Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Subsymbolic Processing and Neural Networks

Tuesday 20\textsuperscript{th} January 2009

Time: 14:00 – 16:00

Please answer any THREE Questions from the FOUR questions provided

The use of electronic calculators is permitted provided they are not programmable and do not store text
1. There is a disease for which some patients respond to treatment A, some respond to treatment B and some don't respond to either treatment and are best left untreated. Due to the expense and side-effects of the treatments, it is desirable to be able to predict this in advance for each patient, with an error rate less than 4% for each class. A questionnaire has been developed consisting of 45 yes/no questions. Your task is to make a system which predicts which treatment (if any) to give to a patient from the questionnaire. You have 300 example questionnaires: 100 from each of the categories: responded to treatment A, responded to treatment B, did not respond to either.

a) Describe how neural networks could be used in this task. (4 marks)

b) Describe in detail: what neural network architectures you would use and how you would go about finding the best performing neural network for this task. (10 marks)

c) Suppose your results do not achieve the desired level of accuracy. Give reasons for this and what could be done to try to rectify it. (3 marks)

d) Suppose your system gives a test error rate of 3% on each class. How confident could you be that the system is below the desired 4% true error rate? Suppose it is required that this level of confidence is greater than 95%. How could you achieve this? (3 marks)
2. A company wishes to predict whether customers will purchase a new product based on products they have purchased in the past. They have a database of 10,000 customers. Of those, they have test-marketed the new product to 200 customers, and found 50 bought the new product and the rest declined to buy. For all customers, there are 100 products (not counting the new product) and the database tells whether they were purchased or not. For customers in the test market, it also tells whether they purchased the new product.

Your task is to make a system to predict the probability that a customer in the database will purchase the new product based on their past purchases.

a) Explain why estimating a full probability model describing the joint probability of all purchases from the database is not feasible without making some assumption. (3 marks)

b) Describe how to make a system to predict the probability that a customer in the database will purchase the new product, based on what purchases they made in the past. (8 marks)

c) For every customer who buys the product, the company makes a net profit of £0.70. For every customer the company markets to who does not buy the product, the company loses £0.30. Given a customer from the database, how should the decision be made whether to market the new product to them? (4 marks)

d) It is believed that the customer base actually falls into two classes: young professionals and retirees, and these have a different probability of purchasing the new product. Unfortunately, there is no information in the database concerning which group each customer is in, and no chance of finding this out. What method could be used to make the probability model in this case? In particular, what algorithm is used and what new variables need to be introduced? (5 marks)
3. Your task is to assign buses to a set of bus routes for a town. Already determined is the set of trips (i.e. the set of bus routes and their schedules). Your task is to determine which physical buses should make each trip such that the minimum number of buses is used. There are \( N \) trips in all. You are given an \( N \times N \) conflict table which tells which trips can share the same bus. If the \( i,j \)th entry of the table contains 1, trip \( i \) and \( j \) must use different buses (because the two trips overlap in time or there is insufficient time to travel between them). If the \( i,j \)th entry of the table contains 0, the two trips can use the same bus. The task is to determine which buses should run on each route such that the minimum number of buses is used. Assume that \( N \) is around 100.

   a) Discuss the feasibility of finding an algorithm which is guaranteed to solve the problem. You don’t need to give an algorithm. (3 marks)

   b) Devise a greedy algorithm to solve this problem. (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)

   d) Suppose in addition to minimising the number of buses, the bus company also wants to minimize the distance travelled by the buses. Fewer buses could mean greater distances travelled for buses used in multiple routes. What method could be used to search for solutions which best trade-off between minimising number of buses used and distance travelled? (3 marks)

4. a) In the lab we used naïve Bayes classifier to make a spam filter. How was the naïve assumption used and why was it used. (4 marks)

   b) Consider other ways of making classifiers studied in this course. What advantages (if any) and disadvantages (if any) would these have over naïve Bayes classification on the spam filtering problem? (6 marks)

   c) One of the issues in making a spam filter is what to use as features. Describe in detail how a hill-climbing algorithm could be used to search for the most effective features. (6 marks)

   d) One approach to go beyond naïve Bayes is to use a mixture model for each class. Describe how the training error and test error would depend on the number of components in the mixture. How would the best number of components be found? (4 marks)

END OF EXAMINATION