ECE521: Inference Algorithms and Machine Learning
University of Toronto

Assignment 4:
Inference and Learning on Graphical Models

TA: Use Piazza for Q&A
Due date: Apr. 9 11:59 pm, 2017
Electronic submission to: ece521ta@gmail.com

General Note:

- In this assignment, you will derive learning and inference procedures for some of the probabilistic models described in class.
- Full points are given for complete solutions, including justifying the choices or assumptions you made to solve the question.
- Homework assignments are to be solved in the assigned groups of two. You are encouraged to discuss the assignment with other students, but you must solve it within your own group. Make sure to be closely involved in all aspects of the assignment.

1 Graphical Models [20 pt.]

1.1 Graphical models from factorization [6 pt.]

Consider a joint distribution that factors in the following form:

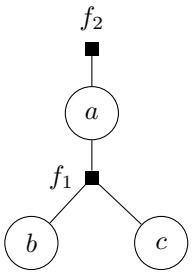
$$P(a, b, c, d, e, f) = P(a|b)P(b)P(c|a, b)P(d|b)P(e|c)P(f|b, e)$$

1. Sketch the corresponding Bayesian network (BN). [2 pt.]
2. Sketch the factor graph representation and label the factors with corresponding distributions. [2 pt.]
3. Sketch the Markov random field (MRF) representation and label all the maximum cliques with corresponding distributions. [2 pt.]

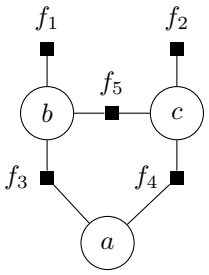
1.2 Conversion between graphical models [10 pt.]

1.2.1 [4 pt.]

a)



b)

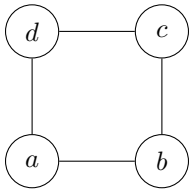


1. For both factor graphs (a) and (b), if it exists, sketch the equivalent BNs that implies the **same conditional independence properties** as the factor graphs and write down the conditional probabilities using the factors, $f_1, \ldots, f_5$. If it does not exist, explain why. [3 pt.]

2. For both factor graphs (a) and (b), if it exists, sketch the equivalent MRFs that implies the **same conditional independence properties** as the factor graphs and write down the maximum clique potentials using the factors, $f_1, \ldots, f_5$. If it does not exist, explain why.[3 pt.]

1.2.2 [4 pt.]

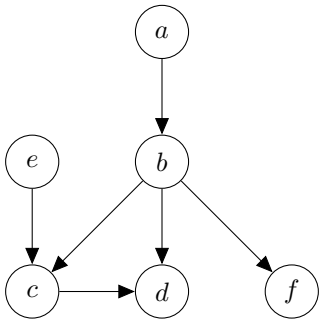
Consider the following MRF:



1. If it exists, sketch the equivalent factor graph representation that implies the **same conditional independence properties** as the MRF. If it does not exist, explain why. [2 pt.]

2. If it exists, sketch the equivalent BN that implies the **same conditional independence properties** as the MRF. If it does not exist, explain why. [2 pt.]

1.3 Conditional Independence in Bayesian Networks [4 pt.]

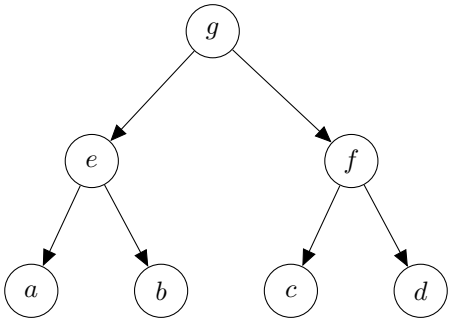


1. Express the joint probability $P(a, b, c, d, e, f)$ in a factorized form corresponding to the BN shown. [1 pt.]

2. Determine whether each of the followings statements is TRUE or FALSE and provide your explanations: $a \perp\!\!\!\perp c$? $a \perp\!\!\!\perp c|b$? $e \perp\!\!\!\perp b$? $e \perp\!\!\!\perp b|c$? $a \perp\!\!\!\perp e$? $a \perp\!\!\!\perp e|c$? [3 pt.]

2 Message-Passing [practice]

Consider a tree structure BN below, which represents a crude statistical model of genes mutations from the parent nodes to the children nodes.



We will focus on the probability of observing two specific genes. The presences of the two genes are denoted using a pair of binary random variables. Thus, the sample space or the set of the possible observations is $\{00, 01, 10, 11\}$ for all the random variables node in the BN. Let $P(g = 00) = 0.5$ and $P(g = 11) = 0.5$. Each gene has 10% chances of changing its state when passed down from a parent node. For example, $P(f = 11 \,|\, g = 00) = 0.01$ and $P(f = 01 \,|\, g = 11) = 0.09$.

