YOUR NAME:
YOUR STUDENT NUMBER:

UNIVERSITY OF TORONTO
Faculty of Arts and Science
APRIL/MAY 2011 EXAMINATIONS
CSC 321 H1S

Duration 2 hours
No Aids Allowed

Answer ALL 10 questions in Part A. Answer EXACTLY 4 questions in part B. Each question in part A is worth 2 points. Each question in part B is worth 5 points.

PART A (20 points)

1. Consider a neural network with two very similar inputs. When using weight decay with a squared weight penalty term, which of the following options is favoured more and why: a) Putting a weight of w on one connection and a weight of 0 on the other. b) Putting a weight of w/2 on both connections.

2. What does the word "conjugate" mean in the expression "conjugate gradient".

3. In supervised learning, we usually try to maximize the sum of the log probabilities of the correct answers. Why is this better than maximizing the sum of the probabilities of the correct answers?

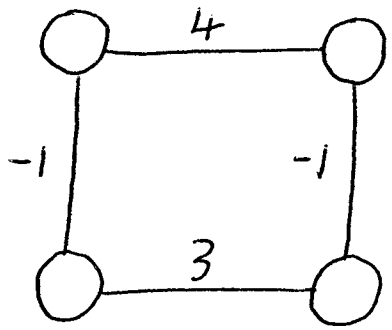
4. Describe as precisely as you can the relationship between Principal Components Analysis and autoencoders.

5. Describe a property of a dataset that makes it particularly appropriate to use a "mixture of experts" model?

6. Under what circumstances is the EM algorithm for fitting a mixture of Gaussians exactly the same as the K-means algorithm?

7. What is a support vector? Be as precise as you can without using math.

8. In the Hopfield net shown below, the units have binary states of 1 or 0, and the biases are all zero. Fill in the binary state that has the second lowest energy.



9. In a Restricted Boltzmann Machine, there are no connections between the hidden units. What is achieved by restricting the connectivity in this way?

10. Explain very briefly how images similar to a query image can be retrieved very quickly when using "semantic hashing" for image retrieval.

PART B (20 points)

Answer EXACTLY 4 questions in this part.

1. Hinton and Shallice designed a neural network to convert strings of visually perceived letters into vectors of semantic features.

   a) (*2 points*) Explain why adding noise to the bottom-up input coming from the letter-detectors can cause the word PEACH to be read as "apricot".

   b) (*2 points*) When training the network, the semantic features of a word can be used as the desired states of the semantic units for the last time-step or for the last three time-steps. Which works better and why?

   c) (*1 point*) Explain why severe damage to the "clean-up" part of the system can allow the system to recognize that the word "HAM" refers to food without knowing that it refers to ham.

2. a) (*2 points*) Suppose that the perceptron convergence procedure is used to train a binary threshold unit on a set of training cases that can be learned perfectly. What quantity is guaranteed to decrease every time the weights are changed?

b) (*3 points*) Suppose that a binary threshold unit has a non-adaptive threshold that is set at zero. It starts with weights of (+1, 0, -2) and it is trained by presenting it with the following sequence of input vectors and desired outputs. Show the weights that it has after each training case, if it is trained using the perceptron convergence procedure.

input vector	desired output	Weights		
		+1	0	-2
(1 0 1)	-> 1			
(0 1 1)	-> 1			
(1 0 0)	-> 0			
(1 0 1)	-> 1			
(1 0 0)	-> 0			

3. a) (*1 point*) Consider a one-dimensional Gaussian with a standard deviation of $\sqrt{2}$. Write down the expression for the probability density at a distance of d from the mean of the Gaussian.

b) (*3 points*) Suppose we are fitting a mixture of two one-dimensional Gaussians to two datapoints with values of -1 and +1. The Gaussians have equal mixing proportions of 0.5 and their initial standard deviations are $\sqrt{2}$. They start with means of -1 and +1 (i.e. they are in the same locations as the two datapoints). After one iteration of the EM algorithm, what are the new means? You should give your answer in terms of expressions that contain the mathematical constant e.

c) (*1 point*) What are the new mixing proportions of the two Gaussians?

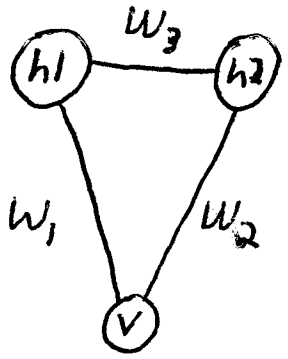
4. a) (*3 points*) Explain in words how a support vector machine uses a kernel function to save computation when fitting a linear model in a very high-dimensional feature space.

b) (*1 point*) State one way in which a support vector machine is better than a neural network.

c) (*1 point*) State one way in which a neural network is better than a support vector machine.

5. a) (*1 point*) The Boltzmann Machine shown below has one visible unit, v, and two hidden units, h1, h2. The biases are all zero. Write down the energies of all eight possible states of the network.

v	h1	h2	Energy
1	1	1	
1	1	0	
1	0	1	
1	0	0	
0	1	1	
0	1	0	
0	0	1	
0	0	0	



b) (*2 points*) If $w_1 = w_2 = w_3$ and $e^{w_1} = 2$, compute the probability, when the network is generating data, that the visible unit is on.

c) (*2 points*) If the network is being trained on a single data vector in which the visible unit is on, what is the derivative of the log probability of the data with respect to w_1?

6. a) (*2 points*) Very briefly describe the "contrastive divergence" learning procedure for a Restricted Boltzmann Machine.

b) (*1 point*) How does the contrastive divergence procedure need to be modified to ensure that it computes a noisy but unbiased estimate of the gradient of the log probability of an individual training case?

c) (*2 points*) Contrastive divergence learning computes a biased estimate of the gradient of the log probability of an individual training case. Is this estimate more biased at the beginning of learning or at the end of learning. Explain your answer.

7. a) (*3 points*) Explain how Restricted Boltzmann Machines can be combined to learn a deep feedforward neural network.

b) (*1 point*) Briefly explain why pre-training a feedforward neural network using a stack of RBMs causes back-propagation to generalize better to new data.

c) (*1 point*) Briefly explain why pre-training a feedforward neural network using a stack of RBMs causes back-propagation to learn faster.

8. a)(*1 point*) Explain the difference between representations that are invariant and representations that are equivariant.

b) (*4 points*) Explain how a "transforming autoencoder" works.

Total marks for both sections = 10x2 + 4x5 = 40
Total pages = 13