NATIONAL UNIVERSITY OF SINGAPORE

CS5340 - Uncertainty Modelling in AI

(Semester 2 AY2018/19)

Time Allowed: 2 Hours

Instructions

- Write your Student Number below, and on every odd page. Do not write your name.
- The assessment contains 4 multi-part problems. You have 120 minutes to earn 120 points.
- The assessment contains 30 pages, including this cover page and 7 pages of scratch paper.
- The assessment is closed book. You may bring one double-sided sheet of A4 paper to the assessment. You may not use your mobile phone or a programmable calculator.
- Write your solutions in the space (box) provided. If you need more space, please use the scratch paper. Do not put part of the answer to one problem on a page for another problem.
- Read each question *carefully*. Don't get stuck on any one problem. The questions are *not* in any particular order of difficulty.
- Show your work. Partial credit will be given. Please be neat; we cannot grade unreadable solutions.
- Don't panic. The problems often look more difficult than they really are.
- Good luck!

Student Number.: ______________________________

Problem #	**Name**	**Possible Points**	**Achieved Points**
1	Yes or No?	20	
2	Red or Blue?	30	
3	Happy or Sad?	45	
4	Fast or Slow?	25	
Total:		120	

Common Probability Distributions

Distribution (Parameters)	PDF/PMF
Normal (μ, σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$
Bernoulli (r)	$r^x(1-r)^{(1-x)}$
Categorical (π)	$\prod_{k=1}^{K} \pi_k^{x_k}$
Binomial (μ, N)	$\binom{N}{x}\mu^x(1-\mu)^{N-x}$
Poisson (λ)	$\frac{\lambda^x \exp[-\lambda]}{x!}$
Beta (α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Gamma (a, b)	$\frac{1}{\Gamma(a)} b^a x^{a-1} \exp[-bx]$
Dirichlet $(\boldsymbol{\alpha})$	$\frac{\Gamma(\sum_k^K \alpha_k)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)} \prod_{k=1}^{K} x_k^{\alpha_k-1}$
Multivariate Normal $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\frac{1}{(2\pi)^{D/2}\lvert\Sigma\rvert^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$
Uniform (a, b)	$\frac{1}{b-a}$

Note: $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$ is the Gamma function.

Problem 1. Yes or No? [20 points]
For each of the questions below, give a **Yes or No** answer and a *brief explanation* justifying your answer.

Problem 1.a. [5 points] Recall that we can fit a parameterized distribution q_θ to a given distribution p by minimizing the KL divergence between q_θ and p,

$$\mathbb{D}_{\mathrm{KL}}[q_\theta \| p] = \int q_\theta(x) \log \frac{q_\theta(x)}{p(x)} dx.$$

Note that the "reversed" KL divergence,

$$\mathbb{D}_{\mathrm{KL}}[p \| q_\theta] = \int p(x) \log \frac{p(x)}{q_\theta(x)} dx$$

takes on a different form (the expectation is taken with respect to $p(x)$ instead of $q_\theta(x)$).

Assume that you are given an initial θ which is *not* the optimal θ^*. Is the following statement true: the solution q_θ obtained by minimizing $\mathbb{D}_{\mathrm{KL}}[q_\theta \| p]$ can *never* be equivalent to the solution obtained by minimizing $\mathbb{D}_{\mathrm{KL}}[p \| q_\theta]$?

Your Answer:

Brief Explanation:

Problem 1.b. [5 points] You are given an oracle that can tell you the treewidth of a Bayesian network (with discrete random variables) and the *optimal* elimination ordering in $O(1)$ time. Then, can you obtain the individual marginal distributions for all the random variables in polynomial time with respect to n (that is, $O(n^d)$) using variable elimination?

Note: n is the number of discrete random variables and d is some constant.

Your Answer:

Brief Explanation:

Problem 1.c. [5 points] Consider the Gaussian Mixture Model (GMM) with observed variables $\mathcal{X}$ and latent variables $\mathcal{Z}$. For independent and identically distributed (*iid*) data $\mathcal{D} = \mathcal{X}$ where $\mathcal{X} = (x_n)_{n=1}^{N}$, is it true that the posterior probability is decomposable, i.e.:

$$p(\mathcal{Z}|\mathcal{X}, \theta) = \prod_{n}^{N} p(z_n|x_n, \theta)?$$

Your Answer:

Brief Explanation: (if yes, provide a proof. If no, provide a counterexample)

Problem 1.d. [5 points] Consider the following transition matrix T for a Markov chain where

$$T = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.1 & 0.2 & 0.7 \\ 0 & 0 & 1.0 \end{bmatrix} \tag{1}$$

Is the transition matrix T *ergodic*?

