UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE	:	**COMPM056**
ASSESSMENT PATTERN	:	**COMPM056A**
MODULE NAME	:	**Graphical Models (Masters Level)**
DATE	:	**27-May-10**
TIME	:	**14:30**
TIME ALLOWED	:	**2 Hours 30 Minutes**

2009/10-COMPM056A-001-EXAM-1

Graphical Models, M056, 2009

Answer THREE of FIVE questions.

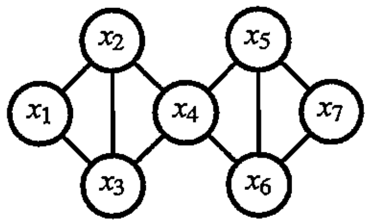
Marks for each part of each question are indicated in square brackets

Calculators are NOT permitted

1. a. Explain the difference between a general Markov Network and a pairwise Markov Network.

[3 marks]

b. Consider the Markov Network defined below.



i. By the use of Bayes' rule and explicit computation of the required conditional distributions, show that $x_1 \perp\!\!\!\perp x_7 | x_4$.

[6 marks]

ii. Show that $x_1 \perp\!\!\!\perp x_4 | \{x_2, x_3, x_5, x_6, x_7\}$.

[6 marks]

iii. Show that $p(x_4|x_1, x_2, x_3, x_5, x_6, x_7) = p(x_4|x_2, x_3, x_5, x_6)$.

[6 marks]

c. Consider the following distribution defined on discrete variables $x_i \in \{0, 1\}$:

$$p(x_1, x_2, x_3, x_4) \propto e^{w_{12}x_1x_2 + w_{123}x_1x_2x_3 + w_{234}x_2x_3x_4}$$

i. Draw a Markov Network for this distribution.

[3 marks]

ii. Draw a factor graph for this distribution.

[3 marks]

iii. Show that $x_1 \perp\!\!\!\perp x_4 | \{x_2, x_3\}$.

[3 marks]

iv. Is there a setting of w_{12}, w_{123} and w_{234} such that $x_1 \perp\!\!\!\perp x_4 | \emptyset$?

[3 marks]

[Total 33 marks]

2. a. Consider the distribution

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

i. Draw the Directed Acyclic Graph that represents this distribution. [4 marks]

ii. Draw the moralised graph. [2 marks]

iii. Draw the triangulated graph. Your triangulated graph should contain cliques of the smallest size possible. [3 marks]

iv. Draw a Junction Tree for the above graph and verify that it satisfies the running intersection property. [4 marks]

v. Describe a suitable initialisation of clique potentials. [4 marks]

vi. Describe the Absorption procedure and an appropriate message updating schedule. [4 marks]

b. Consider the distribution

$$p(y|x_1,\ldots,x_T)p(x_1)\prod_{t=2}^{T}p(x_t|x_{t-1})$$

where all variables are binary.

i. Draw a Junction Tree for this distribution and explain the computational complexity of computing $p(x_T)$, as suggested by the Junction Tree. [6 marks]

ii. Explain how $p(x_T)$ can be computed in $O(T)$ time. [6 marks]

[Total 33 marks]

3. A Markov Decision Process is defined by a distribution on x_2, x_3, x_4 for decisions d_1, d_2, d_3

$$p(x_2, x_3, x_4|x_1, d_1, d_2, d_3) = \prod_{t=1}^{3} p(x_{t+1}|x_t, d_t)$$

The discounted sum of utilities is defined by

$$u(x_2) + \gamma u(x_3) + \gamma^2 u(x_4)$$

where $0 < \gamma < 1$.

a. Draw an influence diagram for this Markov Decision Process.

