional independence, or mathematically:

$$\text{dsep}(X, Y \mid Z) \implies X \perp\!\!\!\perp Y \mid Z.$$

(3 pts) [**bonus question**] Prove completeness of the D-separation property.

(2 pts) Provide a pseudo-code and describe in detail an algorithm for testing D-separation.

(8 pts) Implement an algorithm with the following input-output specification (read input from `stdin` and write output to `stdout`):

Input:

1. First line: `N M Q`.
   Here `N` and `M` are the number of nodes and edges in the BN, respectively, `Q` is the number of D-separation queries that will follow.
2. Next `M` lines: `A B`.
   Each line denotes a directed edge $A \to B$ in the BN graph.
3. Next `Q` lines: `A B | C D E ...`
   Each line denotes a query: whether `A` and `B` are D-separated given `C D E ...`

Output:
For each query, your algorithm should output `True` if the nodes are D-separated or `False` otherwise. Answer to each query should come on a separate line (`Q` output lines in total).

Testing: We provide you 4 simple test cases that you can use to test you algorithm. Your final score for this problem will be based on another 4 tests that we will run your submission on as well as your code. *Please make sure your code is readable and well organized.*

2.2 I-equivalence (20 pts)

In this part, we recall the notion of I-equivalence and ask you to (1) prove a criterion for the I-equivalence of a pair graphs and (2) implement an algorithm that can test whether a given pair of BNs is I-equivalent.

(4 pts) Prove the following proposition:

Proposition 1. *The two BNs, G and G', are I-equivalent if both graphs have the same set of trails and a trail is active in G if and only if it is active in G'.*

Hint: Use the notion of D-separation.

(8 pts) Now, you need to prove a criterion for I-equivalence (i.e., an "if and only if" statement). We will do this in two steps and will need the following definitions.

[1] Meaning that a distribution that factorizes over the given graph necessarily follows that conditional independence.

Definition 1 (Minimal Active Trail)**.** *Consider an active trail $T = X_1, X_2, \ldots, X_m$. This active trail is called* minimal *if no subset of the nodes in T of cardinality less than m forms an active trail between X_1 and X_m. Stated differently, T is minimal if no other active trail between X_1 and X_m "shortcuts" any of the nodes in T.*

Definition 2 (Triangle)**.** *Any three consecutive nodes in a trail $T = X_1, X_2, \ldots, X_m$ are called a* triangle *if their skeleton is fully connected (i.e., forms a 3-clique).*

(3 pts) Prove that a minimal active trail may only contain a triangle of the following form:

* $X_i \rightarrow X_{i-1}$
* $X_i \rightarrow X_{i+1}$
* Either $X_{i-1} \rightarrow X_{i+1}$, or $X_{i-1} \leftarrow X_{i+1}$

Hint: Prove by cases.

(5 pts) Prove the theorem that states the criterion.

Theorem 1. *Two Bayesian networks, G and G', are I-equivalent if and only if G and G' have the same skeletons and the same set of immoralities.*

Hint: Both directions of the criterion can be proved by contradiction.

(2 pts) Provide a pseudo-code and describe in detail an algorithm for testing I-equivalence between BNs.

(6 pts) Implement an algorithm with the following input-output specification (read input from `stdin` and write output to `stdout`):

Input:

1. Line: `N1 M1`.
   `N1` and `M1` are the number of nodes and edges, respectively, in the first Bayesian network.
2. Next `M1` lines: `A B`.
   Each line encodes a directed edge $A \rightarrow B$ in the first Bayesian network graph.
3. Line: `N2 M2`.
   `N2` and `M2` are the number of nodes and edges, respectively, in the second Bayesian network.
4. Next `M2` lines: `A B`.
   Each line encodes a directed edge $A \rightarrow B$ in the second Bayesian network graph.

Output: Return `True` if the graphs are I-equivalent and `False` otherwise.

