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

EXAMINATIONS — 2011

MID-YEAR

COMP 421

Machine Learning

Time Allowed: 3 Hours

Instructions: There are EIGHT questions to choose from, each worth 30 marks.

Answer SIX questions (180 marks, *i.e.* one mark per minute).

If you complete more than 6 questions, only your best 6 will be counted.

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. Graphical Models (I) [30 marks]

Consider the following scenario. Your house has a burglar alarm, which is triggered by burglars breaking in, but also by earthquakes. There are two neighbouring houses, one occupied by John and the other by Mary. If the alarm is triggered, it makes a noise, and there is some chance that one of your neighbours will respond to it by phoning you at work. You can model this situation using a graphical model with nodes **J**`ohn`, **M**`ary`, **B**`urglar`, **E**`arthquake` and **A**`larm`.

(a) [5 marks]
Draw the belief network (with nodes J, M, B, E, A) *and* its the corresponding factor graph.

(b) [4 marks]
Write down the joint probability $P(J, M, B, E, A)$ as a product of factors.

(c) [4 marks]
Prove that your answer to (*b*) implies that Burglar $\perp\!\!\!\perp$ Earthquake (independence).

Suppose we collect a data set of (x, y, z) tuples, and find that $(0, 0, 0)$ occurs 15 times, $(0, 0, 1)$ occurs 7 times, and so on, as the following table shows:

x	y	z	raw count	total count	joint: $P(x, y, z)$	factor: $P(z\|x, y)$
0	0	0	15			
0	0	1	7			
0	1	0	10			
0	1	1	19			
1	0	0	14			
1	0	1	0			
1	1	0	0			
1	1	1	1			
					1.0	

(d) [5 marks] By (copying and) filling in the relevant parts of the table, and including a "pseudocount" of 1, estimate

- $P(x = 0, y = 0, z = 0)$, and
- $P(x = 0, y = 0, z = 1)$.

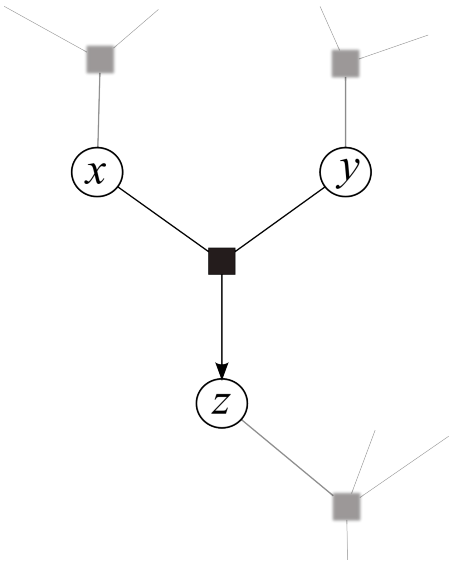
Hint: leave calculations as fractions rather than converting to decimal.

(e) [5 marks] Similarly, estimate $P(z = 0|x = 0, y = 0)$, and $P(z = 1|x = 0, y = 0)$.

(f) [7 marks] Consider training a neural network on a supervised learning problem, using a Training Set of input-output pairs, while keeping a Test Set of similar items. Are there circumstances in which using the conjugate gradient algorithm instead of "vanilla" (standard) back-propagation might lead to *worse* results on a Test Set? Explain your answer.

Question 2. Graphical Models (II) [30 marks]

Consider this fragment from a belief network involving variables x, y and z:



(a) [10 marks]
Copy the picture and show all the messages to and from the central factor on it, in the form of (for example) $P(u, \text{obs})$, $P(u \mid \text{obs})$, or similar. You can use the nomenclature obs_x to refer to observations made in the network at *or beyond* node x, from the perspective of the central factor.

(b) [15 marks]
From first principles if possible, *derive* an expression for $p(x, y, z \mid \text{obs})$ in terms of the above messages and the current factor. The first steps have been completed below:

$$\begin{aligned} p(x, y, z \mid \text{obs}) &\propto p(x, y, z, \text{obs}) \\ &= p(x, y, z, \text{obs}_x, \text{obs}_y, \text{obs}_z) \\ &= p(z, \text{obs}_z \mid x, y, \text{obs}_x, \text{obs}_y)\ \ p(x, y, \text{obs}_x, \text{obs}_y) \end{aligned}$$

Note: if you can't derive it, but know the answer, provide that answer.

