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

EXAMINATIONS — 2005

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.

Use the number of marks for sub-questions as a guide to the depth of answer required.

Non-electronic Foreign-English language dictionaries are permitted.

Question 1. Sampling and various other topics [45 marks]

(a) [8 marks] Explain how it is possible to *integrate* a function $f(\mathbf{x})$ weighted by some probability distribution $p(\mathbf{x})$, merely by taking *samples* of it.

(b) [12 marks] Discuss the differences between standard neural networks trained with backpropagation and Bayesian neural networks implemented using the Metroplis algorithm. (For example, how are they implemented, and what are their strengths and weaknesses?)

(c) [5 marks] Consider a neural network being used to classify input patterns into 1-of-n classes, where $n \geq 3$. Describe the "soft-max" approach to generating multiple outputs from such a network.

(d) [8 marks] The philosopher Daniel Dennett has characterised learning systems in terms of what he called a "tower of generate and test". Describe what he means by Darwinian, Skinnerian and Popperian creatures.

(e) [8 marks] Line-search is a method applicable to optimisation problems where the gradient of the "goodness" function is available at sample points, as well as its value. Explain the problem with using the basic form of line-search (in which searches are carried out in the direction of the local gradient), and how it is solved in *conjugate gradient* line-search.

(f) [4 marks] Describe how insertions are handled in profile HMMs.

Question 2. Bayesian inference [45 marks]

(a) [10 marks] Explain the following - "In Bayesian machine learning, hypotheses are compared via their posterior probabilities. But in a fully Bayesian analysis there is no need to choose between competing hypotheses".

(b) [10 marks] In search problems one has a space of possible solutions $\{x\}$ and a way of arriving at the value $F(x)$ of any given solution. For example, if a search algorithm is allowed to take samples x and their values $F(x)$, then after n samples the algorithm's *best* sample is, say, $F^\star$. One consequence of the No-Free-Lunch theorem is that all search algorithms have the same expected value of $F^\star$, given a new search surface which nothing is known about *a priori*.
How is it possible that hill-climbing has the same performance on average as its opposite, hill-*descent*?

For the next four subquestions, suppose that you want to guess the next three numbers in the sequence $[-1, 3, 7, ...]$

- Hypothesis H_1 is that the sequence is generated by a process $x_t = c_0 + c_1 t$ with two parameters. Using $c_0 = -1$ and $c_1 = 4$ fits the data exactly and predicts $[11, 15, 19]$ for the next three numbers.
- Hypothesis H_2 is that the sequence is generated by a process $x_t = c_0 + c_1 t + c_2 t^2 + c_3 t^3$ with four parameters. Using $c_0 = -1$, $c_1 = 6$, $c_2 = -3$ and $c_3 = 1$ also fits the data exactly, but predicts $[17, 39, 79]$ for the next three numbers in the sequence.

(c) [7 marks] Briefly describe how a Bayesian theorist would go about choosing between these two hypotheses.

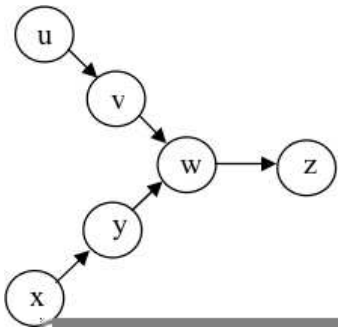
(d) [7 marks] Even with a prior belief $P(H_1) = P(H_2)$, a Bayesian theorist will tend to prefer the simpler of two models if both account for the data equally well. Explain the way in which Bayesian model comparison embodies "Ockham's razor".

(e) [4 marks] Describe how a Bayesian theorist would make predictions using *both* hypotheses instead of choosing just one.

(f) [7 marks] Describe an alternative procedure to predict the next three numbers, using cross-validation instead of the probabilistic approach.

Question 3. Graphical models [45 marks]

Consider the belief network shown at right.



(a) [3 marks] Write down the log probability of the joint distribution, as a sum over factors.

(b) [3 marks] Assuming all variables take exactly 3 values, how many probability values need to be specified / learned?

(c) [3 marks] Are variables `u` and `x` dependent or independent given an observation of `z`? Explain (in words rather than equations) how you arrived at your answer.