Testing: We provide you 4 simple test cases that you can use to test you algorithm. Your final score for this problem will be based on another 4 tests that we will run your submission on as well as your code. *Please make sure your code is readable and well organized.*

3 Exact Inference (Haohan) [20 pts]

3.1 Variable Elimination (5 pts)

For a Hidden Markov Model with T time steps, whose hidden states are denoted as z and observable states are denoted as x, use Variable Elimination Algorithm to derive HMM's classical forward-backward algorithm to inference

$$p(z_t = j | x_{1,\dots T}).$$

You can start with

$$p(z_t = j | x_{1,\dots T}) \propto \alpha_t(j)\beta_t(j),$$

where $\alpha_t(j) = p(z_t = j | x_{1,\dots,t})$ and $\beta_t(j) = p(x_{t+1,\dots,T} | z_t = j)$

3.2 Gaussian Belief Propagation (15 pts)

Let's consider the belief propagation algorithm for Gaussian pairwise MRF, where the potentials are defined as following:

$$\phi_t(x_t) = exp(-\frac{1}{2}A_{tt}x_t^2 + b_t x_t)$$
$$\phi_{s,t} = exp(-\frac{1}{2}x_s A_{st} x_t)$$

where A is the precision matrix and b is just a parameter. Our goal is to derive the message passed into node x_t, i.e. $m(x_t)$.

1 (6 pts) To begin with, we need a fact that the product of two Gaussians is a scaled Gaussian. Show:

$$N(x|\mu_1, \lambda_1^{-1}) \times N(x|\mu_2, \lambda_2^{-1}) = CN(x|\mu, \lambda^{-1}),$$

solve for λ and μ, where C is a constant.

2 (9 pts) Solve for $m(x_t)$.
Hint: First solve for the messages passed into Node t's neighbor nodes from their neighbors excluding t, then use these messaged to represent $m(x_t)$.

You can directly use this result if needed: $\int \exp(-ax^2 + bx)dx = \sqrt{\frac{\pi}{a}} \exp(\frac{b^2}{4a})$

4 Undirected Graphical Models (Maruan) [15 pts]

4.1 Ising Model & Boltzmann Machine [8 pts]

Assume that we are given a graph, $G = (V, E)$, with the vertex set V and edges E. Recall from the lecture that an Ising model (IM) is given by the following distribution:

$$\mathbb{P}(x_1, \ldots, x_n) = \frac{1}{Z} \exp\left\{ \sum_{(i,j)\in E} W_{ij} x_i x_j - \sum_{i \in V} u_i x_i \right\},$$

where each random variable $X_i \in \{-1, 1\}$. The Boltzmann machine (BM) is defined similarly, where the only difference is in the domain of the random variables, $X_i \in \{0, 1\}$.

Show that the Ising model can be re-written as a Boltzmann machine. In particular, for an Ising model with parameters W and $\mathbf{u}$ specify new parameters W' and $\mathbf{u}'$ for a Boltzmann machine and prove that they give the same distribution, assuming that the states $\{-1, 1\}$ are mapped to $\{0, 1\}$.

4.2 Determinantal Point Process [7 pts]

A point process over a finite set $\mathcal{X} = \{x_1, \ldots, x_n\}$ defines a probability distribution over its subsets, $\mathbb{P}(\mathcal{Y} \subseteq \mathcal{X})$. Equivalently, we can think of such point process as a binary Markov random field (MRF) that defines a distribution over n indicator random variables, $Z_1, \ldots, Z_n$, $Z_i \in \{0, 1\}$, each of which corresponds to a unique element of $\mathcal{X}$. Therefore, we can always write the distribution defined by a point process over a finite set in the following form:

$$\mathbb{P}(z_1, \ldots, z_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(Z_c), \tag{1}$$

where $\psi_c(Z_c)$ denotes the potential corresponding to clique $c \in \mathcal{C}$.

Boltzmann machines and Ising models are *pairwise* MRFs, i.e., they only encode unary and binary potentials that correspond to $\mathbf{u}$ and W parameters, respectively. Both models have $n(n+1)/2$ parameters. Now consider the following distribution that defines the *determinantal point process (DPP)* [1]:

$$\mathbb{P}(\mathcal{Y} \subseteq \mathcal{X}) = \frac{\det(S_{\mathcal{Y}})}{\det(S + I)}, \tag{2}$$

where S is some positive semi-definite matrix, $S_{\mathcal{Y}}$ is its submatrix with rows and columns selected by the subset $\mathcal{Y}$, I is the $n \times n$ identity matrix, and $\det(\cdot)$ denotes the standard matrix determinant. Note that DPP is completely described by its positive semi-definite kernel matrix S, i.e., also by $n(n+1)/2$ parameters.

(2 pts) Give a representation of the DPP distribution (2) in the canonical exponential form as a binary MRF.

(7 pts) Is Boltzmann machine on n variables equivalent to a DPP on a set of n elements? In other words, for a DPP with some kernel matrix S, does there always exist W and $\mathbf{u}$ such that the corresponding Boltzmann distribution is equivalent to the one given by DPP? How about the opposite? (*Hint:* Analyze a simple case with 2 and 3 binary random variables).

References

[1] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.