(c) [5 marks] What is the message from z to the factor if no nodes in the network have been observed, and what happens to that message if z is observed?

Question 3. Various questions [30 marks]

(a) [5 marks] In a belief net, is it easy, or hard, to *evaluate* the probability of the hidden state, given some visible state, $P(h|v)$? Explain your reasons.

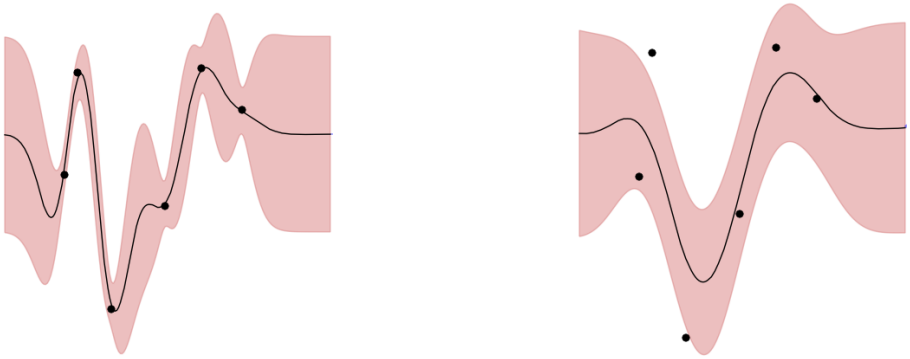
(b) [3 marks] The "Naive Bayes" assumption is that variables $v_1 \ldots v_n$ become conditionally independent of one another, given a classification value ("class") c. Express the Naive Bayes assumption as a graphical model.

(c) [6 marks] In machine learning models where the "output" is assumed to involve Gaussian noise, it is often the case that the learning rule for parameters in the model involves an "error" term of the form

$$(\text{target output} \; - \; \text{predicted output})$$

Explain the connection between the Gaussian noise model and this term in the learning rule.

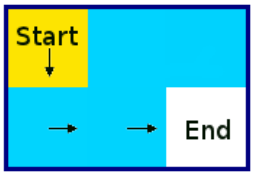
(d) [6 marks] In the course we looked briefly at the Gaussian Process framework for regression problems. The figure below shows the end result of two runs of an algorithm that finds optimal hyperparameters (by gradient ascent of the log likelihood). The 6 dots are data points, the line shows the Gaussian Process predictive mean and the shaded area indicates its variance. The two runs began with slightly different initial values for the hyperparameters, but in all other respects were the same. Explain how and why the two solutions are different.



(e) [10 marks] Boltzmann machines are stochastic neural nets that can be used for unsupervised learning tasks. Learning in a fully connected Boltzmann machine involves the repetition of two phases. Why are Boltzmann machines so slow to train? (There may be several reasons one could give).

Question 4. Reinforcement learning [30 marks]

Consider a reinforcement learner in an extremely small "grid world" consisting of an array of cells, having 2 rows by 3 columns. The walls surrounding these 6 cells are impenetrable and actions that would take the agent through them are not considered. The agent always begins in the top-left cell, and receives reinforcement of -1 per time step in every cell except the bottom-right one (where it receives zero), which also acts as an "exit", prompting a re-start. The figure shows this environment and an example trajectory taken through it.



(a) [10 marks] Suppose an agent follows the trajectory shown for one episode, and employs the SARSA learning algorithm. Assuming that the initial Q values are *all* set to 1 and that discounting of $\gamma = 0.5$ is used, what will the new Q values be, along the trajectory shown? (Show your working).

(b) [5 marks] What is the difference between the SARSA and Q-learning algorithms, and what is the significance of the difference?

(c) [5 marks] In reinforcement learning, an obvious way to use Q values is to choose (from among the available actions) the action that has the largest current Q-value. What is the problem caused by using Q values this way?

(d) [5 marks] Describe how one could use the Q-values learned by an agent to drive "soft-max" action selection.

(e) [5 marks] Function approximators (such as neural networks) can be used to represent the Q-values used in reinforcement learning.
Outline the potential benefits, and problems, with this approach.