Your Answer:

Brief Explanation:

Problem 2. Red or Blue? [30 points] You are given the following Bayesian network:

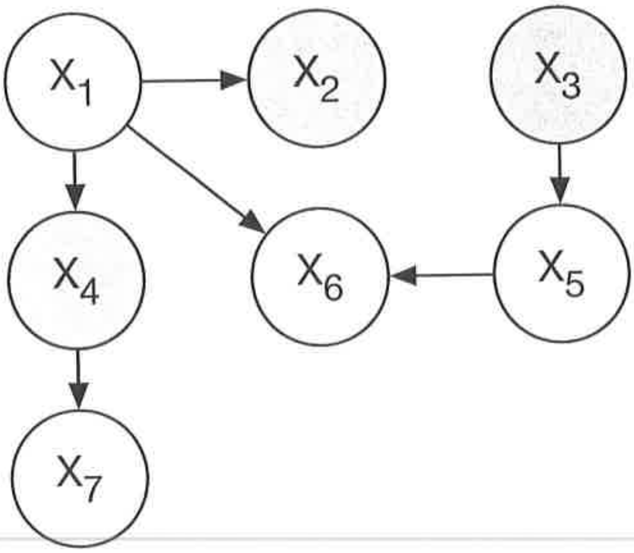


Figure 1: Red-Blue Bayesian Network.

where each random variable $X_1, X_2, \ldots X_7$ takes on either one of two states: RED or BLUE. The (conditional) probability tables associated with the Bayesian network are:

x_1	$p(x_1)$
RED	0.1
BLUE	0.9

x_1	x_2	$p(x_2 \mid x_1)$
RED	RED	0.8
BLUE	RED	0.2
RED	BLUE	0.2
BLUE	BLUE	0.8

x_3	$p(x_3)$
RED	0.8
BLUE	0.2

x_1	x_4	$p(x_4 \mid x_1)$
RED	RED	0.2
BLUE	RED	0.8
RED	BLUE	0.8
BLUE	BLUE	0.2

x_1	x_5	x_6	$p(x_6 \mid x_1, x_5)$
RED	RED	RED	0.1
RED	RED	BLUE	0.9
RED	BLUE	RED	0.9
RED	BLUE	BLUE	0.1
BLUE	RED	RED	0.9
BLUE	RED	BLUE	0.1
BLUE	BLUE	RED	0.9
BLUE	BLUE	BLUE	0.1

x_3	x_5	$p(x_5 \mid x_3)$
RED	RED	0.8
BLUE	RED	0.2
RED	BLUE	0.2
BLUE	BLUE	0.8

x_4	x_7	$p(x_7 \mid x_4)$
RED	RED	0.8
BLUE	RED	0.2
RED	BLUE	0.2
BLUE	BLUE	0.8

Problem 2.a. [8 points] Using the information provided, compute the probability

$$p(X_1 = \textsc{Red} | X_2 = \textsc{Red}, X_3 = \textsc{Blue}, X_4 = \textsc{Red})$$

Your Answer:

Brief Explanation: (provide your working/justification below. Use the scratch sheets if needed)

Problem 2.b. [12 points] Using the information provided, compute the maximum a posteriori (MAP) configuration for (X_1, X_5, X_6, X_7) given that we observe $X_2 = \text{RED}, X_3 = \text{BLUE}$ and $X_4 = \text{RED}$?

Your Answer:

$X_1 = \boxed{}$ $X_5 = \boxed{}$ $X_6 = \boxed{}$ $X_7 = \boxed{}$

Brief Explanation: (provide your working/justification below. Use the scratch sheets if needed)

Problem 2.c. [10 points] Derive the most efficient algorithm you can think of that returns TRUE if each and every possible configuration (given the evidence) has probability that exceeds a given threshold t ?

