UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **COMPGI18**

ASSESSMENT PATTERN : **COMPGI18A**

MODULE NAME : **Probabilistic and Unsupervised Learning**

DATE : **Friday 18 May 2018**

TIME : **14:30**

TIME ALLOWED : **2 hrs 30 mins**

This paper is suitable for candidates who attended classes for this module in the following academic year(s):

Year

2016/17 and 2017/18

EXAMINATION PAPER CANNOT BE REMOVED FROM THE EXAM HALL. PLACE EXAM PAPER AND ALL COMPLETED SCRIPTS INSIDE THE EXAMINATION ENVELOPE

TURN OVER

PROBABILISTIC AND UNSUPERVISED LEARNING
FINAL EXAMINATION
GI18, TERM 1, 2017-18

This examination contains SIX questions, of which you should answer FIVE.

Each question is worth 20 marks. If you have time, you may attempt all 6 questions, but only your 5 best scores will count towards your final mark. Make sure you manage your time carefully, read through ALL questions first before attempting any, and don't spend too much time on any one question. Calculators are allowed, but are unlikely to be useful.

Always provide justification and show any intermediate work for your answers. A correct but unsupported answer may not receive any marks.

In the following, bold symbols represent vectors. When asked to provide algorithms, you may provide equations or use MATLAB/Octave notation for matrix algebra. You may assume the availability of standard matrix functions such as `inv`, `pinv`, `eig` or `svd`. You may **not** assume statistical functions, including `mean` and `cov`. If you choose to use MATLAB notation, you will not lose marks for syntactic errors, as long as your intention is clear and correct.

You might find it useful to consult the following table of standard distributions:

Name	Domain	Symbol	Density or Probability fn
Gaussian (Normal)	$\mathbb{R}$	$x \sim \mathcal{N}\left(\mu, \sigma^2\right)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$
Multivariate Normal	$\mathbb{R}^D$	$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \Sigma\right)$	$\lvert 2\pi\Sigma\rvert^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
Bernoulli	$\{0, 1\}$	$x \sim \mathsf{Bernoulli}(p)$	$p^x(1-p)^{1-x}$
Binomial	$\mathbb{Z}_{0-N}$	$x \sim \mathsf{Binom}(p)$	$\binom{N}{x} p^x (1-p)^{N-x}$
Multinomial	$[\mathbb{Z}_{0-N}]^D$	$\mathbf{x} \sim \mathsf{Multinom}(\mathbf{p})$	$\frac{N!}{x_1!\, x_2! \ldots x_D!} \prod_{d=1}^{D} p_d^{x_d}$
Poisson	$\mathbb{Z}_{0+}$	$x \sim \mathsf{Poisson}(\mu)$	$\mu^x e^{-\mu}/x!$
Beta	$[0, 1]$	$x \sim \mathsf{Beta}(\alpha, \beta)$	$\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
Exponential	$\mathbb{R}_+$	$x \sim \mathsf{Exp}(\lambda)$	$\lambda e^{-\lambda x}$
Gamma	$\mathbb{R}_+$	$x \sim \mathsf{Gamma}(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
Student-t	$\mathbb{R}$	$x \sim \mathsf{Student\text{-}t}(\alpha)$	$\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\pi\alpha}\Gamma\left(\frac{\alpha}{2}\right)} \left(1 + \frac{x^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}$
Inverse-Wishart	$\mathbb{S}_+^p$	$X \sim \mathsf{InvWish}(\Psi, \nu)$	$\frac{\lvert\Psi\rvert^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} \lvert X\rvert^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\mathrm{Tr}[\Psi X^{-1}]}$
Dirichlet	$[0, 1]^D$	$\mathbf{x} \sim \mathsf{Dirichlet}(\boldsymbol{\alpha})$	$\frac{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} x_d^{\alpha_d - 1}$

1. **Conjugate mines**

(a) Define a conjugate prior. *[2 marks]*

(b) How are the parameters of the posterior distribution related to those of the conjugate prior? *[2 marks]*

Suppose that a rather devious type of mine contains an internal clock that ticks with period θ. Once the mine is armed it explodes on the next tick of the clock. Thus, the delay t between a random arming event and the following explosion is distributed:

$$t \sim \mathsf{Uniform}(0, \theta).$$

You arm n mines and (from a safe distance) observe delays to the respective explosions of $t_1 \ldots t_n$. You want to use these data to understand the properties of the internal clocks.

To begin, to assume that all mines have clocks with exactly the same period θ.

