ENGINEERING TRIPOS PART IIB

Friday 25 April 2008 9 to 10.30

Module 4F13

MACHINE LEARNING

Answer not more than ***five*** *questions.*

All questions carry the same number of marks.

The ***approximate*** *percentage of marks allocated to each part of a question is indicated in the right margin.*

The notation $\mathcal{N}(\mu, \Sigma)$ *denotes a univariate (or multivariate) Gaussian distribution with mean* μ *and variance (or covariance matrix)* Σ. I *denotes the identity matrix.*

There are no attachments.

STATIONERY REQUIREMENTS	SPECIAL REQUIREMENTS
Single-sided script paper	Engineering Data Book
	CUED approved calculator allowed

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

1 Consider continuous data $\mathbf{y}$ in D dimensions. We want to use a Factor Analysis model with a scalar (1-dimensional) hidden factor, x.

(a) What is the likelihood function, and how many free parameters does the model have? [20%]

(b) In the Expectation Maximisation (EM) algorithm, we work with a functional $\mathscr{F}$ which is a lower bound of the log likelihood. Write down the $\mathscr{F}$ functional. [20%]

(c) Using the fact that the Kullback-Leibler (KL) divergence between two distributions is non-negative, show that $\mathscr{F}$ is a lower bound on the log likelihood. [30%]

(d) Find an expression for the posterior distribution of the hidden variable. [30%]

2 Bayesian inference in models for machine learning requires evaluation of integrals. If these are intractable, one sometimes resorts to Markov Chain Monte Carlo (MCMC) methods.

(a) State the Metropolis algorithm to sample from a target distribution $p(\mathbf{x})$, using an isotropic (i.e. with covariance $\sigma^2 I$) Gaussian proposal distribution, centred on the current state. [25%]

(b) The efficiency of Metropolis sampling depends on the width of the proposal distribution. Explain what happens if this width is too narrow, or if it is too wide. [25%]

(c) In importance sampling, one draws random samples from a distribution $q(\mathbf{x})$, which is different from the target distribution $p(\mathbf{x})$ of interest. State the importance sampling algorithm for finding the average of a function $f(\mathbf{x})$ with respect to $p(\mathbf{x})$. [25%]

(d) Assume you use an importance sampler to evaluate an integral where the target distribution is heavy-tailed and the proposal distribution $q(\mathbf{x})$ is Gaussian. Why may the importance sampler be slow under these conditions? [25%]

3 Two different parametric models have been trained using Maximum Likelihood on the same training data.

(a) One model has a much higher training likelihood than the other. Is it necessarily the better model for predicting new data? [10%]

(b) In an unsupervised learning task, assume the likelihood has the form $p(\mathbf{y}|\theta)$, where θ is the vector of parameters. Write down the marginal likelihood. [20%]

(c) Approximate Bayesian inference is undertaken in a mixture model with two components. Assume the approximation to the posterior distribution captures only one of two symmetric modes, spaced widely apart in parameter space. How could you adjust the value of the estimated marginal likelihood to compensate for this failure? [30%]

(d) Prove that the marginal likelihood is upper bounded by the maximum likelihood. [40%]

4 Let $x_1, \ldots, x_5$ be binary variables (i.e. $x_i \in \{0,1\}$) and

$$p(x_1, \ldots, x_5) = \frac{1}{Z} \exp\{x_1 x_2 + x_2 x_3 x_4 - 2 x_4 x_5\}$$

where Z is a normalisation constant.

(a) Draw the factor graph for $p(x_1, \ldots, x_5)$. Is it singly connected? [30%]

(b) For each of the following marginal and conditional independence statements, state whether it is true or false for $p(x_1, \ldots, x_5)$:

(i) $x_1 \perp\!\!\!\perp x_3$

(ii) $x_1 \perp\!\!\!\perp x_5$

(iii) $x_1 \perp\!\!\!\perp x_4 | x_3$

(iv) $x_1 \perp\!\!\!\perp x_4 | x_2$

(v) $x_1 \perp\!\!\!\perp x_5 | x_3 = 0$

[30%]

(c) What is the message that the x_1—x_2 factor sends to x_2? [40%]

5 Consider Bellman's optimality equation:

$$V^*(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right]$$

where s and s' represent states, a represents actions, and $V^*(s)$ is the optimal value of state s.

(a) Give an interpretation for $R(s,a)$, γ and $P(s'|s,a)$. [30%]

(b) Describe the value iteration algorithm for solving for $V^*(s)$. [30%]

(c) Assume a Markov Decision Process (MDP) with two states $\{1,2\}$, two actions (`stay` and `jump`), $\gamma = 1/2$ and the following settings for the MDP:

$$\begin{aligned}
P(s'=1|s=1,a=\texttt{stay}) &= 1 \\
P(s'=2|s=2,a=\texttt{stay}) &= 1 \\
P(s'=2|s=1,a=\texttt{jump}) &= 1 \\
P(s'=1|s=2,a=\texttt{jump}) &= 1 \\
R(1,\texttt{stay}) &= 1/2 \qquad R(1,\texttt{jump}) = 0 \\
R(2,\texttt{stay}) &= 2 \qquad R(2,\texttt{jump}) = 1.
\end{aligned}$$

Solve for the optimal value function. [40%]

6 Consider the following model of data (known as a bilinear model):

$$y = xz + \varepsilon$$

where x, z and the noise, ε, are each distributed $\mathcal{N}(0,1)$, so that $y|x,z \sim \mathcal{N}(xz,1)$.

(a) Is the distribution $p(x|y=1)$ Gaussian? Explain your answer. [30%]

(b) Assume we want to approximate $p(x,z|y=1)$ using a variational approximation $q(x,z) = q_1(x)q_2(z)$. Show, using Jensen's inequality, that the following inequality holds:

$$\mathscr{F}(q_1(x), q_2(z)) = \int q_1(x)q_2(z) \ln \frac{p(x,z,y=1)}{q_1(x)q_2(z)} dxdz \leq \ln p(y=1)$$

[30%]

(c) Assume that $q_2(z)$ is distributed according to $\mathcal{N}(2,1)$ and that $q_1(x)$ is also Gaussian distributed. Solve for the parameters of $q_1(x)$ that maximise $\mathscr{F}(q_1(x), q_2(z))$. [40%]

END OF PAPER