

EXAMINATIONS — 2009

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]

(a) [14 marks] The sea squirt is a simple creature which has two phases to its life-cycle: an active juvenile stage when it has a brain (or sorts) and swims, and an adult stage when it attaches itself permanently to a rock. The first thing it does after attaching is dissolve most of its own brain.

Why do you think simple mobile creatures might need more neural tissue than anchored ones?

Consider three binary variables a , b and $c \in \{0, 1\}$ having the joint distribution given in the table below:

a	b	c	$p(a, b, c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

(b) [4 marks] Evaluate the distributions $p(a)$, $p(b|c)$ and $p(c|a)$

(c) [2 marks] For this data, it turns out that $p(a, b, c) = p(a)p(c|a)p(b|c)$. Draw the corresponding belief net.

(d) [4 marks] By showing your working, confirm that $p(a, b, c) = p(a)p(c|a)p(b|c)$ for two specific cases: $(a = 1, b = 0, c = 0)$, and $(a = 1, b = 0, c = 1)$.

(Question 1 continued on next page)

(Question 1 continued)

Boltzmann machines are stochastic neural nets that can be used for unsupervised learning tasks (for instance). Learning in a fully connected Boltzmann machine involves the repetition of two phases.

(e) [6 marks] Describe these two phases.

(f) [5 marks] Boltzmann machines can be very slow to train, but two recent insights have enabled them to become practical learning devices. One of these is to restrict the architecture to two layers and eliminate all the connections within each layer. How does restricting the architecture in this way help?

(g) [5 marks] Describe the *other* insight that has enabled Boltzmann machines to be practical learning devices.

(h) [5 marks] Consider a restricted Boltzmann machine and a data set \mathcal{D} of training examples. Why is it difficult to evaluate the log likelihood $\log P(\mathcal{D})$?

Question 2.

[45 marks]

(a) [2 marks] One way to express the statement “ X and Y are conditionally independent, given Z ” is to write

$$P(X, Y|Z) = P(X|Z) P(Y|Z)$$

What is another way, involving $P(X|Y, Z)$?

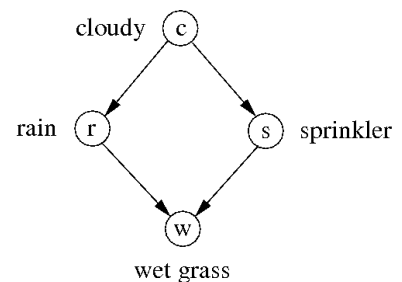
(b) [2 marks] Give a general expression for the joint probability of a belief net (directed graphical model), in terms of the probability of each node given its parents in the graph, $P(x_i|\text{parents}_i)$.

Consider the belief network shown at right.

(c) [2 marks] Assuming all variables take exactly 2 values, how many probability values need to be specified / learned?

(d) [4 marks] Write down the log probability of the joint distribution, as a sum over factors.

(e) [10 marks] Describe THREE ways of performing inference (in the general case: for any node, given observations of any other nodes) in this graph. For each, give an advantage and a disadvantage of using it.



(f) [5 marks] The Metropolis-Hastings algorithm requires that we specify a proposal distribution. What are the main issues surrounding the choice of proposal distribution?

(Question 2 continued)

(g) [5 marks] Consider a deep belief net composed of sigmoid units (stochastic ‘neurons’) and constructed by the greedy, layer-wise procedure we looked at in lectures and in Assignment 2.

For a given input pattern \mathbf{v} , is it easy or difficult to calculate $P(\mathbf{v})$, and why?

(h) [5 marks] Explain how to draw a sample pattern \mathbf{v} from the visible (bottom) layer, under the generative model.

(i) [5 marks] A deep belief network can be “unfolded” and thought of as an auto-encoder, in which case the central bottle-neck layer performs *dimensionality-reduction* of the input patterns. How could you make it do *clustering* instead?

(j) [5 marks] Describe ONE method for untying or otherwise fine-tuning the weights in a deep belief network that has been pre-trained using the greedy layer-wise procedure.

Question 3.

[45 marks]

(a) [4 marks] In belief nets, what is meant by the term Markov Blanket, and what is its significance?

(b) [6 marks] Belief nets are said to simplify the full joint distribution over a set of variables by making conditional independence assumptions. Briefly explain how this is so, by giving an example based on the graph corresponding to the following factorization:

$$p(w, x, y, z) = p(w) p(x|w) p(y|x) p(z|x)$$

(c) [6 marks] What does the term “explaining away” mean, in the context of Belief Nets?