(c) What is the likelihood function for θ? What is the maximum likelihood estimate of θ? Do you think this is a reasonable estimate for small n (e.g. is it unbiased)? *[2 marks]*

(d) Show that a conjugate prior for θ has the form (known as the Pareto distribution):

$$p(\theta) = \frac{1}{Z(\theta_0, k)} \Theta(\theta - \theta_0) \frac{1}{\theta^k}$$

where $\Theta(x)$ is the Heaviside function ($= 1$ for arguments > 0 and 0 otherwise), and θ_0 and k are the hyperparameters.

Calculate the normalising constant $Z(\theta_0, k)$. *[3 marks]*

(e) Give the normalised posterior distribution on θ under this prior. What is the maximum a posterior estimate? *[2 marks]*

Now you wonder whether it is indeed true that all the clocks share the same period. Assume you have access to reasonable values for θ_0 and k.

(f) What is the marginal likelihood of your observed set of delays, under the model assumed thus far of a single θ? *[3 marks]*

(g) Suppose instead that each clock has a period drawn iid from the same Pareto prior. What is the marginal likelihood now? *[3 marks]*

(h) If all the observed delays happened to be identical, which model would be favoured? You may work in the limit $n \gg 1$. *[3 marks]*

2. **Linear Latent Models.** Recall that principal components analysis (PCA) can be obtained from the factor analysis (FA) generative model:

$$\mathbf{y} \sim \mathcal{N}(0, I)$$
$$\mathbf{x} \sim \mathcal{N}(\Lambda \mathbf{y}, \Psi)$$

by letting $\Psi = \epsilon I$ and taking $\epsilon \to 0$.

Suppose you have been given a data set $\{\mathbf{x}_n\}$ with very high dimension D, as well as the associated K-dimensional PCA latent variables $\{\mathbf{y}_n\}$. We are interested in finding the matrix Λ which defines the principal subspace of the data.

(a) Give an algorithm to find Λ from the data $\{\mathbf{x}_n\}$ alone. You may simply state the necessary steps (giving the relevant equations in terms of basic linear algebra operators), or you may write pseudocode. If you use pseudocode, you may assume access to standard linear algebra routines, but not to statistical ones.
[2 marks]

(b) Now give an algorithm to find Λ using both $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$. Again you may give equations or pseudocode. *[3 marks]*

(c) If $D \gg K$, which algorithm is likely to be less computationally expensive? Why?
[2 marks]

Suppose, now, that instead of the PCA latents, you are given the *mean* FA latent values $\{\boldsymbol{\mu}_n\}$ as well as the uniqueness matrix Ψ.

(d) How would you modify your second algorithm above to recover Λ in this case? Be careful not to assume access to any more information than is stated.
[5 marks]

(e) Prove that, given the maximum likelihood parameters Λ and Ψ,

$$\frac{1}{N} \sum_n \mathbb{E}\left[\mathbf{y}_n \mathbf{y}_n^\top | \mathbf{x}_n\right] = I$$

It may help to recall that $\Sigma = I - \beta\Lambda$ where Σ is the posterior covariance on $\mathbf{y}_n$ and β is the linear mapping from $\mathbf{x}_n$ to $\boldsymbol{\mu}_n$. *[8 marks]*

3. **A Random Game**

You are playing a game of exactly T moves. On each move $t = 1 \ldots T$ you may play exactly one of a finite set of actions $a_t \in A$ (the available set A doesn't change). Your chosen action (probabilistically) moves you from your previous square s_{t-1} to a new square s_t based on a conditional distribution $G_t(s_t|s_{t-1}, a_t)$. You start at a fixed square s_0. After all T moves, you win the game with a probability $R(s_T)$ determined by the square you end on.

Given all this randomness, you decide to choose your actions according to a random policy $\pi_t(a_t|s_{t-1})$ which gives the probability of playing a_t given the square you're on, and which may vary through the game.

(a) Draw the graphical model that relates π_t, a_t, s_t and the event of your winning or losing. *[3 marks]*

(b) What are your expected chances of winning in terms of the π_t, G_t and R? *[3 marks]*

(c) Derive an Expectation Maximisation (EM) based approach to efficiently choose the optimal π_t. Clearly specify the "likelihood" being maximised, and give the form of the inference messages and M-step updates. [Hint: consider an analogy to learning the transition matrix of a hidden Markov model.] *[8 marks]*

Now, suppose that instead of winning or losing based on the final square, you win a point after each move with probability $R(s_t)$ (we can assume the function R does not depend on time). Let r_t be an indicator variable for whether or not you earned a point on step t. Your aim is to maximise your expected total score

$$\rho = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right].$$

(d) One thought might be to use EM again to maximise the probability of $r_t = 1$ for all t. Explain why this would not work. *[2 marks]*

(e) How can you modify your algorithm to choose π_t to maximise ρ? Specify the "observations" that you would seek to fit, and describe how the rest of the algorithm would change. *[4 marks]*

4. **Learning Trees**

In this question we will investigate how to learn a tree-structured graphical model from fully observed data.

