Two hours

QUESTION PAPER MUST NOT BE REMOVED FROM THE EXAM ROOM

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Machine Learning and Data Mining

Date: Monday 14th January 2013
Time: 14:00 - 16:00

Please answer ALL Questions

Section A should be answered in an Answer book.

Section B is multiple choice and should be answered directly on the exam paper.
Only answers written in the boxes on the exam paper will be marked.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text.
Section A

Answer all questions.

1. The accuracy and error rate, on both the training and testing datasets, is the most commonly used way of evaluating a learning procedure. Other than these, state two (2) additional factors that could be used to assess the performance of a learning procedure. (2 marks)

2. What is the main difference between the decision boundary induced by a Perceptron, and the decision boundary induced by a linear SVM? (2 marks)

3. State the formula for the entropy. Using whatever log base you wish, calculate the entropy of the feature \( x = \{0, 1, 0, 0, 1, 0\} \). (2 marks)

4. Explain the difference between filters and wrappers for feature selection. (2 marks)

5. Explain the method of cross validation, and what it is used for. (2 marks)

6. The SVM error function is:

\[
E = \sum_{i=1}^{N} \max \left\{ 0, 1 - y_i f(x_i) \right\} + \frac{1}{2} \sum_{j=1}^{d} w_j^2.
\]

Define the terms of this expression, i.e. state what is meant by \( w_j, f(x_i), \) and \( y_i \). State also what is achieved by minimising the first part versus the second part (i.e. before and after the ‘+’ symbol). (5 marks)
7. David travels to the office by one of three modes of transportation: car, bus, or train. If he decides to go by car, due to the high traffic in Manchester, there is a 50% chance he will be late. If he goes by bus, which has special reserved lanes (but is overcrowded 80% of the time, meaning David would have to stand) then the chance of being late is only 25%. The train is almost never late, being punctual 95% of the time, but it costs twice as much as the bus.

David is late to work one day, and his boss wishes to estimate the probability that he drove to work that day by car. He knows that David doesn’t like to drive, so guesses that there is a only a 10% chance of David taking the car. Since he knows that David is quite frugal with money, he guesses that it is twice as likely for David to take the bus as he is to take the train.

What is the boss’s estimate of the chances that David came to work by car? (5 marks)

8. Given the training data in the table below, what probability will a Naive Bayes classifier predict for $p(y = 1|x)$, given input $x = \{1, 1, 0\}$? (5 marks)

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Section B is restricted and cannot be published