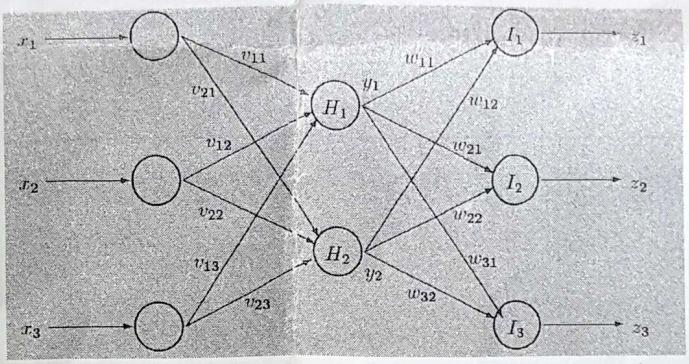
Department of Mathematics
II Semester 2016-2017
MTL 782 Data Mining
Major Examination Date: 2.5.17 Time: 3.30 PM – 5.30 PM
Weightage: 35%

Q1. A training pattern, consisting of an input vector $x = [x_1, x_2, x_3]^T$ and desired output $t = [t_1, t_2, t_3]^T$, is presented to the following neural network.



Suppose the values of different weight vectors are as follows:

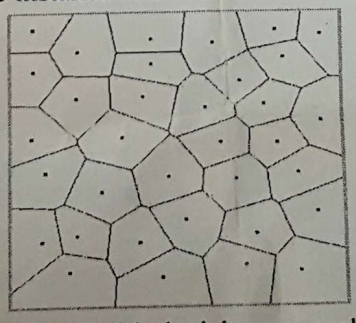
$v_1 = \begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \\ -2 \end{bmatrix}$, $v_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$, $w_1 = \begin{bmatrix} 1 \\ -3.5 \end{bmatrix}$, $w_2 = \begin{bmatrix} 0.5 \\ -1.2 \end{bmatrix}$ and $w_3 = \begin{bmatrix} 0.3 \\ 0.6 \end{bmatrix}$

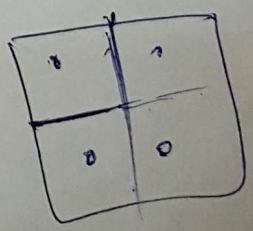
Assume that all hidden and output layer neurons have sigmoid activation functions given by $f(x) = \frac{1}{1+e^{-x}}$ and that each neuron has a bias $\mu = 0$ (zero).

a) If the network is tested with an input vector $x = [2, 3, 1]^T$ then what would be the output of the second output neuron z_2 ?

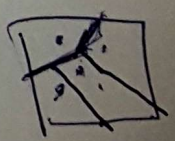
b) Suppose the learning rate of Back-propagation algorithm is 'l' and the loss function is Sum of Errors. Derive the update equation for the weight v_{11} in terms of other terms mathematically. You may define new terms if required. [1.5 + 3.5]

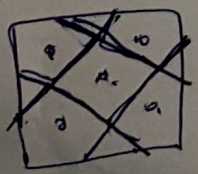
Q2. Let us try and classify data points in 2D Euclidean space. We are given n instances of such points P_1, P_2, …, P_n and the corresponding category for each point C_1, C_2, C_n [where C_1, C_2, .. C_n take values from the set of all possible class labels]. Under the k nearest neighbors classification scheme, each new element Q is simply categorized by a majority vote among its k nearest neighbors in instance space. The 1-NN is a simple variant of this which divides up the input space for classification purposes into a convex region (see figure below for the 1NN decision boundaries under the Euclidean distance measure), each corresponding to a point in the instance set.





a). Is it possible to build a decision tree (with decisions at each node of the form "is x > a", "is x < b", "is y > c", or "is y < d" for any real constants a,b,c,d) which classifies exactly according to the 1-NN scheme using the Euclidean distance measure ? If so, explain how. If not, explain why not.





b). Now assume that the distance measure is not explicitly specified to you. Instead, you are given a "black box" where you input a set of instances $P_1, P_2, .. P_n$ and a new example Q, and the black box outputs the nearest neighbor of Q, say Pi and its corresponding class label C_i. Is it possible to construct a k-NN classification algorithm based on this black box alone? If so, how and if not, why not?

c). If the black box returns the j nearest neighbors (and their corresponding class labels) instead of the single most nearest neighbor (assume $j \neq k$) , is it possible to construct a k-NN classification algorithm based on the black box ? If so how ? [1+1.5+1.5]

Q3. Consider the following algorithm of computing SNN similarity.

```
{ 1. find the k-nearest neighbors of all points.
2. If (two points x and y are not among the k-nearsest neighbors of each other )
similarity(x,y)=0;
  else
    similarity(x,y)=number of shared neighbors; }
```

The calculation of SNN distance does not take into account the position of shared neighbors in the two nearest neighbor lists. In other words, it might be desirable to give higher similarity to two points that share the same nearest neighbors in the same or roughly the same order.

(a) Describe how you might modify the definition of SNN similarity to give higher similarity to points whose shared neighbors are in roughly the same order. [3]

(b) Discuss the advantages and disadvantages of such a modification. [2]

Q4. (a) Write pseudo code for DBSCAN clustering algorithm.

(b) State 1 advantage and 1 disadvantage of DBSCAN clustering algorithm along with brief explanation.

(c) Suppose you are already given an optimal value for the 'Minimum Points' in DBSCAN clustering algorithm. Explain how it can be used to find an optimal value for the distance 'Eps' used in the DBSCAN clustering algorithm.

(d) Describe unsupervised cluster evaluation method which uses correlation and similarity matrix, as discussed in the class. How that method can be used for validating the clusters obtained? What is the intuition behind this evaluation measure? [1+1+1+2]

Q5. Define the recommendation problem. Describe two methods based on Collaborative Filtering for recommender system. Discuss their merits and demerits. [1+3+1]

Q6. The following table summarizes a data set with three attributes A, B, C and two class labels +. and -. Build a two-level decision tree.

A	B	C	Number of Instances	
			+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

(a) According to classification error rate, which attribute would be chosen as first splitting attribute? For each attribute, show the confusion matrix and the gains in classification error rate. [3]

(b) Repeat for the two children of the root node. [3]

(c) How many instances are misclassified by the resulting decision tree. [1]

Q7. What is feature selection? Can Decision tree be used for feature selection? Explain. [3]