Question 5. Interpretation of a published paper [30 marks]

This question tests whether you can take what you have learned in the course and use it to understand material in the research literature.

An interesting looking paper recently appeared with the title "Sum-Product Networks: a new deep architecture" (H. Poon and P. Domingos, 2011).

The beginning of the INTRODUCTION section for this paper is given below, with some words shown in bold for the purposes of the questions that follow:

> "The goal of probabilistic modeling is to represent probability distributions compactly, compute their marginals and modes efficiently, and learn them accurately. Graphical models **represent distributions compactly as normalized products of factors**: $P(X = x) = \frac{1}{Z}\prod_k \phi_k(x_{\{k\}})$, where $x \in \mathcal{X}$ is a d-dimensional vector, each potential ϕ_k is a function of a subset $x_{\{k\}}$ of the variables (its scope), and $Z = \sum_{x\in\mathcal{X}} \prod_k \phi_k(x_{\{k\}})$ is the *partition function*. Graphical models have a number of important limitations. First, there are many distributions that admit a compact representation, but not in the form above. (For example, the uniform distribution over vectors with an even number of 1s.) **Second,** inference is still exponential in the worst case. **Third,** the sample size required for accurate learning is worst-case exponential in scope size. **Fourth,** because learning requires inference as a subroutine, it can take exponential time even with fixed scopes (unless the partition function is a known constant, which requires restricting the potentials to be conditional probabilities)."

(a) [6 marks] The authors state that "Graphical models represent distributions compactly as normalized products of factors". Why is this a compact way to represent a distribution?

(b) [6 marks] Explain what you think the authors mean by "Second, inference is still exponential in the scope size".

(c) [6 marks] Explain what you think the authors mean by "Third, the sample size required for accurate learning is worst-case exponential in scope size."

(d) [6 marks] The authors then state: "Fourth, because learning requires inference as a subroutine, ...". What do you think they mean when they say that learning involves inference as a subroutine?

(e) [6 marks] On this last point, they add the caveat "(unless the partition function is a known constant, which requires restricting the potentials to be conditional probabilities)". Explain what you think they are refering to in this comment.

Question 6. Value Iteration; Hidden Markov Models (I) [30 marks]

Value Iteration iterates the following equation in order to arrive at an optimal value function, and hence an optimal policy in MDPs (Markov Decision Processes).

$$V_{\mathbf{s}}^{k+1} = \max_a \sum_{\mathbf{s}'} P_{\mathbf{s}'|\mathbf{s},a} \left[R_{\mathbf{s}'} + \gamma V_{\mathbf{s}'}^k \right]$$

(a) [10 marks] Outline the conditions under which Value Iteration is a feasible approach to the problem of arriving at a sensible policy.

Hidden Markov models encode relationships between observations $y_1, y_2, \ldots, y_t, \ldots, y_T$ and hidden states $x_0, x_1, \ldots, x_t, \ldots, x_T$. We can use the terminology $y_{1:t}$ as short-hand to refer to $(y_1, y_2, \ldots, y_t)$.

(b) [5 marks] What is the assumption about $P(x_t \mid x_{0:t-1})$ that is made in an HMM?

Consider Hidden Markov models for speech modelling. Hidden nodes take discrete values (corresponding to a finite number of "phonemes"). Visible nodes are encodings for how those phonemes actually sound in terms of a distribution over audible frequencies (you don't need to know the details of this encoding to be able to answer this question). This model allows us to represent sequences of phonemes and the sounds that result from them. However, the "speech" obtained by *generating* from such a model sounds very poor, because real speakers vary their pitch (average frequency) over the course of a spoken sequence - for example many speakers "go up" at the end if they are asking a question.

Similarly real speech also has volume, which may vary during the sequence. These quantities (pitch and volume) vary over the course of an utterance, just as the phonemes do.

(c) [8 marks] Draw the structure of an augmented HMM that represents these other variables and how they interact, assuming they have Markov structure to their dynamics.

(d) [7 marks] Outline the issues this raises for inference and learning in such a model.