Assume we have m joint observations of n discrete variables $\{X_i\}$. Denote the empirical distribution observed for X_i by $d_i(x_i)$, and that for a pair of variables (X_i, X_j) by $d_{ij}(x_i, x_j)$.

(a) Demonstrate that every tree-structured DAG (i.e. directed acyclic graph) with a single root corresponds to an undirected (tree) graphical model and *vice versa.* How would you find the required factors in the undirected model or conditional distributions in the DAG. *[3 marks]*

(b) Suppose we are given a tree-structured graph T on the variables. Show that (assuming there are no further parametric restrictions on the distribution) the maximum likelihood factors given the tree T are those for which

$$\begin{aligned} p_i(x_i) &= d_i(x_i) && \text{for all nodes } i \text{ and values } x_i; \\ p_{ij}(x_i, x_j) &= d_{ij}(x_i, x_j) && \text{for all edges } (ij) \text{ in } T, \text{ and value pairs } x_i, x_j. \end{aligned}$$

where p_i, p_{ij} are the marginal distributions of variables X_i and (jointly) X_i, X_j respectively, under the tree-structured model. *[4 marks]*

(c) Show that the log likelihood at this solution, averaged over the empirical data distribution, is

$$\begin{aligned} &\sum_i \sum_{x_i} d_i(x_i) \log d_i(x_i) + \sum_{(ij)\in T} \sum_{x_i, x_j} d_{ij}(x_i, x_j) \log \frac{d_{ij}(x_i, x_j)}{d_i(x_i) d_j(x_j)} \\ &= -\sum_i H(X_i) + \sum_{(ij)\in T} MI(X_i, X_j) \end{aligned}$$

where $H(X_i)$ is the entropy of variable X_i, and $MI(X_i, X_j)$ is the mutual information of variables X_i, X_j, both under the empirical distribution. *[3 marks]*

(d) Now consider learning the structure of the tree as well as the corresponding factors. Show that the maximum likelihood structure is given by a maximum weight spanning tree for the fully-connected graph on the n random variables, where the weight of edge (ij) is given by $MI(X_i, X_j)$. *[3 marks]*

This approach maximises the likelihood jointly over structure and factors. If we just want to identify the tree structure, it would be preferable to use a Bayesian approach, integrating over the unknown factors. Consider a directed tree T with root r. Choose independent Dirichlet priors for the marginal distribution of the root X_r and for the conditional distributions of X_j given $X_i = x_i$ for each directed edge

$i \to j$ and for each value x_i on the directed tree:

$$p_r(X_r = x_r) = \theta_r(x_r) \qquad \theta_r(\cdot) \sim \text{Dirichlet}(\alpha_r(\cdot))$$
$$p_{ij}(X_j = x_j | X_i = x_i) = \theta_{ij}(x_i, x_j) \qquad \theta_{ij}(x_i, \cdot) \sim \text{Dirichlet}(\alpha_{ij}(x_i, \cdot))$$

Assume that the Dirichlet hyperparameters are all positive and satisfy the following local consistency properties: for each edge $i \to j$,

$$\sum_{x_i} \alpha_{ij}(x_i, x_j) = \alpha_j(x_j) \qquad \sum_{x_j} \alpha_{ij}(x_i, x_j) = \alpha_i(x_i) \qquad \sum_{x_i} \alpha_i(x_i) = \alpha_0$$

for some constant α_0 which denotes the strength of the prior.

(e) Show that the marginal probability of the observed data is:

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + m)} \prod_{x_r} \frac{\Gamma(\alpha_r(x_r) + m d_r(x_r))}{\Gamma(\alpha_r(x_r))}$$
$$\times \prod_{(i \to j) \in T} \prod_{x_i, x_j} \frac{\Gamma(\alpha_{ij}(x_i, x_j) + m d_{ij}(x_i, x_j))}{\Gamma(\alpha_{ij}(x_i, x_j))} \frac{\Gamma(\alpha_i(x_i))}{\Gamma(\alpha_i(x_i) + m d_i(x_i))}$$

[4 marks]

(f) Show that the maximum marginal likelihood tree is again given by a maximum weight spanning tree. Give the form of the edge weights. *[3 marks]*

5. **A Ring of Proteins.**

A biologist friend asks for your help. He is studying a protein complex in which 8 subunits are arranged in a ring-like structure. Each subunit can exist in one of 3 possible configurations; he is interested in understanding how these configurations interact around the ring.

