Candidate No. ....
Seat No. ....

UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE	:	**COMPGI18**
ASSESSMENT PATTERN	:	**COMPGI18A**
MODULE NAME	:	**Probabilistic and Unsupervised Learning**
DATE	:	**02 June 2017**
TIME	:	**10:00 am**
TIME ALLOWED	:	**2 hours 30 mins**

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

2015/16, 2016/17

Under no circumstances are the attached papers to be removed from the examination by the candidate.

TURN OVER

PROBABILISTIC AND UNSUPERVISED LEARNING
FINAL EXAMINATION
COMPGI18, 2016-17

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)	$\mathbb{R}$	$x \sim \mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$
Multivariate Normal	$\mathbb{R}^D$	$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$	$\lvert 2\pi\Sigma\rvert^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
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	$\mathbb{S}_+^p$	$X \sim \mathsf{InvWish}(\Psi, \nu)$	$\frac{\lvert\Psi\rvert^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p(\frac{\nu}{2})} \lvert X\rvert^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\mathsf{Tr}[\Psi X^{-1}]}$
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. **Free Energies.**

In a fit of confusion, one of your colleagues is attempting to derive the expectation maximisation (EM) algorithm for a mixture model using the free energy, but has forgotten the entropy term.

That is, given data $\{\mathbf{x}_i\}$ and a generic mixture model with M components and parameters $\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$:

$$
\begin{aligned}
s_i &\sim \mathsf{Discrete}(\boldsymbol{\pi}) \qquad && s_i \in \{1 \ldots M\};\ \pi_m \geq 0;\ \sum_{m=1}^{M} \pi_m = 1 \\
\mathbf{x}_i | s_i{=}m &\sim P_m(\mathbf{x}_i; \boldsymbol{\theta}_m) \qquad && \mathbf{x}_i \in \mathbb{R}^D ,
\end{aligned}
$$

he has introduced posterior responsibilities $r_i^m \geq 0$ with $\sum_{m=1}^{M} r_i^m = 1$ for each observation i, and is seeking to maximize

$$
\widetilde{\mathcal{F}}(\{r_i^m\}, \{\boldsymbol{\theta}_m\}, \boldsymbol{\pi}) = \sum_i \sum_m r_i^m \log P(\mathbf{x}_i, s_i{=}m | \boldsymbol{\theta}_m, \boldsymbol{\pi}) ,
$$

with respect to $\{r_i^m\}$, $\{\boldsymbol{\theta}_m\}$ and $\boldsymbol{\pi}$.

However, he quickly runs in to a problem. When using the method of Lagrange multipliers to find the optima with respect to r_i^m (as we did in lecture) he obtains derivatives in which the r_i^m do not appear. He calls you over to help.

(a) You suggest thinking geometrically. Taking the case $M = 2$, sketch the isovalue contours for $\widetilde{\mathcal{F}}$ as a function of the responsibilities for the first data point, as well as the constraint surface $\sum_m r_1^m = 1$. Argue that the maximum will typically be found when $r_1^m = 1$ for some m. What determines which m? *[5 marks]*

You both wonder if this result can be understood in another way.

(b) Derive the correct free energy for this model, showing explicitly that it lower bounds the log-likelihood. You should start from the log-likelihood for the model above, and complete the derivation using the specific model notation — **do not** reproduce a generic derivation and substitute the model-specific values at the final step. *[5 marks].*

(c) Show that the incorrect free energy $\widetilde{\mathcal{F}}$ is also a lower bound on the log likelihood. *[3 marks]*

Now consider a continuum of free energies $\mathcal{F}_\tau, \tau \in [0, 1]$ in which the entropy term is premultiplied by τ.

