MACHINE LEARNING SAMPLE EXAM PAPER

4F13 Michaelmas, 2006
Cambridge University Engineering Department

This sample examination consists of nine (9) questions.

The *actual* examination will consist of six (6) questions and you will have to answer 5 questions. Some questions may take longer than others. Make sure you manage your time carefully and don't spend too much on any one question.

Each question is worth 20 marks for a total of 100 marks.

You have a total of 1.5 hours = 90 minutes.

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

---

1. Consider encoding symbols from an alphabet $\{a, b, c, d\}$ with probabilities $\mathbf{p} = (p_1, p_2, p_3, p_4)$ respectively.

   (a) What is the expression for the minimal expected coding cost in bits per symbol as a function of $\mathbf{p}$? *[8 marks]*

   (b) Find a probability distribution $\mathbf{p}$ such that the minimal expected coding cost is 1 bit per symbol. *[6 marks]*

   (c) If we incorrectly code assuming the distribution $\mathbf{q} = (1/4, 1/4, 1/4, 1/4)$ but the true distribution is $\mathbf{p}$ from part (b), how many bits per symbol are we wasting? Explain why. *[6 marks]*

2. Consider the uniform distribution between a and b on the real line, where $b > a$:

$$p(x|a, b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

   Imagine observing a data set $\mathcal{D} = \{-1.3, 0.2, 0.3, 1.1, 2.1\}$. Assume the data points in $\mathcal{D}$ came independently and identically distributed (i.i.d.) from $p(x|a, b)$.

   (a) Write an expression for the likelihood $p(\mathcal{D}|a, b)$. *[6 marks]*

(b) What are the maximum likelihood estimates of a and b? Expain why. *[8 marks]*

(c) Do these estimates seem reasonable to you? Explain why or why not. *[6 marks]*

3. Let $\mathbf{x}_1 \sim \mathcal{N}(0, S)$, that is $\mathbf{x}_1$ is Gaussian distributed with mean 0 and covariance matrix S. Let

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + \mathbf{w}_t$$

where $\mathbf{w}_t \sim \mathcal{N}(0, Q)$ and the $\mathbf{w}$s are uncorrelated with the $\mathbf{x}$s, and A is given.

(a) What is the distribution of $\mathbf{x}_2$? *[6 marks]*

(b) What is the distribution of $\mathbf{x}_2 - \mathbf{x}_1$? *[8 marks]*

(c) What is the distribution of $\mathbf{x}_2$ given $\mathbf{x}_1$? *[6 marks]*

4. Consider the probability distribution

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = \frac{1}{Z} g_1(x_1, x_2) g_2(x_2, x_3) g_3(x_1, x_3) g_4(x_3, x_4, x_5) g_5(x_3, x_6) \tag{1}$$

where the g_i are non-negative functions and Z is a normalization constant.

(a) Draw the appropriate factor graph for this distribution. *[6 marks]*

(b) For each of the following marginal and conditional independence relations, state whether it is true or false: *[8 marks]*

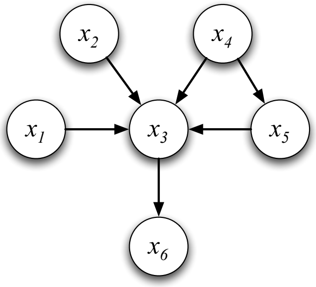
i. $x_1 \perp\!\!\!\perp x_5 | x_3$

ii. $x_2 \perp\!\!\!\perp x_6$

iii. $x_4 \perp\!\!\!\perp x_2 | \{x_3, x_6\}$

iv. $\{x_1, x_2, x_6\} \perp\!\!\!\perp \{x_4, x_5\} | x_3$

(c) List three conditional or marginal independencies in this directed graph which are not present in the above distribution (eqn 1): *[6 marks]*



5. Let x be drawn uniformly from the interval $[0, 1)$. Let $y = -\log(1-x)$.

   (a) What is the density of y, $p(y)$? Use the fact that $p(y)|dy| = p(x)|dx|$ when x and y are related by a one-to-one monotonic transformation, which implies that $p(y) = p(x)/|dy/dx|$. *[8 marks]*

   (b) Assuming we can draw $z \sim p(z) = e^{-z}$ for $z \geq 0$, describe a Monte Carlo method for estimating

   $$f = \int_0^\infty \sin(z^2)e^{-z}dz$$

   [6 marks]

   (c) Using rejection sampling, and the above assumption, describe a Monte Carlo method for estimating

   $$g = \int_0^5 \sin(z^2)e^{-z}dz$$

   Note that the upper limit of the integral is different from part (b). *[6 marks]*

6. You are a Statistician working for the National Health Service (NHS). You measure patient waiting times for two local councils, doing surveys of 100 patients in Council A, and 60 patients in Council B. You find that in Council A, 60 out of the 100 patients waited one week or more to see their doctor, and 40 patients waited less than a week. While in council B, 40 out of 60 patients waited one week or more, and 20 patients waited less than one week.

