Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Subsymbolic Processing and Neural Networks

Monday 14th January 2008
Time: 14:00 – 16:00

Please answer any THREE Questions from the FOUR questions provided

This is a CLOSED book examination

The use of electronic calculators is permitted provided
they are not programmable and do not store text
1. A client wants to predict whether customers will purchase their product, based on a questionnaire consisting of 45 yes/no questions the client obtained in a market survey. The client has collected 200 examples of customers who have made the purchase and 200 examples of customers who have not. The client has heard about neural networks and has hired you to produce a neural network to predict whether a potential customer will purchase the product. They ask whether it is possible to obtain an error rate of 5% or less.

   a) Describe in detail the procedures you would apply to find the best performing MLP for this problem. Describe the architectures you would use and how you would determine which was best. (8 marks)

   b) How would you estimate how well the system will perform in use? How would you evaluate the expected accuracy of this estimate? (4 marks)

   c) Suppose the answer found in part b) is worse than the 5% error desired by the client. Give possible reasons that the performance is less than ideal. (4 marks)

   d) Suppose instead of 200 examples of each class, there were 40 examples of each class. Would the task be feasible with this amount of data? Why or why not? (4 marks)

2. A new medicine is introduced to treat a serious disease. This medicine is found to be effective only in some types of patients; in others it fails to treat the disease and causes unpleasant side effects. Thus, it would be very useful to be able to predict in advance those patients most likely to respond positively to this medicine, so that those unlikely to experience positive results could be given an informed choice whether to take the medicine. Your task is to build a system to predict how likely it is that the medicine will work for a given patient.

   You have records from 1000 patients who have been given this medicine, 40% of whom have responded positively and the rest have not. There are 80 attributes describing each patient: 60 attributes are binary; the remaining 20 are numbers which represent counts of appearances of symptoms.

   a) Consider a Bayesian classifier for this problem. Describe the probability model to be used, identify which probabilities need to be estimated from data, and show how Bayes rule is used to classify new patients. (6 marks)

   b) How would you estimate the probabilities needed for the model described in part a)? (6 marks)

   c) The goals are to give the medicine to as many people who it will help as possible, while at the same time minimizing the number receiving who will not be helped. How would you evaluate this system in such a way as to help balance these two competing goals? (4 marks)

   d) The financial cost of this medicine is cheap compared to the alternative treatment. However, if it fails, the alternative treatment must be applied as well. Assume that your system is effective at estimating the probability that the treatment will work. Give a formula for the threshold at which the new medicine becomes cost effective to apply. Assume that the new treatment costs $E_n$ and the old treatment costs $E_o$ per patient. (If you wish, you can put specific numbers in for $E_n$ and $E_o$.) (4 marks)
3. A set of files needs to be backed up to a set of disks. Each file must be stored entirely on a disk and all files must be backed up. You can buy as many disks as you need, and the goal is to minimize the total cost of the backup disks. Backup disks come in several capacities, and cost per MB is lower as the disk capacity increases. You are asked to consider several different ways to solve this problem. To make this more concrete, consider that there are N files to be backed up; their sizes are denoted $F_i$ for each value of $i$. There are three sizes of disk, 1GB, 2 GB and 4 GB and their costs are $C_1$, $C_2$, and $C_4$ respectively. You wish to find the configuration of files to disks that minimizes the total cost of backup disks.

a) Consider the following three methods for searching for solutions to this problem: exhaustive search, gradient descent and hillclimbing algorithms. For each, state whether the approach can be applied at all, whether the approach is feasible and whether the approach will always find the best solution. (6 marks)

b) Devise a greedy algorithm for this problem. Describe it in pseudo-code if you can. (6 marks)

c) Describe an evolutionary algorithm for this problem. Be sure to describe the representation and the fitness function you would use. (8 marks)
4. You are developing a two-class classifier. One of the attributes is age. You believe that age is a positive indicator for class 1 if it is low and if it is high, but intermediate values are indicative of class 2. In other words, you think the situation looks roughly like this.

a) Could a perceptron using age as a numerical input perform this task? Why or why not. (4 marks)

b) Could a naive Bayes classifier using age as a numerical input perform this task? Why or why not. (4 marks)

c) How would introducing a latent variable help make Bayesian classification more feasible? How does the latent variable change the way the probabilities are estimated? (6 marks)

d) One of the problems in training an MLP is to select the architecture. Choose either a genetic algorithm or a hill-climbing algorithm and describe how it could be used to search over MLP architectures to find the one that performs best. Describe the fitness function in detail and how the algorithm will develop new architectures and evaluate them. (6 marks)