CRIB

2008 IIB 4F13 MACHINE LEARNING PROF Z GHAHRAMANI

ENGINEERING TRIPOS PART IIB

? ?

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 attachmens.

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

Version: 2

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%]

ANSWER: $p(y|\Lambda, \Psi) = \mathcal{N}(0, \Lambda\Lambda^\top + \Psi)$. *There are 2D free parameters.*

(b) In the EM algorithm, we work with a lower bound functional $\mathscr{F}$ to the log likelihood. Write down the $\mathscr{F}$ functional. [30%]

ANSWER: $\mathscr{F}(\theta, q) = \int q(x) \log(\frac{p(\mathbf{y},x|\theta)}{q(x)})dx.$

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

ANSWER: *Write* $\mathscr{F}$ *as*

$$\mathscr{F}(\theta, q) = \int q(x) \log(\frac{p(x|\mathbf{y},\theta)p(\mathbf{y}|\theta)}{q(x)})dx \;=\; \log(p(y|\theta)) - \mathrm{KL}(q(x), p(x|\mathbf{y},\theta)).$$

(d) Find an expression for the posterior distribution of the hidden variable. [20%]

ANSWER: *The un-normalized posterior is the likelihood times the prior,* $p(\mathbf{y}|x,\theta)p(x)$.

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 $p(\mathbf{x})$, using an isotropic (i.e. with covariance $\sigma^2 I$) Gaussian proposal distribution, centered on the current state. [25%]

ANSWER: *The current state is* $\mathbf{x}$. *Iterate the following:*

- *propose a candidate* $\mathbf{x}^*$ *from* $\mathscr{N}(\mathbf{x}, \sigma^2 I)$.
- *accept the proposed state* $\mathbf{x}^*$ *if* $p(\mathbf{x}^*)/p(\mathbf{x}) > u$, *where* u *is sampled uniformly in* $[0;1]$, *otherwise retain the old state.*

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

ANSWER: *Too narrow: slow exploration of p. Too wide: very low acceptance rate.*

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

ANSWER: *Use* $\frac{1}{T}\sum_t f(x^{(t)})\frac{p(x^{(t)})}{q(x^{(t)})}$, *where* $x^{(t)}$ *are samples drawn from* $q(x)$.

(d) You use an importance sampler to evaluate an integral where the target distribution is heavy tailed; the proposal distribution q is Gaussian. Why may the importance sampler be slow under these conditions? [25%]

ANSWER: *The importance weights may have occasional, very large values; the variance of the estimate will be large.*

3 Two different parametric models have been trained using Maximum Likelihood on the same training data. One model has a much lower training error than the other.

(a) Which model is best? [10%]

ANSWER: *It is not possible to tell from the training error alone.*

(b) In an unsupervised learning task, the likelihood is $p(\mathbf{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters. Write down the marginal likelihood. [20%]

ANSWER: *Marginal likelihood:* $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta$.

(c) Approximate Bayesian inference is undertaken in a mixture model with two components. The approximation to the posterior ditribution 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%]

ANSWER: *Double the value of the estimate.*

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

ANSWER:

$$\int p(y|\theta)p(\theta)d\theta \leq \int p(y|\theta_{\text{ML}})p(\theta)d\theta = p(y|\theta_{\text{ML}}) \int p(\theta)d\theta = p(y|\theta_{\text{ML}}).$$

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

$$p(x_1,\dots,x_5) = \frac{1}{Z}\exp\{x_1x_2 + x_2x_3x_4 - 2x_4x_5\}$$

where Z is a normalisation constant.

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

ANSWER: *Picture goes here. Yes.*

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

(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%]

ANSWER: *No. No. No. Yes. Yes.*

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

ANSWER: *Let f_{12} refer to the factor between x_1 and x_2, then*

$$\mu_{f_{12}\to x_2}(x_2) = \sum_{x_1} \exp\{x_1x_2\}\mu_{x_1\to f_{12}}(x_1)$$

where $\mu_{x_1\to f_{12}}(x_1) = 1$.

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%]

ANSWER: *$R(s,a)$ is the immediate reward obtained in state s after taking action a, γ is the dicount factor for future rewards, and $P(s'|s,a)$ is the probablity of transitioning from state s to state s' after taking action a.*

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

ANSWER: *Initialise the values arbitrarily (e.g. $V^*(s) = 0$) and iterate the above equation for each state until convergence.*

(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%]

ANSWER: *For each state we max over the two actions (stay and jump):*

$$V(1) = \max[0.5 + 0.5V(1), 0 + 0.5V(2)]$$

$$V(2) = \max[2 + 0.5V(2), 1 + 0.5V(1)]$$

To solve this, first assume $0.5 + 0.5V(1) \geq 0.2V(2)$. Then $V(1) = 0.5 + 0.5V(1)$ which means $V(1) = 1$. So $V(2) = \max[2 + 0.5V(2), 1.5] = 2 + 0.5V(2)$ which means $V(2) = 4$ which is a contradiction.

Therefore, $0.5 + 0.5V(1) < 0.5V(2)$, so $V(1) = 0.5V(2)$ which means that $V(2) = \max[2 + 0.5V(2), 1 + 0.25V(2)] = 2 + 0.5V(2)$, therefore $V(2) = 4$ and $V(1) = 2$.

END OF PAPER