   Assume that whether a patient waits a week or more is independent of the waiting time of other patients in the council (this is probably unrealistic), but might depend on what council the patient lives in.

Let θ_A be the probability that a typical person in council A waits one week or more; similarly, let θ_B be the probability that a person in council B waits one week or more.

(a) Use Bayesian inference to reason about whether $\theta_B > \theta_A$. State your assumptions, and write down any relevant expressions you would have to evaluate. *[20 marks]*

7. Consider the following joint distribution $p(s_1, s_2)$ over 2 binary variables s_1 and s_2, written as a table over the 4 settings of (s_1, s_2):

	$s_2 = 0$	$s_2 = 1$
$s_1 = 0$	1/8	1/2
$s_1 = 1$	1/4	1/8

Let

$$q(s_1, s_2) = \lambda_1^{s_1}(1-\lambda_1)^{(1-s_1)}\lambda_2^{s_2}(1-\lambda_2)^{(1-s_2)} = q_1(s_1)q_2(s_2)$$

be a fully factorized approximation to $p(s_1, s_2)$, with $0 \leq \lambda_1 \leq 1$ and $0 \leq \lambda_2 \leq 1$.

(a) Compute the first derivatives of

$$KL(p(s_1, s_2)\|q(s_1, s_2)) = \sum_{s_1}\sum_{s_2} p(s_1, s_2) \log \frac{p(s_1, s_2)}{q(s_1, s_2)}$$

with respect to λ_1 and λ_2. *[8 marks]*

(b) Use the fact that, for fully factored q, $KL(p\|q)$ is minimized when q matches the marginals of p to solve for λ_1 and λ_2. *[6 marks]*

(c) Confirm that the values of λ_1 and λ_2 found in part (b) set the derivatives found in part (a) to zero. *[6 marks]*

8. Re-visiting the same setup as in the previous question: We have a joint distribution $p(s_1, s_2)$ over 2 binary variables s_1 and s_2, written as a table over the 4 settings of (s_1, s_2):

	$s_2 = 0$	$s_2 = 1$
$s_1 = 0$	1/8	1/2
$s_1 = 1$	1/4	1/8

and

$$q(s_1, s_2) = \lambda_1^{s_1}(1 - \lambda_1)^{(1-s_1)}\lambda_2^{s_2}(1 - \lambda_2)^{(1-s_2)} = q_1(s_1)q_2(s_2)$$

is a fully factorized approximation to $p(s_1, s_2)$, with $0 \leq \lambda_1 \leq 1$ and $0 \leq \lambda_2 \leq 1$.

Now consider minimizing *this* KL divergence:

$$KL(q(s_1, s_2)\|p(s_1, s_2)) = \sum_{s_1}\sum_{s_2} q(s_1, s_2) \log \frac{q(s_1, s_2)}{p(s_1, s_2)}$$

(a) Write down $KL(q(s_1, s_2)\|p(s_1, s_2))$ as a function of λ_1, λ_2, and the elements of the above 2×2 table. *[6 marks]*

(b) Solve for λ_1 as a function of λ_2 by taking derivatives of $KL(q\|p)$. [Note that we can similarly solve for λ_2 as a function of λ_1, but you do not need to do this]. *[8 marks]*

(c) Describe an iterative algorithm for minimizing $KL(q(s_1, s_2)\|p(s_1, s_2))$ with respect to (λ_1, λ_2) and argue whether this algorithm is guaranteed to converge or not. *[6 marks]*

9. Consider a sequential decision making problem with 2 states, A, and B, and two actions, L, and R. Let s_t denote the state at time t, and let a_t be the action taken at time t. Assume that the state transition probabilities are:

$$P(s_{t+1}|s_t, a_t = L) = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

$$P(s_{t+1}|s_t, a_t = R) = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$$

where each row represents the state (A or B) at time t and each column represents the state at time $t + 1$. That is, $P(s_{t+1} = A|s_t = B, a_t = R) = 1/4$.

Assume that the reward for being in state A is $r(s_t = A) = +1$, and for state B is $r(s_t = B) = -1$, and let γ be the discount factor.

(a) What is the expected discounted return of starting in state A and taking two actions R, R? *[6 marks]*

(b) Using the self-consistency of value functions,

$$V^\pi(s) = \sum_a \pi(a|s) \sum_{s'} P(s'|s,a) \left[r(s') + \gamma V^\pi(s') \right]$$

calculate the values $V^\pi(A)$ and $V^\pi(B)$ under the policy which always takes the action R, i.e. $\pi(a = R|s = A) = 1$ and $\pi(a = R|s = B) = 1$ in terms of γ. *[8 marks]*

(c) What is the optimal policy in this sequential decision problem? Explain why. *[6 marks]*