Question 7. Hidden Markov Models (II) [30 marks]

In HMMs there are several problems that one might want solutions for:

1. **model evaluation**: what is $p(y_{1..T})$ under the current model?
2. **prediction**: what is $p(y_{t+1}|y_{1..t})$?
3. **state estimation**: what is $p(x_t|y_{1..T})$?
4. **trajectory estimation**: what is the most likely trajectory through states, given $y_{1..T}$?
5. **learning**: how can the factors $p(x_1)$, $p(x_t|x_{t-1})$ and $p(y_t|x_t)$ be improved in the light of data?

Describe how each of these problems are approached.

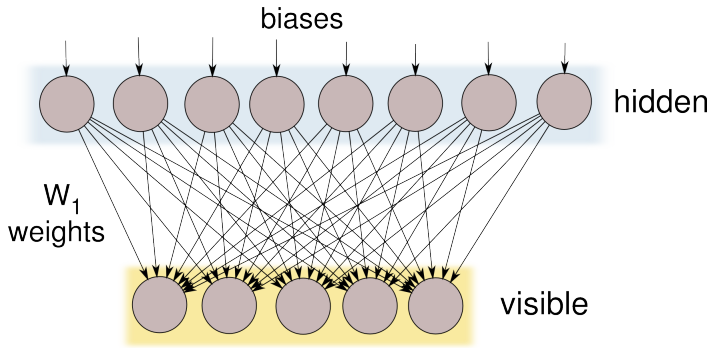
Question 8. Bayesian model comparison; Deep belief nets [30 marks]

Suppose we have a data set $\mathcal{D}$, and a number of generative models $M_1, M_2, \ldots, M_N$ to choose from. The models are from a model space $\mathcal{H}$, and each model M_i involves a vector of real-valued parameters $\{\theta\}_i$.

(a) [4 marks] Give an expression for $P(M_i|\mathcal{D})$ involving θ.

(b) [6 marks] Outline in general terms how the Metropolis algorithm could be used to help someone *choose between the available models*.

Sigmoid belief nets are stochastic neural nets that can be trained using a data set of vectors describing their "output" nodes. The diagram below shows a sigmoid belief net with one layer of hidden units:



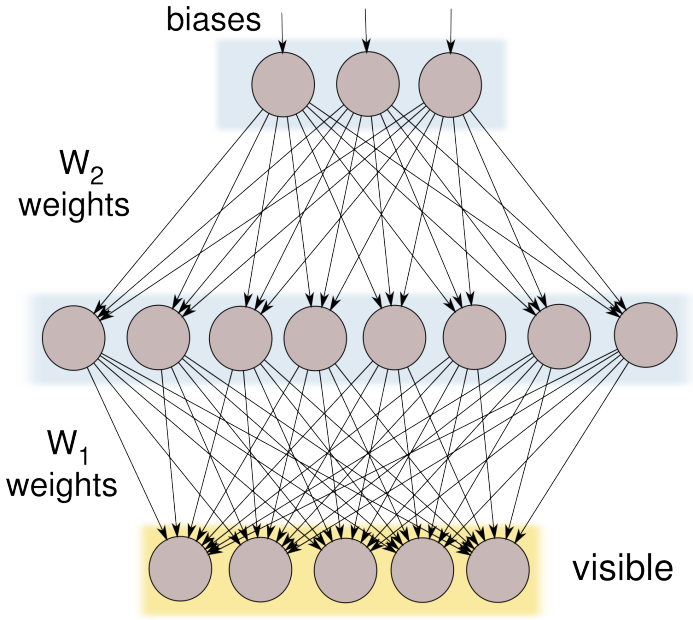
(c) [6 marks] Given a large data set of input patterns, such a network could be trained to maximize the likelihood by a stochastic form of the EM algorithm. In contrast to the Boltzmann machine, for the above network there would be no "sleep" phase to the learning algorithm. Why not?

(d) [6 marks] Despite only requiring a "wake" phase, learning is difficult in such networks however. Explain why this is so.

(Question 8 continued on next page)

(Question 8 continued)

The diagram below shows a sigmoid belief net with two layers of hidden units:



One way to train such a network is to learn the bottom layer on its own. A "training set" for the second layer is then generated by going through the (original) training set and sampling from the posterior over the first layer hidden units. The second layer weights are then trained on this data, which should result in an improvement in the overall generative model's performance.

(e) [8 marks] *Even if* the bottom layer was trained successfully, in most cases this procedure does *not* lead to substantial improvements in the overall generative model. Why not?

****