TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

EXAMINATIONS — 2006

MID-YEAR

COMP 421

Machine Learning

Time Allowed: 3 Hours

Instructions: There are 5 questions to choose from: each question is worth 45 marks.

Answer FOUR questions (180 marks).
If you answer more questions, only your best 4 will be taken.

Pay close attention to the number of marks for each sub-question, which gives an indication of the depth of answer that is expected.

Non-electronic Foreign-English language dictionaries are permitted.

Question 1. [45 marks]

Bayesians sometimes refer to the quantity $P(D|H)$ as the "evidence" for H. Here D is data and H denotes a "hypothesis" space that captures a whole family of possible predictors $P(D^{\text{new}}|\theta, H)$ corresponding to different settings of parameters θ.

(a) [5 marks] Give an expression for $P(D|H)$ in terms of parameters θ.

(b) [12 marks] Explain why the evidence $P(D|H)$ is *not* required for sampling from the posterior distribution over θ, *or* for optimising θ, but *is* required for comparing two hypotheses H_1 and H_2.

(c) [10 marks] For a discrete ("1-of-N") variable x the Kullback-Lieber divergence measure between two distributions P and Q is

$$KL(P, Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Discuss the relationship between this and the "expected surprise" of an agent that believes events happen with probabilities $Q(x)$ when in fact they happen with probabilities $P(x)$.

(d) [8 marks] Suppose a Hopfield network with initial weights set to zero uses the Hebb rule to update weights when exposed to a single pattern X composed of elements chosen from $\{-1, +1\}$. Show that X is a stable state of the network after this update.

Suppose that the world generates some number x from a zero-mean Gaussian distribution with variance σ_x^2. This is not directly observable itself, but emits an observation y that is simply x corrupted by measurement noise. This noise is additive and zero-mean Gaussian with variance σ_y^2.

(e) [4 marks] Given a single datum, y, show that the posterior distribution is Gaussian as well.

(f) [6 marks] Given a single datum, y, show that the MAP (maximum a posteriori) estimate for x is

$$x_{\text{map}} = \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} \right) y$$

Hint: differentiate the log posterior and set the gradient to zero.

Question 2. [45 marks]

(a) [6 marks] Metropolis sampling involves selecting an appropriate proposal distribution to use, for which a common choice is the spherical Gaussian with variance σ^2. Discuss the issues involved in determining a suitable value for σ^2.

(b) [2 marks] What is the relationship between the Metropolis algorithm and Simulated Annealing?

Consider data consisting of a set of $(\mathbf{x}, y)$ pairs, where "inputs" $\mathbf{x}$ may be vectors and "outputs" y are scalars. A Gaussian process model could be used to predict y given a new $\mathbf{x}$, by giving $P_{\text{gp}}(y|x)$. Suppose the Gaussian process uses the following covariance function:

$$C(x, x') = \theta_1 \exp\left[\frac{-|x - x'|}{\theta_2}\right] + \theta_3 \, \delta_{x,x'}$$

where $\delta_{x,x'} = 1$ if $x = x'$ and zero otherwise. θ_1, θ_2 and θ_3 are hyperparameters.

(c) [6 marks] Describe the role played by each of the hyperparameters θ_1, θ_2 and θ_3 in such a Gaussian process.

(d) [6 marks] Briefly describe a method for learning hyperparameters such as θ from data.

(e) [20 marks] Discuss the pros and cons of using Gaussian process inference versus Bayesian neural networks (in which predictions are made by integrating out weights using MCMC) for regression problems.

(f) [5 marks] Basic Q-learning only takes reinforcement r_{t+1} into account in updating $Q_{a_t,\mathbf{x}_t}$ (this is the "TD trick"). Give a Q-learning algorithm that uses both r_{t+1} and r_{t+2} to update $Q_{a_t,\mathbf{x}_t}$.

Question 3. [45 marks]

(a) [5 marks] Bayesians often need to calculate integrals over probability distributions, such as

$$\mathbb{E}_p[f] \;=\; \int f(x)\; p(x)\;\; dx$$

Only rarely do such integrals have analytic solutions, however. Show how to approximate such an integral by drawing samples $x^{(i)} \sim p(x)$.

(b) [6 marks] Importance sampling is a technique for approximating integrals over probability distributions such as the above when $p(x)$ is easy to evaluate but difficult to sample from directly. Explain how importance sampling achieves this approximation, using a second distribution $q(x)$ that *is* easy to draw samples from.

(c) [6 marks] Describe how you could use Gibbs Sampling to perform inference in a *directed* graphical model with discrete states.

(d) [8 marks] Would Gibbs sampling would work in an *undirected* graphical model (with discrete states)? If so, explain how. If not, explain why not.

(e) [20 marks] Suppose an agent is at a position $\mathbf{x}$ in a 2 dimensional world, and it has 4 possible actions, labelled "North", "South", "East" and "West". Each of these actions has the effect of deterministically moving $\mathbf{x}$ in the corresponding direction by 1 step of some predetermined, fixed size.

In reinforcement learning scenarios one is interested in the expected reward that follows from being in any given state x. Suppose we try to represent this with a Gaussian process model, $p_{\text{gp}}(r|x)$.