For example, the algorithm should return TRUE if for all possible configurations (X_1, X_5, X_6, X_7),

$$p(X_1 = \text{RED}, X_5 = \text{RED}, X_6 = \text{RED}, X_7 = \text{RED} | X_2 = \text{RED}, X_3 = \text{BLUE}, X_4 = \text{RED}) > t$$
$$p(X_1 = \text{RED}, X_5 = \text{RED}, X_6 = \text{RED}, X_7 = \text{BLUE} | X_2 = \text{RED}, X_3 = \text{BLUE}, X_4 = \text{RED}) > t$$
$$\vdots$$
$$p(X_1 = \text{BLUE}, X_5 = \text{BLUE}, X_6 = \text{BLUE}, X_7 = \text{BLUE} | X_2 = \text{RED}, X_3 = \text{BLUE}, X_4 = \text{RED}) > t$$

It should return FALSE if the probability for *any* of the possible configurations is less than (or equal to) t. Give a general algorithm that could work on *any* valid Bayesian network and evidence set.

Your Answer:

(More space on next page...)

(*Continued for Problem 2.c.*)

Problem 3. Happy or Sad? [45 points] Congratulations, you've just landed a new job at Foogle — the hottest AI startup in town! Your first task is to model data generated by Foogle's new EmoSense sensor. EmoSense is designed to detect whether a given person is HAPPY or SAD.

EmoSense selects a real-valued reading (Y) between either W or V, depending on the person's HAPPY/SAD state (Z). If a person is HAPPY, Emosense picks W, which follows a normal distribution with mean μ_w and variance σ_w^2. If a person is SAD, Emosense picks V, which also follows a normal distribution, but with mean μ_v and variance σ_v^2.

The Emosense model is partially summarized by the following graphical model for N *iid* data points:

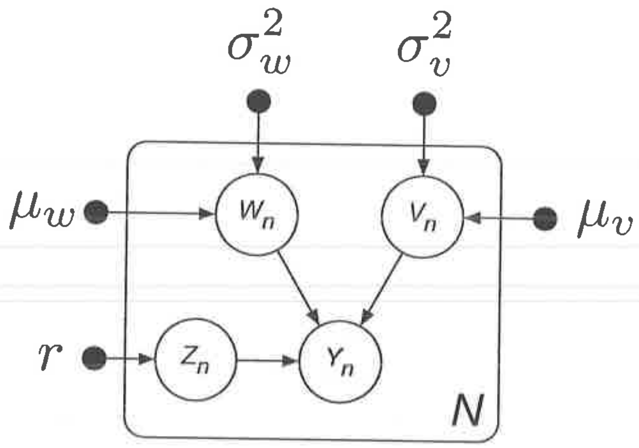


Figure 2: EmoSense Bayesian Network.

along with the distributions:

$$p(z_n|r) = \text{Bern}(r) \tag{2}$$
$$p(w_n|\mu_w, \sigma_w^2) = \mathcal{N}(w_n|\mu_w, \sigma_w^2) \tag{3}$$
$$p(v_n|\mu_v, \sigma_v^2) = \mathcal{N}(v_n|\mu_v, \sigma_v^2) \tag{4}$$

The parameters of the model are $\theta = \{r, \mu_w, \sigma_w^2, \mu_v, \sigma_v^2\}$.

Problem 3.a. [2 points] Given the description of how EmoSense works, derive the conditional distribution $p(y_n|z_n, w_n, v_n, \theta)$. You may introduce new parameter definitions if it makes your model description more succinct.

Your Answer:

Problem 3.b. [5 points] Derive the joint distribution $p(y, z, w, v|\theta) = \prod_n^N p(y_n, z_n, w_n, v_n|\theta)$.

Your Answer:

Problem 3.c. [8 points] Only y_n's are observed and you want to learn the unknown parameters θ using Expectation-Maximization (EM). Given the information provided, write down the function $Q(\theta, \theta_t)$ that is maximized during the M step.

Note: Be precise; an acceptable solution is specific to this subproblem. Expectations can be left unresolved *only if* the expectation is for one of the distributions listed on page 2.

Your Answer (the Q Function):

Brief Explanation: (provide your working/justification below. Use the scratch sheets if needed)

(More space on next page...)

(*Continued for Problem 3.c.*)

Problem 3.d. [7 points] Your boss at Foogle thinks your EM estimates may overfit the data. She wants you to infer the latent parameters using *Variational Bayes* instead. For this sub-problem, assume that σ_w^2 and σ_v^2 are *known*, i.e., you do not have to infer them.