He tells you that the probability of finding the ring structure in a particular configuration (defined by the configurations of each of the subunits) can be written as a Gibbs distribution. That is, if x_i denotes the configuration of the ith subunit

$$P(x_1 \dots x_8) = \frac{1}{Z} e^{-E(x_1 \dots x_8)/kT}$$

where $E()$ is an energy function; k is a constant; T is the temperature; and Z is a normaliser. Furthermore, on theoretical grounds, E can be written as a sum of 8 terms, each of which depends on the configuration of two adjacent subunits $E_1(x_1, x_2), E_2(x_2, x_3), \dots, E_8(x_8, x_1)$.

In a triumph of biotechnology, he has managed to attach 8 different flourescent ligands to the subunits. Each gives a noisy signal y_i reflecting the configuration of its corresponding subunit (x_i). The probabilistic mapping of subunit configuration to fluorescence value (i.e. $p(y_i|x_i)$) is known.

(a) Given this information, write the joint distribution over $(x_1 \dots x_8, y_1 \dots y_8)$ as a product of factors, each one of which involves a minimal number of variables. *[2 marks]*

(b) Draw an undirected graphical model reflecting this factorisation. *[2 marks]*

Your friend has made repeated measurements of fluorescence signals from individual molecules of the protein complex, all at the same temperature. He would like to use these to characterise the terms in the energy function. You realise that you can use the EM algorithm to help. You start by building a junction tree for your graph.

(c) What is the output of the junction tree algorithm? How will this help you implement EM? *[3 marks]*

(d) Triangulate the undirected graph you drew above. *[4 marks]*

(e) Obtain a junction tree from your triangulation. *[4 marks]*

Your friend, who knows a little bit about graphical models, has been looking over your shoulder. He sees you finish your implementation of the junction tree algorithm and says, "Great! Now all I have to do is maximise each E_i separately after averaging over the posteriors."

(f) Write down the gradient that must be optimised in the M-step. Your friend is wrong about the E_i being separately maximisable. Explain why the E_i terms have to be maximised jointly. *[5 marks]*

6. **Periodic Gaussian Processes**

Recall that a Gaussian Process (GP) prior over functions $f(t) \sim \mathcal{GP}(0, K)$ can (in certain cases) be derived from the Bayesian linear regression model

$$\beta \sim \mathcal{N}(0, I)$$
$$f(t) = \beta^{\mathsf{T}} \phi(t)$$

where $\phi(t) = [\phi_1(t), \phi_2(t), \dots]$ is a vector of nonlinear feature functions, and I is an identity matrix.

(a) Derive the relationship between $\phi(t)$ and the GP covariance kernel $K(t, t')$? *[4 marks]*

(b) Note that we have written the linear mapping above without noise. If we had incorporated noisy observations:

$$y \sim \mathcal{N}\left(\beta^{\mathsf{T}} \phi(t), \sigma^2\right)$$

what would be the (marginal) covariance matrix for a set of observations $\{y_i\}$ at inputs $\{t_i\}$? *[2 marks]*

The standard choice of kernel function for a Gaussian Process prior on periodic functions $f(t)$ is

$$K(t, t') = s^2 e^{-\frac{2\sin^2(\pi|t-t'|/p)}{l^2}}$$

(c) Give interpretations of the hyperparameters s, p and l. *[3 marks]*

(d) Sketch a draw from a GP with the product kernel

$$K(t, t') = s^2 e^{-\frac{2\sin^2(\pi|t-t'|/p)}{l^2}} e^{-\frac{(t-t')^2}{2r^2}}$$

indicating the roles of s, p, l and r. *[4 marks]*

An alternative periodic kernel is simply $K(t, t') = s^2 \cos(2\pi(t - t')/p)$.

(e) Show that this is a valid kernel by explicitly constructing a corresponding feature vector $\phi(t)$. *[2 marks]*

(f) Based on this decomposition, describe the set of all functions that have non-zero density under the corresponding GP prior. *[3 marks]*

(g) Why do you think the first form given above is the more commonly used? *[2 marks]*