Suppose one takes a collection of coins, K of them in total. Each coin c has a certain probability of coming up Heads (its “bentness”), b_c . Consider a data set \mathcal{D} arising from the following process, which generates a set of sequences:

Do several times:

1. a coin c is chosen at random, with probability μ_c
2. it is thrown it a few times, generating a sequence \mathbf{S} , such as HHTTHTHTTTHT . . . HHHHT

(d) [3 marks] Give a general mathematical expression for $P(\mathbf{S})$, the probability of generating arbitrary sequence \mathbf{S} .

(e) [3 marks] Suppose an observer of these sequences knows how many coins are involved but not their μ_c and b_c values, so they estimate values (μ_c^* and b_c^*) for these.

Give an expression for $\log P(\mathcal{D})$ under the model given by the μ_c^* and b_c^* values.

(f) [8 marks] Outline how you could improve the μ_c^* and b_c^* values based on a data set of sequences.

(Question 3 continued on next page)

(Question 3 continued)

In reinforcement learning the SARSA learning rule is

$$\Delta Q_{x_t, a_t} = \eta [r_{t+1} + \gamma Q_{x_{t+1}, a_{t+1}} - Q_{x_t, a_t}]$$

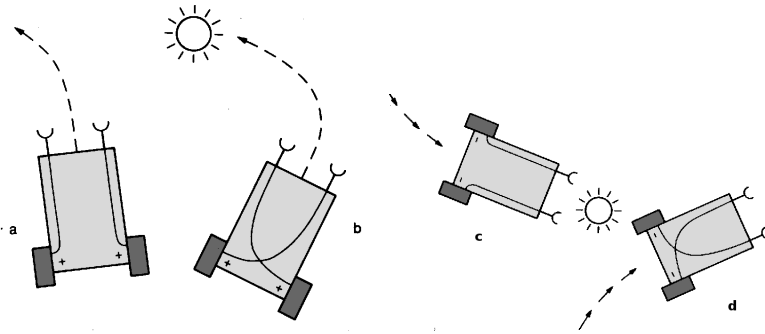
(g) [5 marks] What is the “TD trick” being used here?

(h) [5 marks] The algorithm called “Q-learning” is different from SARSA. Describe the difference.

(i) [5 marks] What is the advantage of using Q-learning compared to SARSA?

Question 4.

[45 marks]



The figure shows several "Braitenberg vehicles", which are simple robots that exhibit very basic behaviours. The difference between those shown here can be seen in terms of different values on weighted connections going from both sensors to both actuators (motors).

Suppose that the light source is constantly moving in some non-trivial way, and that a scalar signal r is available to the agent at every time step, with $r = -1$ if the vehicle is facing toward the light (within some tolerance such as $\pm 45^\circ$ for example), and $r = 0$ otherwise.

- (a) [10 marks] Describe a *policy gradient* ("direct") learning algorithm that would allow the agent to learn to avoid the light over time, if it starts off with random initial weights.
- (b) [5 marks] Outline one major *disadvantage* of the algorithm you described above. Note this question is about the algorithm, not the architecture of the vehicles.
- (c) [5 marks] Suggest an enhancement to the learning architecture used by the agent (the mapping between sensors and actuators) that would help.
- (d) [5 marks] What is the relationship between the sum-product algorithm and the max-sum algorithm in belief networks?
- (e) [5 marks] Hidden Markov models (HMM) are equivalent to belief nets, with the addition of some assumptions about the nature of the graph. What are these extra assumptions?

(Question 4 continued on next page)

(Question 4 continued)

The following questions concern the *general* Boltzmann machine (with all possible connections included). The joint probability distribution of a Boltzmann machine with neuron states \mathbf{s} is

$$P_{\mathbf{s}} = \frac{e^{-E_{\mathbf{s}}}}{Z} \quad \text{where "energy"} \quad E_{\mathbf{s}} = -\frac{1}{2} \sum_{i,j} s_i s_j w_{ij}$$

with symmetric weights and no "self" connections.

The difference between the energy of a joint state in which the i^{th} neuron takes $s_i = 1$ and the same joint state except with $s_i = 0$ is

$$\Delta E = \phi_i$$

where ϕ_i is the weighted sum of inputs to the i^{th} neuron.