[3 marks]

Defining

$$U(d_1) \equiv \sum_{x_2} \max_{d_2} \sum_{x_3} \max_{d_3} \sum_{x_4} p(x_2, x_3, x_4|x_1, d_1, d_2, d_3) \left[u(x_2) + \gamma u(x_3) + \gamma^2 u(x_4)\right]$$

b. Show that

$$U(d_1) = \gamma \sum_{x_2} p(x_2|x_1, d_1) \max_{d_2} \sum_{x_3} p(x_3|x_2, d_2) \left[u(x_3) + u_{3\leftarrow 4}(x_3)\right] + \sum_{x_2} p(x_2|x_1, d_1) u(x_2)$$

where the message $u_{3\leftarrow 4}(x_3)$ is defined as

$$u_{3\leftarrow 4}(x_3) \equiv \gamma \max_{d_3} \sum_{x_4} p(x_4|x_3, d_3) u(x_4)$$

[6 marks]

c. Show that

$$U(d_1) = \sum_{x_2} p(x_2|x_1, d_1) \left[u(x_2) + u_{2\leftarrow 3}(x_2)\right]$$

where we define the message

$$u_{2\leftarrow 3}(x_2) = \gamma \max_{d_2} \sum_{x_3} p(x_3|x_2, d_2) \left[u(x_3) + u_{3\leftarrow 4}(x_3)\right]$$

[4 marks]

d. For a Markov Decision Process as above, extended to T timesteps, define the value as

$$v_t(x_t) \equiv \begin{cases} u(x_t) + u_{t \leftarrow t+1}(x_t) & t < T \\ u(x_T) & t = T \end{cases}$$

for suitably defined messages $u_{t \leftarrow t+1}(x_t)$. Derive the recursion

$$v_{t-1}(x_{t-1}) = u(x_{t-1}) + \gamma \max_{d_{t-1}} \sum_{x_t} p(x_t | x_{t-1}, d_{t-1}) v_t(x_t)$$

[5 marks]

and show that the optimal decision d_t^* is then given by

$$d_t^* = \operatorname*{argmax}_{d_t} \sum_{x_{t+1}} p(x_{t+1} | x_t, d_t) v(x_{t+1})$$

[4 marks]

e. In the limit $T \to \infty$ and assuming a stationary value $v(x)$, derive the relation

$$v(x) = u(x) + \gamma \max_{d} \sum_{x'} p(x' | x, d) v(x')$$

[5 marks]

Explain how Value and Policy iteration may be used to solve for $v(x)$.

[6 marks]

[Total 33 marks]

4. A Markov chain is defined on variables $x_t \in \{1,\dots,X\}$, $t = 1,\dots,T$, by

$$p(x_1,\dots,x_T) = p(x_1)\prod_{t=2}^{T} p(x_t|x_{t-1})$$

for a transition

$$\theta_{j,i} \equiv p(x_t = j|x_{t-1} = i), \qquad i,j \in \{1,\dots,X\}$$

a. i. For a dataset of observed transitions $\mathcal{X} = \{x_1,\dots,x_T\}$ derive the maximum likelihood setting for the transition

$$\theta_{j,i} = \frac{\sum_{t=2}^{T} \mathbb{I}[x_t = j]\,\mathbb{I}[x_{t-1} = i]}{\sum_j \sum_{t=2}^{T} \mathbb{I}[x_t = j]\,\mathbb{I}[x_{t-1} = i]}$$

[5 marks]

ii. In a Bayesian approach to learning, one places a prior distribution $p(\theta)$ on the transition to form a joint distribution formed from the likelihood and the prior:

$$p(x_1,\dots,x_T,\theta) = p(\theta)p(x_1)\prod_{t=2}^{T} p(x_t|x_{t-1}|\theta)$$

Draw a Belief Network for this distribution.

[4 marks]

iii. Explain the concept of conjugacy and derive the appropriate conjugate distribution for θ.

[5 marks]

b. The stationary distribution p of a Markov chain is defined by

$$p_j = \sum_{i=1}^{X} \theta_{j,i} p_i$$

i. Explain how to make a basic search engine based on interpreting the stationary distribution of a Markov chain as the 'rank' of a website.