1. Sketch the factor graph representing the BN. [practice]

2. Calculate the numerical values of the probabilities $P(e)$ and $P(e|a = 01)$ using sum-product message-passing rules. Show the intermediate steps for computing the local messages. [practice]

3 Hidden Markov Models [14 pt.]

Hidden Markov Model (HMM) is a class of probabilistic generative models for sequence data. Similar to Mixtures of Gaussians, HMM also has a set of latent mixture components. In addition, the latent states in HMM evolve over time to capture the temporal dependencies in the observed sequences. Consider a dataset of N sequences of observations and each has length D. For the n^{th} data point, the observed variable $x^{(n)} = \{x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \cdots, x_D^{(n)}\}$ is a sequence of D elements. HMM uses the latent variables $z_t^{(n)}$ to represent the hidden state assignments for each time t. An important difference between HMMs and MoGs is that the latent states in HMMs are related through shared transition probabilities, $P(z_t^{(n)} \mid z_{t-1}^{(n)})$ for each time t. For the n^{th} training sequence in the dataset, the HMM represents the factorized joint distribution: $P(x_1^{(n)}, x_2^{(n)}, \cdots, x_D^{(n)}, z_1^{(n)}, z_2^{(n)}, \cdots, z_D^{(n)}) = P(z_1^{(n)}) \prod_{t=2}^{D} P(z_t^{(n)} \mid z_{t-1}^{(n)}) \prod_{t=1}^{D} P(x_t^{(n)} \mid z_t^{(n)})$. We will derive the learning and the inference algorithms for HMMs in the following sub-sections.

3.1 Factor graph representation [2 pt.]

1. Sketch a BN representing a HMM of over five observed variables in a sequence $\{x_1, x_2, x_3, x_4, x_5\}$. Label the latent state variables $\{z_1, z_2, z_3, z_4, z_5\}$. [1 pt.]

2. Sketch the factor graph representation for the BN above. Annotate each of the factors using either the prior $P(z_1)$, the transition probabilities $P(z_t \mid z_{t-1})$ and the likelihoods $P(x_t \mid z_t)$. [1 pt.]

3.2 Inference by passing messages [2 pt.]

We define the factor between variable nodes a and b as $f_{ab}(a, b)$, e.g. the factor between x_3 and z_3 is $f_{x_3 z_3}(x_3, z_3)$. In a message-passing scheme, messages from node c to node d are written as $\mu_{c \to d}$. The local messages of a node in a factor graph include all of its incoming and outgoing messages from the neighboring nodes. We continue the example from Section 3.1 and derive the sum-product algorithm for the HMM:

1. Write down the message-passing rule to compute the message from the variable node z_4 to the factor node $f_{z_3 z_4}$, that is $\mu_{z_4 \to f_{z_3 z_4}}(z_4)$, in terms of the other local messages at z_4: $\mu_{z_4 \to f_{z_4 z_5}}(z_4), \mu_{f_{z_4 z_5} \to z_4}(z_4), \mu_{z_4 \to f_{x_4 z_4}}(z_4), \mu_{f_{x_4 z_4} \to z_4}(z_4), \mu_{f_{z_3 z_4} \to z_4}(z_4)$ [2 pt.]

2. Write down the message-passing rule to compute the posterior distribution $P(z_3 \mid x_1, x_2, x_3, x_4, x_5)$ in terms of the local messages at z_3. [practice]

3. We expand the HMM to a new variable x_6 in the sequence that has not been observed yet. x_6 has its latent state z_6. We can make predictions about x_6 given the past observations by computing the predictive distribution $P(x_6 \mid x_1, x_2, x_3, x_4, x_5)$. Write down the expression for the predictive distribution in terms of $\mu_{f_{z_3 z_4 \to z_4}}(z_4)$ and the sums and the products of other local messages. [practice]

3.3 Message-passing as bi-direction RNNs [10 pt.]

Consider an HMM with K latent states and observed variables may take on M discrete values $\{1, 2, \cdots, M\}$. Under a particular latent state $z_t = k$, each observation value may occur with $P(x_t = m \mid z_t = k)$. We represent the likelihood function $P(x_t \mid z_t)$ concisely in a $M \times K$ matrix W, where $P(x_t = m \mid z_t = k)$ is the element on the m^{th} row and the k^{th} column. Furthermore, we define a transition matrix $T \in \mathbb{R}^{K \times K}$ contains the transition probability $P(z_t = i \mid z_{t-1} = j)$ on its i^{th} row and j^{th} column. The prior distribution over the initial latent state z_1 is then defined as a vector $\pi = [P(z_1 = 1), \cdots, P(z_1 = K)]^\top$. Notice that the same matrices W and T are used for all time t in the sequence $\{x_t\}$, $\{z_t\}$.

In a factor graph without loops, there are two messages at each edge of the graph. They are sent by the nodes connected by the edge. In order to compute all the messages, a message-passing algorithm has to visit each node twice in both forward and backward directions. In particular, running message-passing algorithms on HMMs can be thought of as running a bi-directional recurrent neural network (bi-RNN) from both ends of the sequence, where W are the bottom up input weights and T are the recurrent weights in an RNN with linear hidden units.

1. We continue the example from Section 3.1. Assume the observed variables x_t are one-hot vectors. Write down the expression to compute a vectorized message $\mu_{f_{z_2 z_3} \to z_3} = [\mu_{f_{z_2 z_3} \to z_3}(z_3 = 1), \cdots, \mu_{f_{z_2 z_3} \to z_3}(z_3 = K)]^\top$ in terms of x_1, x_2, W, T, π. [5 pt.]
2. Write down the expression for $\mu_{z_3 \to f_{z_2 z_3}}$ in terms of x_3, x_4, x_5, W, T, π. [5 pt.]