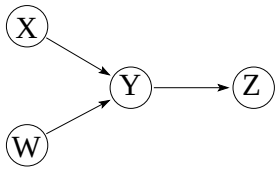
The agent also has uncertainty about its current position, represented by some distribution $p(\mathbf{x})$. You may assume the agent can "look up" this probability for any given $\mathbf{x}$, and that it uses a Gaussian process model to capture its knowledge about worthwhile states via $p_{\text{gp}}(r|\mathbf{x})$, with r being the *immediate* reward that follows from being in state $\mathbf{x}$.

What quantity should the agent calculate in order to decide what action to take? Give an expression for this quantity in terms of $p(\mathbf{x})$ and $p_{\text{gp}}(r|\mathbf{x})$, and indicate how you might go about calculating it in practice.

Question 4. [45 marks]

(a) [5 marks] Give **one** advantage, and **one** disadvantage, of using Gibbs sampling compared to the Metropolis algorithm, for drawing samples from some known probability distribution $p(x)$.

(b) [6 marks] Give the joint probability distribution embodied by the probabilistic graphical model shown here, and list the conditional independencies that are being assumed in adopting this model.



(c) [6 marks] Draw the belief network (not the factor graph) corresponding to the following factorisation of the joint probability over hidden variables $x_{0:T}$ and observations $y_{1:T}$

$$p(x_{0:T}, y_{1:T}) \;=\; p(x_0) \prod_{t=1}^{T} \; p(x_t|x_{t-1}) \;\; p(y_t|x_t, x_{t-1})$$

and describe *in words* (a sentence or two) what this model assumes about the origins of the data.

(d) [8 marks] Could the graphical model $X \longrightarrow Y \longrightarrow Z \longrightarrow X$ represent a joint probability distribution? If so, give an example. If not, explain why not.

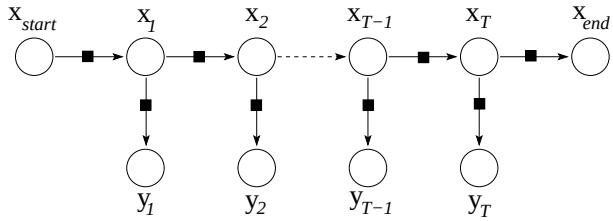
(e) [4 marks] Consider two observable variables X and Y, which are known to be dependent. Use probability theory to explain why measurements of X and Y alone provide no evidence at all as to whether X is a cause of Y.

(f) [8 marks] Suppose there is another observable variable, Z, that is thought to relate to X and Y, and that you collect a set of observations of all three. What relationships (dependencies and independencies) would these observations have to exhibit in order for you to conclude that X is *not* a cause of Y?

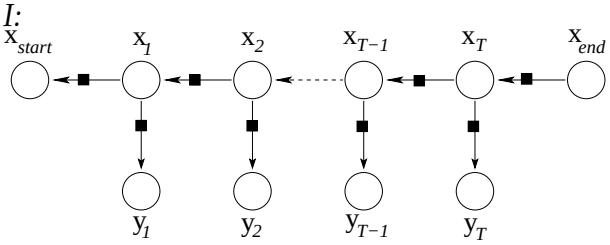
(g) [8 marks] Now suppose that you are certain *a priori* that an observable variable W is a cause of X. What relationships (dependencies and independencies) would your observations need to exhibit in order for you to conclude that X *is* a cause of Y? Explain your answer.

Question 5. [45 marks]

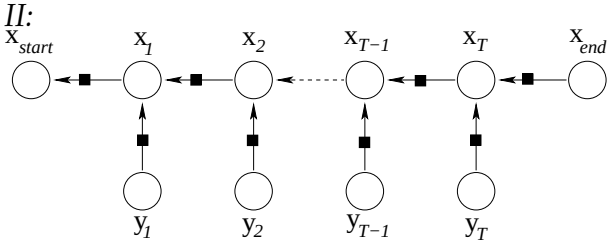
(a) [12 marks] Consider the simple HMM with start and end states shown on the right, in which the black squares refer to factors. Could either of the two graphical models shown below be used to model the *same joint distribution* as the one defined by this graph? Explain your answer.



I:



II:

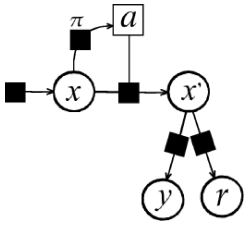


(b) [8 marks] Consider searching a continuous space for the maximum of some function $F(x)$ amounting to the height of a surface at point x. Suppose you start from a random position and make a path of n steps of fixed size. Making these steps "up" the gradient $\nabla_x F(x)$ must lead to a higher point on the surface that is at least as high as making the steps *down* the gradient, from any given starting position. How can this be reconciled with the No Free Lunch theorem, which states that all search algorithms perform the same when averaged over all surfaces?

(c) [8 marks] Explain why the sum-product algorithm is incorrect for loopy graphs.

(d) [6 marks] Explain how variables can be merged in order to remove loops, and point out the main disadvantage involved in doing so.

Consider the graphical model shown on the right, which relates states (x and x'), action (a), observation (y) and reinforcement signal (r). The black squares refer to factors, one of which is a "policy" π. Others are $p(x)$, $p(x'|x,a)$, $p(y|x')$ and $p(r|x')$.



(e) [4 marks] Assuming that states and actions are discrete, give an expression for the overall transition probability $M_{x \to x'}$.

(f) [7 marks] Now suppose that the above graph is repeated over time, so that x' becomes x for the next time step and so on for a long sequence of actions and observations. How could you go about finding the expected reward under the current policy?

****