[4 marks]

ii. Describe a fast approximate way to compute the stationary distribution in the search-engine case, exploiting the properties of the structure of the web.

[2 marks]

iii. Your boss knows that it is possible to navigate from website 1 to website 100 in 50 clicks. Derive an efficient algorithm to find the most probable path from website 1 to website 100 in 50 clicks.

[8 marks]

iv. Derive an efficient algorithm to find the most probable path from website 1 to website 100 in any number of clicks.

[5 marks]

[Total 33 marks]

5. The symptoms a patient shows can be expressed by the binary vector s with elements

$$\mathbf{s} = (s_1, \dots, s_S)^\mathsf{T} \qquad s_i \in \{0,1\}$$

so that patient n has symptom i if $s_i^n = 1$. If patient n does not display symptom i then $s_i^n = 0$. There are a set of diseases

$$\mathbf{d} = (d_1, \dots, d_D)^\mathsf{T} \qquad d_j \in \{0,1\}$$

so that patient n has disease j if $d_j^n = 1$. Otherwise $d_j^n = 0$ if the patient does not have disease j. A patient may have more than one symptom and more than one disease.

A model for this situation is given by

$$p(\mathbf{s}, \mathbf{d}) = \frac{1}{Z} e^{\mathbf{s}^\mathsf{T}\mathbf{W}\mathbf{d} + \mathbf{a}^\mathsf{T}\mathbf{s} + \mathbf{b}^\mathsf{T}\mathbf{d}}$$

where Z is a normalisation constant and $\mathbf{W}, \mathbf{a}, \mathbf{b}$ are parameters.

a. Draw a Markov Network for this model.

[5 marks]

b. Show that

$$p(\mathbf{d}|\mathbf{s}) = \prod_{j=1}^{D} \sigma\left((2d_j - 1) \left[b_j + \sum_{i=1}^{S} W_{i,j} s_i \right] \right)$$

where $\sigma(x) \equiv e^x/(1+e^x)$.

[5 marks]

c. There are a set of N patients, each with a record $(\mathbf{s}^n, \mathbf{d}^n)$, $n = 1, \dots, N$. Assuming the patient records are independently and identically distributed according to the model, show that the log likelihood is given by

$$L = \left[\sum_{n=1}^{N} (\mathbf{s}^n)^\mathsf{T} \mathbf{W} \mathbf{d}^n + \mathbf{a}^\mathsf{T} \mathbf{s}^n + \mathbf{b}^\mathsf{T} \mathbf{d}^n \right] - N \log Z$$

[5 marks]

d. Show that the derivative of the log-likelihood is

$$\frac{dL}{dW_{i,j}} = \sum_{n=1}^{N} s_i^n d_j^n - Np(s_i = 1, d_j = 1)$$

Hint: use the fact that the variables are binary 0,1. Derive similar expressions for

$$\frac{dL}{da_i}, \qquad \frac{dL}{db_j}$$

[5 marks]

e. Explain if the derivatives above can be computed efficiently for general $\mathbf{W}$.

[2 marks]

f. For the special case $D = S$, describe structures of the matrix $\mathbf{W}$ and draw the corresponding Markov Network for cases in which Z can be computed in time which is linear is D.

[4 marks]

g. A colleague suggests to define a new model

$$p(\mathbf{d}, \mathbf{s}) = p(\mathbf{d}|\mathbf{s}) \prod_{i=1}^{S} p(s_i)$$

where $p(\mathbf{d}|\mathbf{s})$ is defined in question 5(b). Describe a procedure to learn $\mathbf{W}$, $\mathbf{b}$ and $p(s_i)$, $i = 1, \ldots, S$, for this distribution, based on identically and independently distributed data $(\mathbf{s}^n, \mathbf{d}^n)$, $n = 1, \ldots, N$. You should discuss issues related to computational tractability and the geometrical structure of the objective function.

[5 marks]

Discuss any potential drawbacks of this simplified model.

[2 marks]

[Total 33 marks]

END OF PAPER