Two hours

QUESTION PAPER MUST NOT BE REMOVED FROM
THE EXAM ROOM AND MUST BE RETURNED

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Machine Learning and Optimisation

Date:	Wednesday 15th January 2014
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 a SEPARATE answerbook for each SECTION.

This is a CLOSED book examination

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

[PTO]

Section A of this examination is restricted

Section B

Answer *one* question only from this section.

1. a) *Sensitivity* and *Specificity*, are two measures used in ROC analysis. The first step in calculating them is to build a *confusion matrix*, as below, where y means the true label, and $\hat{y}$ means the *prediction* of the label.

	$y = 0$	$y = 1$
$\hat{y} = 0$	165	34
$\hat{y} = 1$	12	96

Given this table, *how many examples* did we have in total?
What is the value of the *sensitivity*, and the *specificity*? (5 marks)

b) State the learning rule for the perceptron algorithm, giving definitions for all the mathematical terms used, and stating which term(s) can be altered by a user to control the learning. (5 marks)

c) Given the following text, answer the question below.

> *Charles is a singing teacher. Sometimes his students sing too much, damage their vocal chords, and should really take a rest for a few weeks. This medical condition is quite difficult to diagnose, unless you are an expert like Charles, who has many years' experience. Charles would like to apply Machine Learning to the task.*
>
> *He has a device which measures several vocal characteristics, converting them to digital form, when someone speaks. He records voice segments from his students: 90 who have no problems, and the minority 10 who do have vocal damage. The* 20 *features he records are all continuous valued, in the range* $[0,1]$*, apart from one which is in the range* $[0,50]$*. To correct for this, he rescales that last feature to be the same as the others.*
>
> *He then uses feature selection, measuring the correlation of each feature to the class label. He discards features that have a close to zero correlation to the label, as this is likely to cause overfitting. This leaves him with 15 features.*
>
> *Since he discarded some features, he makes sure to use all 100 students to train a Perceptron learning algorithm, using a learning rate of* $\alpha = 0.25$*. He tunes this learning rate, slowly decreasing it to* 0.1*, as this will take smaller steps in the search space. Accordingly, he increases the number of iterations to compensate for this.*
>
> *Because of the Perceptron convergence theorem, he knows that it should be able to correctly classify at least* $0.1 \times 100 = 10$ *students. In fact it vastly exceeds his expectations, getting an accuracy of* 90%*. Charles happily concludes that the data must be linearly separable, but would still like better performance, as he thinks it may be overfitting.*
>
> *Due to his conclusion on the Perceptron, he decides to apply a linear classifier, so picks the K-NN model. He sets* $k = 100$*, and finds that he can again only correctly classify 90 examples, but is not sure why. Charles remembers that the perceptron has has lower computational complexity at testing phase than the K-NN, so on balance recommends the perceptron as the model to use in practice.*

State 5 things that are wrong with Charles's methodology and understanding of Machine Learning, giving reasons for each, and explain to him why his K-NN rule behaved in the way it did.

(10 marks)

2. a) Write out full pseudo-code for the ID3 algorithm, being sure to state base cases, and state *precisely* (i.e. with a mathematical equation) how you would determine the most important feature at each splitpoint. (6 marks)

   b) What is the main difference between a *filter* method, and a *wrapper* method for feature selection? Give pseudo-code for the method of *forward selection* with a wrapper method. (4 marks)

   c) Describe the principle of an 'ensemble' learning algorithm. State 2 examples of such algorithms, being sure to (i) give full pseudocode, such that someone could implement the algorithm, and (ii) state the differences between them. (10 marks)

Section C

Answer *one* question only from this section.

3. Supervised learning is one of the most important machine learning paradigms and there are a variety of learning algorithms for different situations.

   a) To train a linear classifier, *Support Vector Machine* (SVM) is a modern learning algorithm while Perceptron is a traditional one. Describe the following important concepts used in SVM: (i) support vector and (ii) margin. For a given linearly separable classification task, explain why for a given training data set, the SVM learning always ends up to the unique solution regardless of initialization but the Perceptron learning cannot, and why the generalisation performance of the SVM is always not worse than the Perceptron. (10 marks)

   b) *Incremental* learning is one of various supervised learning forms where a trained learning model can be updated by using a new example only without the need of re-training a learning model by the whole training data set previously used. Analyse four learning algorithms learnt from this module, i.e., *Decision Tree*, the *Perceptron*, *Naïve Bayes* and the *SVM*, to decide if they can be easily used in incremental learning. It is essential to provide details of your analysis to justify your answer. (10 marks)

4. Clustering analysis is an unsupervised learning process. Clustering analysis algorithms share some common issues with those used in supervised learning.

   a) A class of partitioning clustering algorithms are derived by minimising a cost function as follows:

$$J = \sum_{k=1}^{K} \sum_{\mathbf{x} \in D_k} d^2(\mathbf{x}, \mathbf{m}_k), \tag{1}$$

   where $d(\cdot,\cdot)$ is the Euclidean distance, D_k is the set of all data points belonging to cluster k, and $\mathbf{m}_k$ is the centroid of cluster k.
   Given a data set of four points, $\mathbf{x}_1 = (0\ 2)$, $\mathbf{x}_2 = (2\ 0)$, $\mathbf{x}_3 = (4\ 0)$ and $\mathbf{x}_4 = (0\ 6)$, there are three candidate partitions of two clusters: (i) $D_1 = \{\mathbf{x}_1, \mathbf{x}_2\}, D_2 = \{\mathbf{x}_3, \mathbf{x}_4\}$, (ii) $D_1 = \{\mathbf{x}_1, \mathbf{x}_3\}, D_2 = \{\mathbf{x}_2, \mathbf{x}_4\}$ and (iii) $D_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}, D_2 = \{\mathbf{x}_4\}$.

   Calculate the value of J for each of these three partitions, and state which is the best in terms of the cost function.

   (6 marks)

   b) The *K-Means* algorithm provides a heuristic solution to finding a partition corresponding to a minimum of the cost function J. Describe main steps of the algorithm. Based on the cost function J, point out one of its limitations in clustering analysis. (6 marks)

   c) *K-Means* and *K Nearest Neighbours* (*K-NN*) are two simple yet popular machine learning algorithms. However, both have to deal with an issue, e.g., distance metric, whenever are applied to a real problem. Suppose there are two training data sets, i.e., $X_1 = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ and $X_2 = \{(\mathbf{z}_1, y_1), \cdots, (\mathbf{z}_N, y_N)\}$ where y_n is the label of $\mathbf{z}_n$ $(n = 1, \cdots, N)$, for two different learning tasks. Explain which algorithm, K-Means or K-NN, should be applied to X_1 and X_2, respectively, for learning. For two candidate distance metrics, $d_1(\cdot,\cdot)$ and $d_2(\cdot,\cdot)$, describe a procedure to decide which one is a better distance metric used in K-means and K-NN algorithms, respectively, in terms of the aforementioned training data sets. (8 marks)

END OF EXAMINATION