(d) Derive the optimal responsibilities for $\mathcal{F}_\tau$ when $\tau > 0$. Then take the limit $\tau \to 0$ and show that you obtain the same result as you found geometrically. *[5 marks]*

(e) What well-known algorithm will result if you maximise $\widetilde{F}$ for a mixture of Gaussians with fixed identity covariance matrices and unknown means, and fixed mixing proportions $\pi_m = 1/M$. *[2 marks]*

2. **HMM posteriors.**

Consider a hidden Markov model (HMM) for observations $\mathbf{x}_t \in \mathbb{R}^D, t = 1 \dots T$ with:

hidden states	$s_t \in \{1 \dots K\}, t = 1 \dots T,$
initial state distribution	$P(s_1{=}i) = \pi_i,$
transition matrix	$P(s_t{=}j \mid s_{t-1}{=}i) = \Phi_{ij},$
and output distributions	$P(\mathbf{x}_t \mid s_t{=}i) = A_i(\mathbf{x}_t)\,.$

(a) Give the general Bayesian filtering recursion for $P(s_t|\mathbf{x}_1 \dots \mathbf{x}_t)$ in terms of the parameters and $P(s_{t-1}|\mathbf{x}_1 \dots \mathbf{x}_{t-1})$. How does this recursion depend on conditional independence in the HMM? *[4 marks]*

(b) Given forward and backward messages:

$$\alpha_t(i) = P(\mathbf{x}_1, \dots, \mathbf{x}_t, s_t{=}i|\Phi, \{A_j\}, \boldsymbol{\pi})$$
$$\beta_t(i) = P(\mathbf{x}_{t+1}, \dots, \mathbf{x}_T|s_t{=}i, \Phi, \{A_j\})$$

and a prior on the rows of Φ:

$$\Phi_{i,\cdot} \sim \mathsf{Dirichlet}(\boldsymbol{\lambda}) \qquad \boldsymbol{\lambda} \in \mathbb{R}_+^K$$

derive the *maximum a posteriori* (MAP) M-step update for the transition matrix Φ by finding the analytic maximimum of the free energy. *[8 marks]*

(c) Now introduce a set of diagonal matrices

$$\mathcal{A}_t = \begin{bmatrix} A_1(\mathbf{x}_t) & 0 & 0 & \dots & 0 \\ 0 & A_2(\mathbf{x}_t) & 0 & \dots & 0 \\ 0 & 0 & \ddots & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & A_K(\mathbf{x}_t) \end{bmatrix}$$

and derive a closed-form expression for the likelihood as a (long!) matrix-vector product. *[5 marks]*

(d) Based on this expression, write down a formal expression for the posterior distribution on the transition matrix Φ (assuming that all the other parameters are known). Do you think direct maximisation of this posterior is possible? Why (not)? *[3 marks]*

3. **Protocol sniffing.**

A secure network router handles 5 different address encryption protocols. The time taken to handle the ith incoming packet (and send it on to the next gateway) is made up of two parts:

- the routing time τ_{p_i} which depends on the protocol p_i that the current packet employs
- and a protocol-dependent startup time σ_{0,p_1} for the first packet following a reset, and a state-switching time σ_{p_{i-1},p_i} for $i \geq 2$, which depends on both the current protocol (p_i) and the protocol employed by the previous packet (p_{i-1}).

By monitoring radio-frequency emissions from the router you are able to obtain the times at which outgoing packets are sent, the differences between which (t_i) correspond to the processing delays above, plus an exponentially-distributed random wait for the next incoming packet (with parameter λ).

You wish to use the sequence $\{t_i\}_{i=1}^N$ to infer the packet protocols $\{p_i\}_{i=1}^N$. Assume that the packet protocols are *a priori* independent, with frequencies given by the probability vector $\boldsymbol{\pi}$.