(d) [2 marks] "In a belief net, a variable is conditionally independent of *all* other variables in the net given its Markov Blanket". What is meant by the term `Markov Blanket`?

(e) [5 marks] Explain *why* the statement quoted in the previous subquestion is true.

(f) [10 marks] Discuss whether "explaining away" can occur in undirected graphical models. If so, how, and if not, why not? [It may be useful to include a simple example by way of explanation.]

(g) [12 marks] The probability propagation algorithm uses message passing to arrive at conditional distributions for variables in graphs with no cycles. Describe THREE ways of performing such inference in graphs that *do* contain cycles ("loopy" graphs).

(h) [7 marks] What is the relationship between the Viterbi algorithm and the forward pass of standard probability propagation in an HMM (Hidden Markov Model)? (You *may* find it useful to refer to the most likely path up to time t ending with state i as $\text{path}_{1..t,i}$ or similar).

Question 4. Reinforcement learning [45 marks]

Given an optimal policy, the long-term discounted values ("values to go") of states obey the following recursion, known as the Bellman equation:

$$V^{\star}_{\mathbf{s}} = \max_{a} \sum_{s'} P_{\mathbf{s}'|\mathbf{s},a} \ [R_{\mathbf{s}'} \ + \ \gamma V^{\star}_{\mathbf{s}'}]$$

Here $\mathbf{s}$ represents the current state, a is an action taken in the current state, and $\mathbf{s}'$ is a state at the next time-step. R is the expected immediate reward. The '$\star$' indicates that this is the *optimal* value function.

(a) [7 marks] Give a different expression for $V^{\star}_{\mathbf{s}}$ that is *not* recursive.

(b) [7 marks] Describe 'Value Iteration'.

(c) [7 marks] Value Iteration requires knowledge of $P_{\mathbf{s}'|\mathbf{s},a}$. Describe how you could learn a model of $P_{\mathbf{s}'|\mathbf{s},a}$.

(d) [7 marks] Value Iteration also requires knowledge of $R_{\mathbf{s}}$. Describe an online (as opposed to batch) learning algorithm for estimating $R_{\mathbf{s}}$.

(e) [10 marks] A discount rate γ is used in both 'direct' reinforcement learning and the various forms of TD (temporal difference) learning, but this number is often chosen in way that is *ad hoc*. One way to think of γ is that it captures the fact that an agent's future may not go "according to plan". Using this interpretation, can you think of a way to *learn* an appropriate value to use for γ, over many episodes?

(f) [7 marks] In the context of reinforcement learning, what is the "exploit-vs-explore" dilemma?

Question 5. Mixture models and Gaussian processes [45 marks]

(a) [8 marks] Describe the use of the EM algorithm for unsupervised learning of K clusters in data, under a mixture of Gaussians model. Assume K is known.

(b) [3 marks] Suggest a method for identifying K if it is not known *a priori*.

Consider a learning system being used to predict the distribution of a single real-valued output value, given the current input. The learner is provided with a training set of examples consisting of input vectors and their observed (scalar) outputs. One approach to such a problem is to use Gaussian processes, whose "output" or prediction is characterised by:

$$\texttt{mean} = \mathbf{k}^T\mathbf{C}^{-1}\mathbf{t} \qquad \texttt{variance} = \kappa - \mathbf{k}^T\mathbf{C}^{-1}\mathbf{k}$$

In these equations $\mathbf{C}_{ij} = Cov(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{k}$ is the vector of individual covariances $\mathrm{k}_j = Cov(\mathbf{x}_j, \mathbf{x})$ between the new input $\mathbf{x}$ and each of those in the data set. κ is $Cov(x, x)$.

(c) [8 marks] Describe one advantage of using Gaussian process predictors, compared to Bayesian neural networks.

(d) [8 marks] Describe one *dis*advantage of using Gaussian process predictors, compared to Bayesian neural networks.

Now suppose that a "mixture-of-experts" system uses Gaussian process predictors as its "experts".

(e) [8 marks] Describe how the combining of multiple Gaussian process predictions in this way could improve the quality or usefulness of predictions over an alternative system that used only one Gaussian process.

(f) [10 marks] Describe the steps you would need to carry out in making a prediction using this system.

****