Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Machine Learning

Tuesday 20\textsuperscript{th} January 2009
09:45 – 11:45

Answer any THREE questions from the FOUR questions provided

Electronic calculators may be used, provided that they cannot store text.
Answer any THREE questions from the questions provided

1. a) Give the equation for the Perceptron, and explain its parameters as far as possible, including what effect they have on the decision boundary. You do not need to describe the learning algorithm. (3 marks)

b) Describe the general principle of a Support Vector Machine, including how it differs from a Perceptron classifier. Be sure to give a definition of (i) a support vector, (ii) the margin, (iii) a kernel. (6 marks)

c) Give 1 advantage and 1 disadvantage of Support Vector Machines. (2 marks)

d) The SVM optimisation problem is phrased as trade-off between two objectives:

\[ L = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i \{ t_i f(x_i) - 1 \} \]

i) What is the purpose of the \( \frac{1}{2} w^T w \) part? (2 marks)

ii) What is the purpose of the \( \sum_{i=1}^{N} \alpha_i \{ t_i f(x_i) - 1 \} \) part? (2 marks)

2. a) Give a definition of “overfitting”, including how it can be detected, and how it can occur in (i) decision tree classifiers, (ii) k-nearest neighbour classifiers. (4 marks)

b) Outline the differences between “filter” and “wrapper” feature selection, including any advantages/disadvantages of each. (4 marks)

c) The World Bank is analysing financial trades over the last year to determine what caused the current financial crisis. It has compiled a dataset of 500 banks and financial institutions worldwide, and asked the Chief Financial Officer of each bank 10 yes/no questions on his management style, and 10 questions about the amounts of their current investments (so these answers were continuous variables). In addition they have records of whether each bank is currently judged as being “in crisis” or “not in crisis”.

The bank asks you for advice on using decision trees to assess the situations of banks in the future. Give a technical description of the ID3 algorithm, such that a programmer within the bank could apply it to the data above. (7 marks)
3. a) Explain the difference between a supervised and unsupervised approach to machine learning. Give one example of a task that can be addressed by each approach. (3 marks)

b) Below is the probability density function for a mixture model with K components. Each component is associated with a set of parameters denoted by \( w_j \),

\[
p(x) = \sum_{j=1}^{K} p(j) p(x \mid w_j)
\]

i) In addition to \( w_j \), what are the other parameters of the mixture model? (1 mark)

ii) The Expectation-Maximization (EM) algorithm is an iterative algorithm that can be used to learn the parameters of a mixture model from data. Describe the two steps that characterize each iteration of the algorithm (you can either use general terms or consider one specific class of mixture model in your answer). (4 marks)

iii) Which quantity is the EM-algorithm guaranteed to not decrease at each iteration? Give the definition of this quantity. (2 marks)

iv) You use a Gaussian mixture model in order to find clusters in a multivariate data set. The software used to implement the mixture model has an option to select the class of covariance matrix as isotropic, diagonal or full-rank. Explain how this option affects the model. How would you determine the best choice for this option? (5 marks)
4. Below is a hidden Markov model (HMM) of DNA sequence data.

![Diagram of HMM]

a) Explain what the numbers in the diagram represent. (2 marks)

b) The numbers associated with the arrows leaving a state sum to one. Explain why that must be the case for the diagram to describe a valid HMM? (1 mark)

c) What is the probability of generating a sequence which passes through state 1 twice, generating a symbol C each time, and then terminates? (2 marks)

d) What is the probability of generating the sequence “CC” from the model allowing for all possible paths through the model? (3 marks)

e) Given that a sequence “CC” has been generated by the model, what is the probability that both symbols were generated by state 1. (Hint: use the results from the previous two sections of the question) (2 marks)

f) The above HMM is used to locate regions of a long DNA sequence with an unusually high ratio of A to C composition.

i) Which algorithm could be used to carry out this task?

ii) What does this algorithm compute?

iii) What is the computational complexity of this algorithm for this model?

iv) A naïve implementation of the algorithm is observed to return incorrect results due to a numerical underflow problem. Explain how this problem can be resolved by improving the implementation. (5 marks)