YOUR NAME:
YOUR STUDENT NUMBER:

UNIVERSITY OF TORONTO
Faculty of Arts and Science
APRIL/MAY 2012 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)

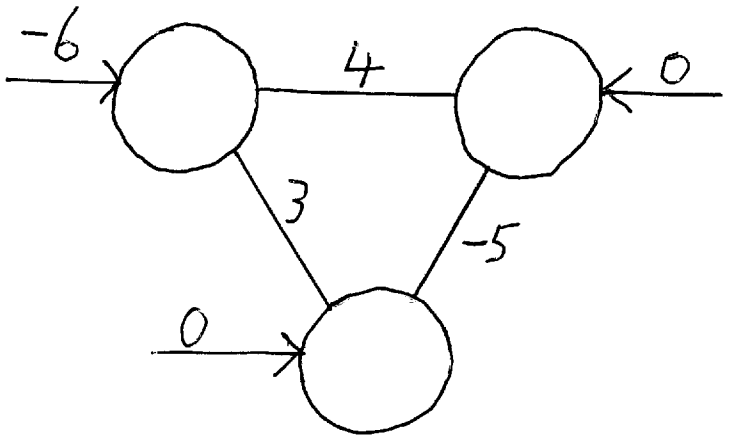
1. Suppose we have a single linear neuron with two inputs. We are trying to minimize a cost that is equal to the squared difference between the actual output of the neuron and the desired output, summed over all training cases. Assuming that the two weights on the inputs correspond to horizontal axes and the cost corresponds to height, draw a horizontal cross-section and a vertical cross-section of the error surface.

2. Explain very briefly the problem that arises if a network represents a quiet American by activating a unit representing "quiet" and a unit representing "American" at the same time as it is representing a noisy Canadian by activating a unit for "noisy" and a unit for "Canadian".

3. Explain why the backpropagation algorithm cannot be used with binary hidden units.

4. What is the relationship between Principal Components Analysis and the squared reconstruction error of an autoencoder?

5. In the Hopfield net shown below, the units have binary states of 1 or 0, and their biases are shown on the arrows. Fill in the **maximum** energy binary state.



6. In the K-means clustering algorithm, there is a quantity which decreases every time a data-point is assigned to a different cluster. The same quantity also decreases every time the center of a cluster is moved during the the refitting step. State what this quantity is as precisely as you can.

7. What is the **difference** between the maximum likelihood training procedure for a restricted Boltzmann machine and the contrastive divergence training procedure.

8. Briefly explain how a transforming autoencoder differs from a standard autoencoder.

9. Briefly explain the "bag of words" method of representing a document and say why it is frequently used in machine learning.

10. Suppose that a datavector is clamped on the visible units of a Boltzmann machine and we want to get an unbiased sample from the posterior distribution over all possible binary hidden vectors. Why is this easier to do in a Restricted Boltzmann Machine than in an unrestricted one?

PART B (20 points)

Answer EXACTLY 4 questions in this part.

1. Briefly describe three different ways of making the learning go faster when using back-propagation to learn a feedforward multilayer neural network.

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

a) (*3 points*) Explain in detail 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?

3. a) (*1 point*) When fitting a mixture of Gaussians model, it is possible to keep all of the variances extremely small and fixed and to only update the means and mixing proportions. If the variances are extremely small, what can you say about the posterior distribution over the Gaussians for a training case?

b) (*2 points*) When we are using the EM algorithm to fit a mixture of two Gaussians to data, what is the rule for updating the mixing proportions of the two Gaussians?

c) (*2 points*) Suppose that we start fitting a mixture of Gaussians by placing the means of the Gaussians at random locations. After one full step of the EM algorithm, what interesting property do all of the means have?

4. a) (*2 points*) In a mixture of experts model, the gating network computes the probability p_i^c of picking each expert, i, for a particular training case, c. When the mixture of experts model is being fitted to data, p_i^c will increase for some experts and decrease for other experts. What determines whether p_i^c increases or decreases?

b) (*1 point*) If we make the input vectors identical for all of the training cases (which seems like a very weird thing to do), the mixture of experts model becomes identical to another model you have learned about. What is this other model?

c) (*2 points*) Write down the two different objective functions that are used in the two different versions of mixtures of experts described in the paper by Jacobs et. al.

5. Support vector machines work by finding a "maximum margin" separating hyperplane in a high-dimensional feature space.

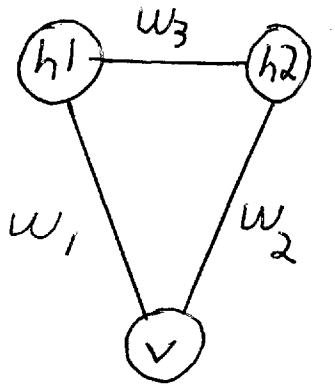
   a) (*2 points*) Why is the maximum margin hyperplane better than other separating hyperplanes?

   b) (*2 points*) Support vector machines use a special trick for computing scalar products efficiently when fitting a hyperplane in a high-dimensional feature space. What is this trick?

   c) (*1 point*) What is the relationship between the maximum margin hyperplane and a support vector?

6. 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 expressions for 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 $e^{w_1} = 3$, 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?

7. a) (*2 points*) Briefly explain how Restricted Boltzmann machines can be combined with backpropagation to learn a deep autoencoder.

b) (*2 points*) Briefly explain the document retrieval technique called "semantic hashing".

c) (*1 point*) Briefly explain how the learning of a deep autoencoder needs to be modified to learn the codes required for semantic hashing.

Total marks for both sections = 10x2 + 4x5 = 40
Total pages = 12