EECS
KTH
Probabilistic Graphical Models DD2420
Exam 14:00-19:00 March 10, 2020

Aids: None, no books, no notes, nor calculators
Observe:

- Name and person number on every page
- Answers should be in English or Swedish
- Only write on one side of the sheets
- Specify the total number of handed in pages on the cover
- Be careful to label each answer with the question number and letter
- All questions should be answered briefly but do motivate your answer and clearly state any additional assumptions you may need to make.

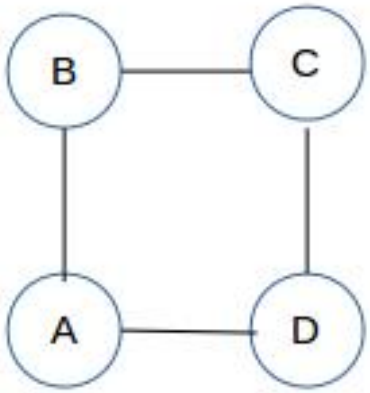
Responsible: John Folkesson, 08-790-6201

Part A (40 points - passing is 50%) Approximate Inference

1. Loopy Belief Propagation (8 points):

a) The running intersection property in clique trees ensures information from a random variable can reach all relevant cliques. How is this modified for loopy cluster graphs? Give an intuitive explanation of why this modification is needed? (4p)

b) A cluster graph with loops can not run exact inference using message passing.



Draw the cluster graph for the above undirected network (MRF) where each maximal clique in the graph forms a cluster. Indicate the sepsets on each edge. Why can we not just remove one edge from the cluster graph to break the loop and do inference using exact message passing on the clique tree? (2 p)

c) The 'calibration' of a clique tree assures us that two clique beliefs connected by an edge in the tree will agree on the marginals of their intersection set. Describe how this gets weakened for loopy cluster graph belief propagation. (2p)

2. Variational methods (12 points):

$p(\mathbf{Z} \mid \mathbf{X} = \mathbf{x})$ is a distribution with $\mathbf{x}$ being some observed data and $\mathbf{Z}$ a vector of hidden (latent) random variables. The joint distribution is $p(\mathbf{Z}, \mathbf{X})$ and the prior on $\mathbf{X}$ is $p(\mathbf{X})$. $q(\mathbf{Z})$ is our variational approximation (and should not be in any way confused with some marginal of $p(\mathbf{Z}, \mathbf{X})$). We will simplify the notation to replace $\mathbf{X} = \mathbf{x}$ with $\mathbf{x}$.

a) Show that

$$\mathbf{D}_{KL}(q(\mathbf{Z}) \mid\mid p(\mathbf{Z} \mid \mathbf{x})) = ln(p(\mathbf{x})) - \mathcal{L}(q).$$

Where $\mathcal{L}(q)$ is the so called variational lower bound (aka ELBO) expressed in terms of the two distributions $q(\mathbf{Z})$ and $p(\mathbf{Z}, \mathbf{x})$). This can be done in four lines starting from the definition of the Kullback-Liebler divergence. (8p)

b) Describe how one might use (a) to find better approximations of $p(\mathbf{z} \mid \mathbf{x})$ by q($\mathbf{z}$). That is conceptually how could you lower the Kullback-Liebler divergence between q and p. (2p)

c) Now assume that the data and latent variables are one dimensional (i.e. there is a single data point and it is scalar). Also
$p(Z, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(Z-x)^2}{2}}$
$q(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(Z-\alpha)^2}{2}}$

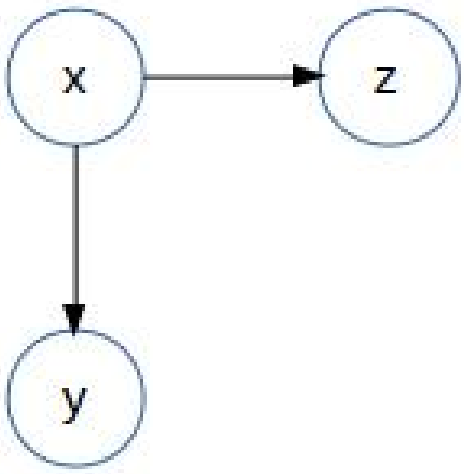
You will need to notice that these are both normal distributions with respect to z with variance of 1.0.

Show the use of the variational reasoning as described in (b) to determine what value of α will best fit to $p(Z \mid x)$. (2p)

3. Sampling based Approximation (10 points):

a) One wants to compute the expectation value of some function of random varibles. Explain why and how sampling from the distribution to compute an expectation value might be done instead of analytically computing it. (4p)

Use this Bayesian Net model for the following questions:



x	P(x)
0	.75
1	.25

$P(y \mid x)$	y=0	y=1
x=0	.42	.58
x=1	.21	.79

$P(z \mid x)$	z=0	z=1
x=0	.91	.09
x=1	.73	.27

b) Draw 6 samples of x given the following numbers drawn uniformly from the interval (0,1): $\{0.9, 0.2, 0.6, 0.8, 0.1, 0.5\}$ (2p)

c) We have evidence that $y = 1$. Use the samples of (x) to form sample pairs of (x, y) with rejection sampling given the evidence and this list of uniform random numbers: $\{0.5, 0.4, 0.3, 0.2, 0.5, 0.3, 0.9, 0.1, 0.6, 0.5, 0.3, 0.8\}$ (2p)

d) If z=1 what is the normalized importance weight of each of your samples from (b)? (2p)

4. Markov Chain Monte Carlo Methods, MCMC (10 points)

a) What is the Markov transition probability (also called the kernel) and how is it used in MCMC? (2p)

b) Show that the detailed balanced condition implies the stationary condition in MCMC. (2p)

c) How might 'mixing' and 'burn in time' be related? Be sure that you convince me that you understand the meaning of the two terms. (4p)

d) Why is the stationary condition on the MCMC kernel not enough and what additional condition would be enough? (2p)

Part B (20 points - passing is 50%) Learning

5. Estimation (10 points)

   a) What is the difference between the maximum likelihood estimate, MLE, and the maximum a posteriori, MAP, estimate of model parameters given some data? (2p)

   b) Derive the MLE of the binomial distribution's one parameter given the outcome of some trials? (2p)

   c) How does the conjugate prior simplify computing the full predictive distribution given some data? (2p)

   d) What would be the a conjugate prior distribution for (b)? (2p)

   e) Show that the distribution you named in (d) is indeed the conjugate prior of the binomial distribution? (2p)

6. Partially Observed Data (10 points)

   a) Give an example of 'missing completely at random', MCAR, and an example that is not MCAR. (2p)

   b) Give an example of 'missing at random', MAR, that is not MCAR. (2p)

   c) What is good about data that is MAR compared to data that is not, in regards to parameter estimation? (2p)

d) How might we proceed if we are trying to estimate the conditional probability table values for the Bayes Nets model below using observed data tuples (x_i, y_i, z_i, w_i)? Where i runs from 1 to m, the number of data tuples. A complication is that each tuple has one of the four varibles missing at random. Which variable is missing is random each time. All the variables are binary, i.e they can be either 0 or 1 (or ?). Hint EM and sufficient statistics are useful here. (4p)