EECS
KTH
Probabilistic Graphical Models DD2420
Exam 08:00-12:00 March 13, 2019

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) Approximate Inference

1. Loopy Belief Propagation (7 points):

   a) In a cluster graph with loops, how and where does one start and end the message passing? (2 p)

   b) What is the generalization of the running intersection property? (2p)

   c) How are the cluster beliefs and the marginal probabilities in a calibrated loopy graph related? (1p)

   d) Write out an expression for the unnormalized joint in terms of the cluster beliefs and sepset beliefs, that is invariant under message passing. (2p)

2. Variational methods (14 points):
   $p(\mathbf{z} \mid \mathbf{x})$ and $q(\mathbf{z})$ are two distribution with $\mathbf{x}$ being some data and $\mathbf{z} = (z_1, z_2, z_3)$ a set of hidden (latent) variables.

   a) Write out the expression for the entropy of $q(\mathbf{z})$. (2p)

   b) How does the entropy of a Gaussian distribution depend on its parameters? (2p)

   c) What is the relative entropy (also called the Kullback-Leibler Divergence) between q and p, $\mathbf{D}_{KL}(q(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{x})) =$? (2p)

   d) Define the M and I projections of p onto q in terms of their relative entropy. (3p)

   e) Write out an expression to show the relationship between our data likelihood, the evidence lower bound (ELBO) and a relative entropy. (2p)

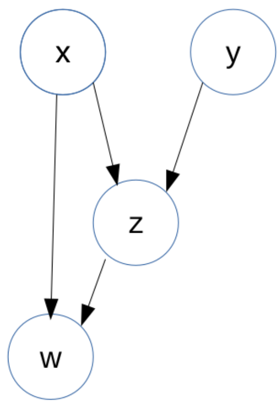
   f) When we use a mean field approximation for q, what is its form? (2p)

   g) Using the coordinate ascent (one variable at a time) method to optimize the mean field approximation, as in question (f), what would be the result of the step of optimizing with respect to the distribution of z_2? (a formula or good conceptual statment is needed) (1p)

3. Sampling based Approximation (9 points):

a) What is the 'Monte Carlo Principle', that is how would we express the expectation value of a function $f(\mathbf{x})$ using it? (2p)

Use this Bayesian Net model for the following questions:



x	P(x)
0	.3
1	.7

y	P(y)
0	.5
1	.5

$P(z=0 \mid x,y)$	y=0	y=1
x=0	.1	.9
x=1	.2	.6

$P(z=1 \mid x,y)$	y=0	y=1
x=0	.9	.1
x=1	.8	.4

$P(w=0 \mid x,z)$	z=0	z=1
x=0	.1	.2
x=1	.1	0

$P(w=1 \mid x,z)$	z=0	z=1
x=0	.9	.8
x=1	.9	1

b) Draw 4 samples of x given the following numbers drawn uniformly from the interval (0,1): $\{0.2, 0.6, 0.8, 0.1\}$ and 4 samples of y given similar uniform random numbers: $\{0.6, 0.9, 0.1, 0.4\}$ (3p)

c) We have evidence that $z = 1$. Use the samples of (x, y) to form samples of (x, y, z) with rejection sampling given the evidence and uniform random numbers $\{0.5, 0.4, 0.3, 0.2, 0.5, 0.3\}$ (2p)

d) If w=0 what is the normalized importance weight of each of your samples? (2p)

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

   a) Describe the general concept of MCMC methods including the two parts of the name. (2p)

   b) If we are doing inference on a probability distribution $P(\mathbf{x})$ state the stationary and detailed balance equations for the kernel. (3p)

   c) If we want to use some arbitrary kernel $Q(\mathbf{x} \mid \mathbf{x}')$ and the Metropolis-Hastings algorithm, how would we form a proper MCMC kernel for $P(\mathbf{x})$ from it? (2p)

   d) If our target distribution is given by $p(x_1, x_2, x_3, ..., x_n)$ what would Gibbs sampling look like? (2p)

   e) Explain the concept of mixing and why it is important in MCMC. (1p)

Part B (20 points) Learning

5. Maximum Likelihood Estimation (8 points)

   $\mathbf{X}$ is a random variable for which we have m i.i.d. observations and a parameterized model of its probability with parameters θ:

   a) What is the log-likelihood function for θ? (2p)

   b) What is the maximum likelihood estimate, MLE, for θ? (2p)

   c) If $\mathbf{X}$ is a scalar (1 dimensional) discrete random variable that can take on 3 possible values and we model its distribution using a multinomial distribution what would be the MLE of parameters and what are the sufficient statistics? (2p)

   d) If $\mathbf{X}$ is a scalar continuous random variable modeled using a Gaussian distribution what would be the MLE of the parameters and what are the sufficient statistics? (2p)

6. Bayesian Parameter Estimation (8 points)

   Again, $\mathbf{X}$ is a random variable for which we have m i.i.d. observations and a parameterized model of its probability with parameters θ:

   a) Write down an expression for the predictive distribution for new data $\mathbf{x}$ using Bayesian parameter estimation. (2p)

   b) What is meant by the term conjugate prior? (2p)

   c) Give two examples of a conjugate prior. (4p)

7. Partially Observed Data (4 points)

   a) Define and contrast the concepts of missing at random, MAR and missing completely at random, MCAR. (3p)

   b) Define the concept of identifiability. (1p)