Introduce random variables for each of the remaining parameters in θ and state the *conjugate* prior distributions. Draw out the resulting graphical model.

Your Answer: (draw the graphical model and write down each of the prior distributions)

Problem 3.e. [15 points] Assume a mean-field variational distribution. Write down the resulting evidence lower bound (ELBO). For this sub-problem, assume that σ_w^2 and σ_v^2 are *known*, i.e., you do not have to infer them. You are free to select a desired q or derive it from the independence assumptions. Again, you may re-parameterize the model if re-parameterization makes the model more succinct.

Note: Be precise; an acceptable solution is specific to this subproblem. If you have an expectation $\mathbb{E}[p(x)]$ or $\mathbb{E}[\log p(x)]$, and $p(x)$ is one of distributions on Page 2, you can leave the expectation unresolved.

Your Answer (the ELBO):

(More space on next page...)

(*Continued for Problem 3.e.*)

Problem 3.f. [8 points] For this sub-problem assume that r, σ_w^2 and σ_w^2 are known. Your colleague Dlorah decides to infer the remaining parameters using the following variational distribution:

$$q(\mu_w, \mu_v) = \mathcal{N}(\mu_w|m, s)\mathcal{N}(\mu_v|m, s)$$

Both $q(\mu_w)$ and $q(\mu_v)$ are parameterized with the *same* distribution $\mathcal{N}(\mu_w|m, s)$. Describe briefly the potential problems (if any) that may occur when using this variational distribution.

Your Answer (brief explanation):

Problem 4. Fast or Slow? [25 points]

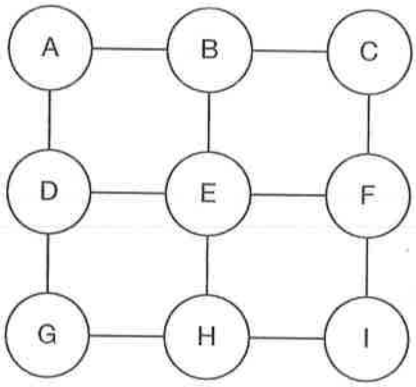


Figure 3: A Lattice Markov Random Field (MRF).

You are given the above Markov random field (MRF). Assume each random variable $A, B, ..., I$ is discrete with K possible states. The MRF is completely parameterized with only pairwise potential functions $\phi(\cdot, \cdot)$ between neighboring nodes.

Problem 4.a. [5 points] Given the elimination ordering $O = (I, H, G, F, E, D, C, B, A)$, draw out the *reconstituted graph.*

Your Answer: (Draw out the reconstituted graph.)

Problem 4.b. [10 points] Using your reconstituted graph, draw out the *junction tree.*

Your Answer: (use the scratch sheets if you need more space for your intermediate steps)

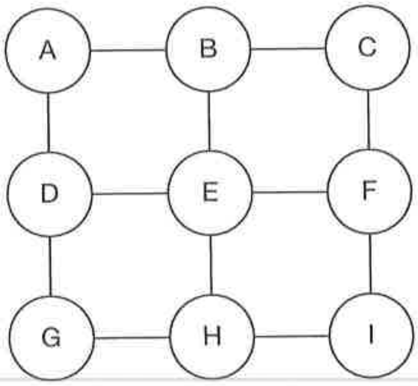
Problem 4.c. [5 points] We can "grow" the graph structure in Fig. 3 to larger sizes by increasing the number of nodes n in the lattice structure; we can add a row and column of nodes. Do you expect inference (via the junction tree algorithm) on graphs with this type of lattice structure to be **fast** (polynomial in the number of nodes n) or **slow** (exponential in n)? Provide a brief justification.

Hint: *How does the computation time depend on the properties of the junction tree? Consider the given elimination ordering; is it a good one?*

Your Answer: (Fast or Slow)

Brief Explanation:

Problem 4.d. [5 points] We can perform inference on the lattice MRF (Fig. 3 and shown below) using Gibbs sampling. For each of the nodes below, state the *minimal* set of nodes that have to be conditioned upon in the Gibbs sampling step, i.e., for a given node X, what nodes must constitute X_E for $p(X|X_E)$ that we sample from?



Node A

Node B

Node E

Node F

Node H

Scratch Paper

Scratch Paper

Scratch Paper

Scratch Paper

Scratch Paper

Scratch Paper

Scratch Paper

End of Paper