Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Machine Learning and Data Mining

Date: Friday 21st January 2011
Time: 14:00 - 16:00

Please answer ALL Questions
For full marks your answers should be concise as well as accurate.
Marks will be awarded for reasoning and method as well as being correct.

This is a CLOSED book examination
The use of electronic calculators is permitted provided they are not programmable and do not store text.

[PTO]
1. a) When assessing a machine learning procedure, the generalisation error is obviously very important. State 3 (three) additional factors that might be used to assess the performance of a learning procedure.

(3 marks)

b) How can we control overfitting in decision trees?

(2 marks)

c) State the equation for the entropy of a feature, and calculate it for a binary feature with \( p(X=1) = 0.75 \).

(3 marks)

d) Describe the id3 decision tree algorithm in as much detail as you can, including pseudo-code if necessary.

(7 marks)
2. 

a) Rebecca is building a spam filter. She notes from some historical data in her inbox that of 500 spam emails, 400 of them contain the subject line “Want a job?”. She therefore takes this as a good indicator of an email being spam. However, she also notes that of the 500 non-spam emails, 25 contain the same phrase, as she has recently been applying for a new job.

Rebecca observes that a new mail has just arrived, with the subject line “Want a job?”.
Using Bayes rule, show how to calculate the probability that this new mail is spam.

(b) A doctor can run a test for the horrible disease Examophobia - the test has three possible outcomes – “positive”, “negative”, and “unknown”. It is known that among all students, if Examophobia is present, the test comes out positive 80% of the time, negative 15% of the time, and unknown 5% of the time. If Examophobia is not present, the test comes out positive 1% of the time, negative 70%, and unknown 29%.

Among the general student population, Examophobia is known to occur in 30% of all students.

A student enters the clinic and tests positive for the disease. What is the probability they really have Examophobia?

b) A doctor can run a test for the horrible disease Examophobia - the test has three possible outcomes – “positive”, “negative”, and “unknown”. It is known that among all students, if Examophobia is present, the test comes out positive 80% of the time, negative 15% of the time, and unknown 5% of the time. If Examophobia is not present, the test comes out positive 1% of the time, negative 70%, and unknown 29%.

Among the general student population, Examophobia is known to occur in 30% of all students.

A student enters the clinic and tests positive for the disease. What is the probability they really have Examophobia?

(6 marks)

b) A doctor can run a test for the horrible disease Examophobia - the test has three possible outcomes – “positive”, “negative”, and “unknown”. It is known that among all students, if Examophobia is present, the test comes out positive 80% of the time, negative 15% of the time, and unknown 5% of the time. If Examophobia is not present, the test comes out positive 1% of the time, negative 70%, and unknown 29%.

Among the general student population, Examophobia is known to occur in 30% of all students.

A student enters the clinic and tests positive for the disease. What is the probability they really have Examophobia?

(6 marks)

c) What is the Naïve Bayes assumption?

(3 marks)
3. a) State the equations for “sensitivity”, and “specificity” explaining the concepts and why they are important.

   (4 marks)

b) State and describe the model for a linear classifier.

   (3 marks)

c) Write pseudo-code of the learning algorithm for a linear classifier, including descriptions of any parameters that have to be set by hand rather than learnt.

   (6 marks)

d) The Support Vector Machine optimisation problem is phrased as tradeoff between two objectives:

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

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

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

   (2 marks)