(a) Draw the directed graphical model that relates the random variables $\{p_i\}$, the measured delatys $\{t_i\}$, and the parameters. *[3 marks]*

(b) Consider the moralised graph on the latents $\{p_i\}$ and write down the form of the corresponding clique potentials. *[5 marks]*

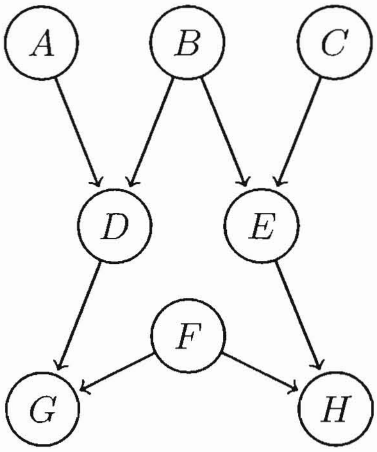
(c) This posterior factorisation is the same as would be obtained for state inference in a hidden Markov model (HMM). However, there is no equivalent standard HMM with state p_i and observations t_i that would yield the same posterior. Use the ideas of conditional independence to show why. *[2 marks]*

(d) However, it *is* possible to construct an equivalent HMM by augmenting the state. Show how, giving the corresponding transition matrix and output distribution. *[3 marks]*

(e) Derive an efficient (i.e. order N) algorithm to estimate from $\{t_i\}$ the most likely sequence of protocols that the router has handled. *[7 marks]*

4. **Bayes nets and belief propagation.**

Consider this directed acyclic graph (DAG):



(a) Is $C \perp\!\!\!\perp H \mid E$? *[2 marks]*

(b) Is $A \perp\!\!\!\perp H \mid \{F, G\}$? *[2 marks]*

(c) Is $A \perp\!\!\!\perp G \mid \{D, H\}$? *[2 marks]*

(d) Which nodes are in the Markov boundary of E? *[2 marks]*

(e) How many edges must be added to moralise the graph? *[2 marks]*

(f) Is it possible to use belief propagation on this graph (i.e. without forming a junction tree) to compute $P(B|G, H)$, the marginal distribution on B given that G and H are observed? Explain. *[3 marks]*

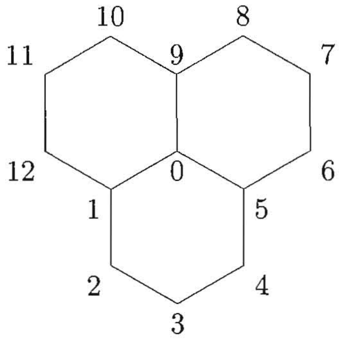
(g) What about $P(B|F, G)$? *[3 marks]*

(h) Now suppose we observe only F. How would you compute $P(B|F)$? *[2 marks]*

(i) What is the smallest possible size of the largest clique in a junction tree for the full graph with no nodes observed (i.e. the tree width of the graph + 1). *[2 marks]*

5. **Synthetic chemistry.**

A tricyclic molecular backbone has the form:



Each of the 13 carbon atoms (represented by the numbered vertices in the diagram) may have one of K different ligand species attached. Represent the ligand at location i by $l_i \in \{1 \dots K\}; i = 0 \dots 12$. The energy of the resulting molecule depends on the configuration of these ligands. Let $\mathcal{E}$ be the edgeset of the connectivity graph of the molecule (i.e. the set of bonds). Then

$$E(l_0 \dots l_{12}) = \sum_{(ij) \in \mathcal{E}} \varepsilon(l_i, l_j)$$

where the interaction energy function ε is the same for each edge. The probability of a particular molecule forming in a bath of temperature T is

$$P(l_0 \dots l_{12}) \propto e^{-E(l_0 \dots l_{12})/T} \prod_i C_{l_i} ,$$

where C_k is the concentration of ligand k in the bath.

A chemist would like to find the probability that $l_0 = k$ given observations of the ligands at positions 3, 7 and 11 (assuming that ε and the C_k are known); and from the dim recesses of memory recalls that the junction tree (JT) algorithm discussed in that machine learning class she once audited might be useful.

