Department of Computer Science
Hong Kong Baptist University
COMP7650

Assignment 2 – 6 Marks
Due: 12/November. Submission on or before 18:30 (6:30pm) via the subjects' Web site.

Theme of assignment 2: Machine Learning in Data Mining and Knowledge Discovery.

Answer all of the following three questions.

Note:

- **This is an individual assignment. Plagiarism will result in having 0 marks for all students involved.**
- **Late submissions will be marked with 25% deducted for each late day. Only full days are counted. Fractions of a day will always be rounded up (i.e. a submission 2 hours late will count as a one day). Submissions 4 or more days late will not be marked.**

Question 1: Multi-Layer Perceptron **50%**

1. Separability.

- Load the training data (assignment2_train.txt) and only use the first two feature dimensions. Make a plot that allows to see which data point belongs to which class.
- Train a linear MLP (a single layer MLP) on the training data and report what weight values of w and bias b you obtain.
- Plot the decision boundary $y(x) = wx+b = 0$ on top of the plot of the data.

2. Long term dependency

- MLPs that have many hidden layers, recurrent and recursive MLPs can suffer from a problem known as the "long term dependency problem". Explain the reasons for this (in your own words), and propose one approach which would minimize the effects of the long term dependency problem. The recommended reading tnn-94-gradient.pdf, available of the subject web site will help you to answer this question.

3. Dimension reduction and projection

- Load the training data (assignment2_train.txt) then train an MLP-auto-associative memory with two hidden layer neurons in a single hidden layer. Plot the sum square error as it decreases with the iterations until convergence is observed.
- After training the MLP, plot for each input vector the output of the 2-dimensional hidden layer. Is this projection linearly separable? Explain.

Question 2: Self-Organizing Maps **30%**

1. Clustering

- Consider the training examples containing 16 vectors as is defined in Table 1. In the literature it is reported that a Self-Organizing Map of size 10 x 10 would cluster this set of data as is depicted in Figure 1. Verify such a finding by training your own SOM on this set of data. In practise, you will find that you are unable to produce a result as shown in Figure 1. Explain why. Describe how you chose the training parameters, and which set of training parameters helped to produce the best results. Explain your observations, and illustrate your findings. Explain how you assessed the quality of the result. For this experiment, you can use any software you like. For example, the som_pak (available from http://www.cis.hut.fi/research/som_pak/) may be the best documented and easiest one to use. Windows and Linux binaries as well as a source code is

available.

2. Unsupervised learning

- In unsupervised learning, there is no supervising signal, no teacher who tells us the meaning of a given set of data. List at least three different evaluation techniques which are suitable for measuring the performance of an unsupervised machine learning technique. Explain and compare the three evaluation measures.

Question 3: Complexity **20%**

- Assume you were to simultaneously train a 10x10 dimensional SOM and an MLP with 5 hidden layer neurons in each of two hidden layers. Assume that the SOM and MLP is trained on 13 dimensional input vectors for 3000 iterations. There are N input vectors. For the MLP, the targets are of dimension 5. Will the SOM or the MLP be quicker in completing the training iterations?
- Provide a discussion on the computational efficiency of SOM and MLP. How well does MLP and SOM scale with the size of a dataset?

Note: For the experiments, I have provided c-sources of an MLP, and a source code package for SOM. These can be downloaded from the subject's web site. However, the use of this code is by all means optional. You may use any software that you like or find on the Web, or write your own code.

Table 1: Animal names and their attributes

Animal		Dove	Hen	Duck	Goose	Owl	Hawk	Eagle	Fox	Dog	Wolf	Cat	Tiger	Lion	Horse	Zebra	Cow
is	small	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
	medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	big	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
has	2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4 legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	mane	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
likes to	hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
	run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
	fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

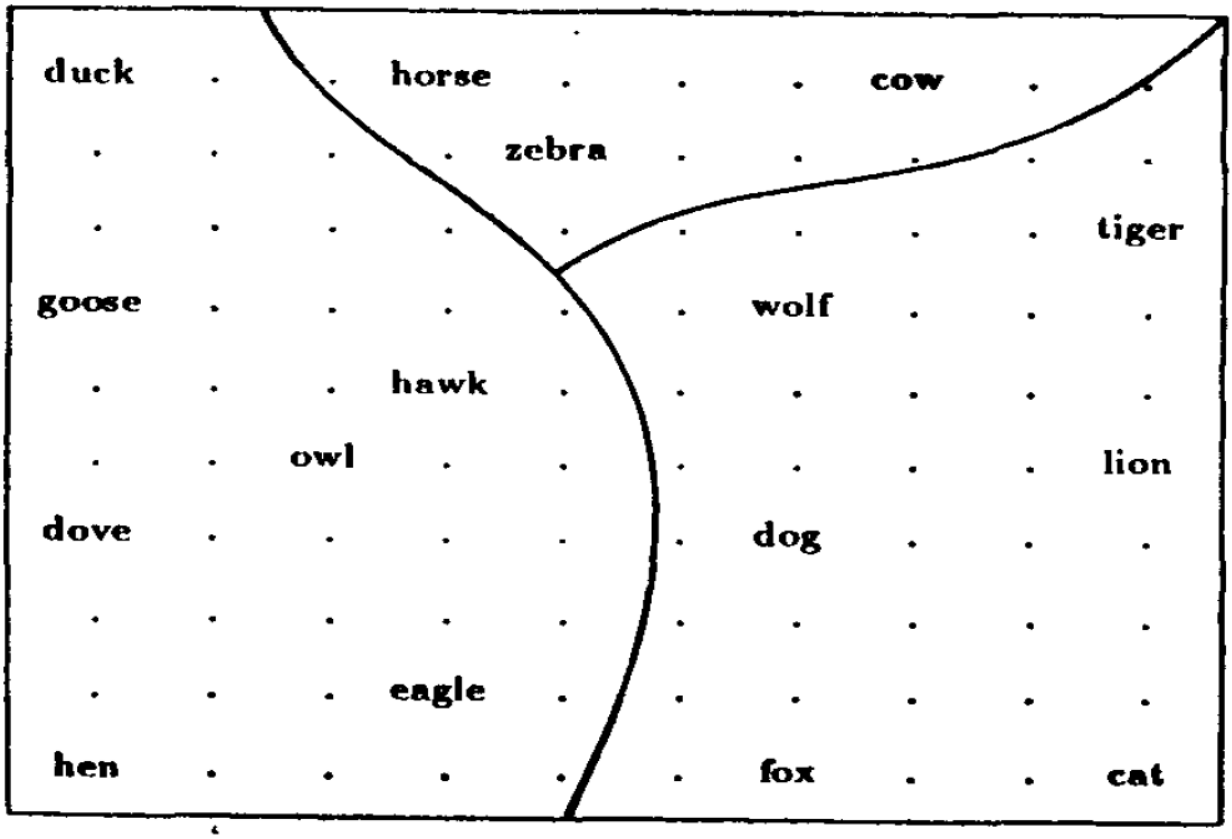

Figure 1: The mapping of the training data after the SOM network is trained on the dataset shown in Table 1. The network trained was of size 10x10, and used a rectangular neighboorhood.