Two hours

QUESTION PAPER MUST NOT BE REMOVED FROM THE EXAM ROOM

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Machine Learning and Optimisation

Date:	Monday 24th January 2011
Time:	14:00 - 16:00

Answer ALL 10 short questions in Section A
Answer ONE question from Section B
Answer ONE question from Section C

Use SEPARATE Answerbooks for EACH section

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.

Section A (Answer ALL questions)

Section A is a multiple choice section and is therefore restricted

Section B (Answer ONE question only from this section)

B1. 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.`

(5 marks)

d) Describe the id3 decision tree algorithm in full, including pseudo-code, and details of how the decision to split a node is taken.

(10 marks)

[PTO]

B2. a) Why do we need confidence intervals when plotting the performance of learning algorithms?

(1 marks)

b) Explain the terms “cross-validation” and “bootstrap”, in the sense of data sampling schemes.

(4 marks)

c) State the learning rule for perceptrons, describing how and why it works, including an example update.

(5 marks)

d) Read the following passages, then answer the question following.

Rebecca is teaching an industrial training session all about Machine Learning algorithms. She says:

“K-nearest neighbour classifiers are a great example of a linear classifier, though they can in fact solve non-linearly separable problems. You should be careful however not to set the K value too high though, as it will likely not converge in a finite number of iterations. The reason is that with large K and lots of datapoints, it is a very computationally intensive algorithm, but this can be offset with some careful feature extraction. A great feature extraction algorithm is the method of ROC analysis, which is essential when there are highly imbalanced classification costs. The main measurement in ROC is called ‘true positives’, which measure generalisation accuracy very efficiently. We can use true positives to calculate sensitivity, and true negatives to calculate specificity. A perceptron could have been used instead, though of course these can only solve linearly separable problems – the nice thing is of course that the convergence theorem tells us that it will find a linear separating decision boundary if it exists – I could of course extend this by applying a linear SVM, which is almost certainly more reliable than a Perceptron, since it will guarantee a larger margin.”

She goes on to describe how she worked applying k-nn classifiers to a dataset recently.

“I had a dataset of 10,000 examples, 9,000 of which were class 1, and 1000 class 0. This meant I was able to allocate a large K value without too much trouble, though I did watch out to ensure overfitting did not happen. I noticed that with a K value of 7,000, I could classify the dataset with 90% accuracy, so I finished my investigations – I was never quite sure why this was the case though.”

List 5 things wrong with Rebecca’s understanding of ML algorithms.
Explain why she was able to get 90% accuracy, mentioned in the second paragraph.

(10 marks)

Section C (Answer ONE question only from this section)

C1. *K nearest neighbours* (KNN), *naïve Bayes* (NB) and *support vector machine* (SVM) are three popular machine learning algorithms for classification.

(a) Given a training data set of *N* examples, $\{(\boldsymbol{x}_i, y_i) | i=1,\ldots,N\}$, for a classification task, describe how the KNN, NB and SVM classifiers are constructed.
(8 marks)

(b) For a test data set on the same classification task mentioned in (a), we found that KNN and NB outperform SVM. Explain why this situation can happen and describe how you can improve the performance of SVM for this task. **(4 marks)**

(c) During the module coursework we always had a fixed dataset, supplied before learning begins. An alternative form of learning is "online" learning, where examples arrive one-by-one in a continual stream, and you are required to provide a prediction for the example arriving at each step before the true label is revealed to you. The online learning scenario is challenging as each example can be used only once – i.e. you have a queue of length 1.

Given your knowledge of K-nn, SVM, and Naive Bayes, which of these algorithms are simple to adapt for the online learning scenario, and which are difficult to adapt? Give full descriptions of the alterations you might make**.** **(8 marks)**

[PTO]

C2. Clustering analysis is an unsupervised learning process that groups a set of physical or abstract objects into clusters of similar or coherent objects. There are different methodologies for clustering.

(a) Briefly describe the following clustering methodologies: partitioning clustering and hierarchical clustering. Give an example for each methodology**.** **(6 marks)**

(b) K-means is a popular clustering algorithm. Describe this algorithm in detail and list its advantage and two disadvantages**.** **(6 marks)**

(c) Suppose that BT is going to allocate a certain number of automatic teller machines (ATMs) in the suburb of south Manchester region. Households or places of work may be clustered so that typically one ATM is assigned per cluster. You are asked to use what you have learned to help them make an allocation plan. Describe your method that deploys ATMs so that each of them must cover at least a number of household or working places, e.g. 1,000. It is essential to justify your method and address issues to be considered for satisfactory allocation. **(8 marks)**

END OF EXAMINATION