CS 35000 – Introduction to AI
The University of Chicago, Winter 2005

(Please attempt all problems by yourself without consultation with classmates or friends. If any question is unclear, contact me (niyogi@cs) or the TA (matveeva@cs).)

Problem Set 2. Due: Thursday, 2/16

You will be given two data sets. Use the first one to test and train. Use the second one just to train and hand in the trained solution.

You can download the data sets from the CS35000 page.

Data Sets

(1) *Breast-cancer data.* The data is of the form (x, y) where x is a 9-dimensional quantity and y is either 2 (class 1) or 4 (class 2). Each line has 11 fields and denotes a single data point. The first field is a identity number of a patient which you must discard. Fields 2 through 10 are the first 9 components of the vector x. Field 11 is the label (either 2 or 4). Split this data into training and test sets for the problems. There are 699 data points in all. Use the first 400 to train and the last 299 to test.

(2) *Speech-data.* Each line of this file denotes data point x. The y labels have not been provided in the data set because the data are ordered by class. The first 133 data points are positive data ($y = +1$) and the rest are negative data. Use this only as training data.

1. MLPs and Kernels – Programming

Implement the Backpropagation Algorithm to train a Multilayer Perceptron with three layers. An input layer will take the input data. A hidden layer will transform the data. An output layer will compute the final outputs.

The program should take as input:

1. the number of input nodes
2. the number of hidden nodes
3. the number of output nodes
4. the data set of (x, y) pairs

It should output:

1. the weights after backpropagation training is done.
2. the performance on the test set.
3. the training error as a function of the number of iterations of gradient descent.
4. the training error as a function of different choices of step size.

2. Theoretical

2. Consider the combinatorics of exhaustive inspection of clusters on n samples into c clusters.

(a) Show that there are exactly

$$\frac{1}{c!}\sum_{i=1}^{c} C_i^c(-1)^{(c-i)} i^n$$

such distinct clusterings.

(b) How many clusters are there for $n = 100$ and $c = 5$?

(c) Find an approximation to your answer for (a) for the case $n >> c$. Use your answer to estimate the number of clusterings of 1000 points into 10 clusters.

3. Suppose you were given n examples in a non metric space such that only a similarity measure $s(x_i, x_j)$ (as opposed to a distance measure) existed between the various tokens.

(a) What do you think are reasonable properties a similarity measure should have?

(b) Suppose you were to get k clusters, what clustering scheme would you use that is appropriate for the similarity measure developed in (a)?

(c) Can you think of clustering problems in practical applications where a natural distance measure would be hard to define but a similarity measure would be quite natural?

4. Consider the following production rules associated with a stochastic context free grammar. The objects $s, np, pp, vp, noun, verb, prep$ are all non-terminals while lexical items "like","swat", etc. are all terminals. The probabilities associated with each production rule are shown in brackets.

1. $s \rightarrow np\ \ vp(0.8)$
2. $s \rightarrow vp(0.2)$
3. $np \rightarrow noun(0.4)$
4. $np \rightarrow noun\ \ pp(0.4)$
5. $np \rightarrow noun\ \ np(0.2)$
6. $vp \rightarrow verb(0.3)$
7. $vp \rightarrow verb\ \ np(0.3)$
8. $vp \rightarrow verb\ \ pp(0.2)$
9. $vp \rightarrow verb\ \ np\ \ pp(0.2)$
10. $pp \rightarrow prep\ \ np(1.0)$
11. $prep \rightarrow like(1.0)$
12. $verb \rightarrow swat(0.2)$
13. $verb \rightarrow like(0.4)$
14. $verb \rightarrow flies(0.4)$
15. $noun \rightarrow swat(0.05)$
16. $noun \rightarrow flies(0.45)$

17. $noun \rightarrow ants(0.5)$

Consider the sentence “swat flies like ants”. How many valid derivations (parses) can you come up with for this sentence? What are the probabilities associated with each parse? Do these different derivations have different meanings?