(a) She begins by triangulating the graph, but her first attempt is to connect node 0 to each of the others. Explain why this does not yield a triangulated graph. *[2 marks].*

(b) Realising her error she seeks advice, and learns that triangulation by variable elimination using minimum deficiency search is often a good idea. Suggest an appropriate variable elimination order, and show the triangulated graph that would result. *[5 marks]*

(c) Draw the JT obtained from this graph. Remember to check that it satisifies the running intersection property. *[4 marks]*

(d) Explain how to use belief propagation on the JT to obtain the desired probability $P(l_0 = k|l_3, l_7, l_{11})$. Specify the relevant clique potentials, and the form of the messages that must be passed. *[4 marks]*

(e) Assume now that the ligand concentrations $\{C_k\}$ in a bath are unknown, but that M molecules have been formed in the bath (at the same temperature), and for each one the triplet of ligands (l_3, l_7, l_{11}) has been observed. You have a chemical reaction simulator which is able to predict the distribution of ligands at any site given $\{C_K\}$, ε and T. Explain how this information can be used to infer the concentrations after the fact. *[5 marks]*

6. **Nonlinear regression.**

Consider the Bayesian nonlinear regression model:

$$y_i = \boldsymbol{\beta}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_i) + \epsilon \qquad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$
$$\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, A^{-1}\right) \qquad A = \mathsf{diag}\left[\alpha_1 \ldots \alpha_K\right]$$

where $\boldsymbol{\phi} : \mathbb{R}^D \to \mathbb{R}^K$ is vector-valued function from the D-dimensional input space of $\mathbf{x}$ to a K-dimensional feature n space, and $\boldsymbol{\beta}$ is a K-dimensional weight vector. It will be useful to collect the vectors $\boldsymbol{\phi}(\mathbf{x}_i)$ for points $\{\mathbf{x}_i\}_{i=1}^N$ into a $(K \times N)$-dimensional matrix Φ, and the corresponding values y_i into a $(1 \times N)$ row vector Y.

(a) Write down the joint distribution $P(Y|\Phi, \boldsymbol{\beta}, \sigma^2)$ as a vector normal on Y^T. *[3 marks]*

(b) Find the posterior distribution on $\boldsymbol{\beta}$ given the N observed $(\mathbf{x}_i, y_i)$ pairs. *[5 marks]*

(c) Find the marginal likelihood $P(Y^\mathsf{T}|\Phi, A, \sigma^2)$ by first arguing that it is normal, and then computing the mean and variance using the standard properties of moments of conditional distributions. *[3 marks]*

(d) Consider optimising the marginal likelihood with respect to the hyperparameters $\{\alpha_i\}$. Reasoning by analogy to automatic relevance determination (ARD) for linear regression, describe the form of the result you expect to obtain after optimisation. How would your interpretation of this result differ in the context of this regression model (rather than linear regression)? *[3 marks]*

(e) Differentiating the marginal likelihood directly is possible, but not necessarily the most efficient method of optimisation. One alternative is to treat $\boldsymbol{\beta}$ as a latent variable, and use expectation maximisation (EM). Compute the M-step update for A in this approach. *[3 marks]*

An alternative parameterisation would be to fix the prior on $\boldsymbol{\beta}$ to $\mathcal{N}(0, I)$, and instead introduce a D-dimensional hyperparameter $\boldsymbol{\gamma}$ to the regression model:

$$y_i = \boldsymbol{\beta}^\mathsf{T} \boldsymbol{\phi}(\boldsymbol{\gamma}^{-1} \circ \mathbf{x}_i) + \epsilon$$

where $\circ$ is the Hadamard or elementwise product. Again, it is possible to optimise the marginal likelihood with respect to $\boldsymbol{\gamma}$. This approach has the advantage of being compatible with Gaussian process (GP) regression, where the product $\boldsymbol{\phi}(\mathbf{x}_i)^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_j)$ is replaced by a covariance kernel $K(\mathbf{x}_i, \mathbf{x}_j)$.

(f) Consider using the exponentiated quadratic kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}.$$

Interpret the alternative hyperparameters $\boldsymbol{\gamma}$ within this regression model. If they behave in the same way as the αs of ARD, what can be concluded about the regression result? *[3 marks]*