(f) [2 marks] What is Z ? (Give an expression).

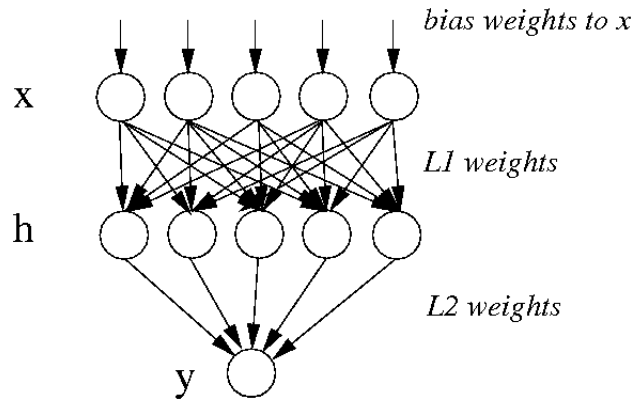
(g) [5 marks] Although Z is hard to evaluate, Gibbs sampling can generate samples from $P_{\mathbf{s}}$ regardless of Z 's value. So *why is it a problem* that evaluating Z is hard?

(h) [8 marks] Prove that the Gibbs sampling algorithm in this situation reduces to stochastic sigmoidal neurons (i.e. neurons that output 1 with probability given by the sigmoid function of ϕ).

Question 5.

[45 marks]

Consider the following architecture composed entirely of stochastic binary neurons:



In the generative (top-down) model, each neuron computes a sigmoid function of the sum its weighted inputs, and uses this as the probability of producing a 1 (rather than a 0) as output. The top layer only has “bias” inputs, equivalent to an extra input (not shown) which is always set to 1.

This architecture could be used to perform classification of input patterns x into two classes given by y . Suppose that you have a training set of (x, y) pairs, and a test set of input patterns x^* that you want to classify.

- (a) [5 marks] Given an x^* pattern in the test set, and a certain set of trained weights, how would you go about using the classifier to produce a distribution over y ? *Hint: you don't need to do any Gibbs sampling.*
- (b) [10 marks] Describe how would you go about learning better values for the weights in this classifier.
- (c) [8 marks] Some of the x^* patterns in the *test* data are missing some of their values. Is this a problem for the classifier? Explain why or why not.

(Question 5 continued on next page)

(Question 5 continued)

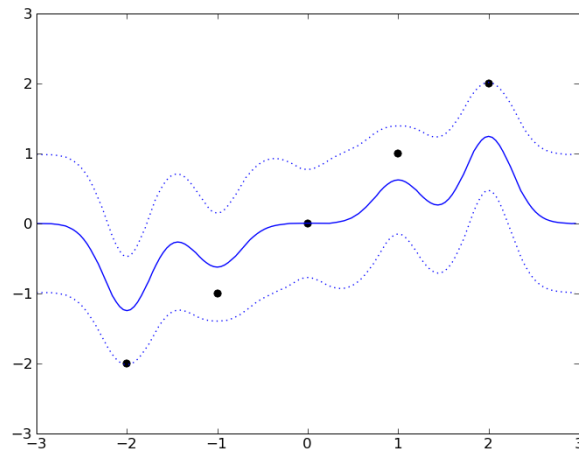
A Gaussian process with hyperparameters θ provides a mean and variance for a scalar output value y given an input vector \mathbf{x} , conditional on a training set of other (\mathbf{x}, y) pairs.

An example from the lectures is shown on the right.

A common procedure is to set θ to values that maximize the log likelihood of the training set.

A *mixture* of (say) K such Gaussian processes would be more powerful model, because now the distribution over y could be modelled as consisting of 2 or more modes, not just one.

Each Gaussian process (indexed k) in the mixture would have its own hyperparameters θ_k , and a mixing proportion μ_k .



(d) [5 marks] In the style of the picture shown here, draw an example showing the new predictive distribution given some data.

(e) [10 marks] Suggest how you might go about “training” this model. Where possible give details, and otherwise indicate the main ideas.

(f) [7 marks] Discuss how ONE of the following relates to learning by machines:

- Price’s basic equation:

$$\bar{w}\Delta\bar{z} = \text{Cov}(w, z)$$

- Hill climbing and exploit-versus-explore as *emergent* properties of a generalised approach to optimization.
