Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Machine Learning and Optimisation

Date:	Friday 18th January 2013
Time:	09:45 - 11:45

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.

Section A

Answer *all* questions. Each question is worth one (1) mark.

1. Overfitting is the theoretical ideal for any learning algorithm. True or False? (1 mark)

2. If you have a weight vector $\mathbf{w} = \{0.2, 0.3, 0.1, -0.7\}$, and threshold $t = 1.5$, what label would the perceptron model predict for $\mathbf{x} = \{1, 2, 3, 4\}$? (1 mark)

3. I have a training dataset of 150 examples: 50 have label $y = 1$, 100 have $y = 0$. I use this for a K-NN model with $k = 150$. I receive a new testing data point and test the model. What does it predict: $y = 0$, or $y = 1$? You **must** explain why it predicts this instead of other label. (1 mark)

4. Briefly explain the method of ‘cross-validation’. (1 mark)

5. A linear SVM is guaranteed to always outperform the K-NN rule in terms of generalisation error due to the kernel trick. True or False? You must **give a reason why**. (1 mark)

6. Explain the essential difference between supervised and unsupervised learning. (1 mark)

7. Explain the essential difference between Bayesian and naïve Bayesian classifiers. (1 mark)

8. Briefly describe the two ultimate goals of clustering analysis. (1 mark)

9. Given two objects a = (9, 13, 1, 3) and b = (8, 12, 0, 2), calculate their Manhattan distance. (1 mark)

10. Given two clusters $\{0, -1\}$ and $\{2, 3\}$ of one dimensional objects, use the single link with Manhattan distance to calculate their distance. (1 mark)

Section B

Answer *one* question only from this section.

1. a) Give pseudo-code for building a decision tree. Be sure to state the base case for the recursion and be precise when discussing the split criterion. (7 marks)

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

   c) Describe the principle of an 'ensemble' 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 key important differences between them. (10 marks)

2. Jennifer works as a professional diamond assessor. She attempts to automate some of her job by using machine learning – she receives some data from her boss, where the aim is to predict if the price of a diamond is over £10,000 or not, given its features.

   She has 750 examples of diamonds, with 50 features, where the label $y = 1$ means the diamond was priced over £10,000. She splits this into a training and testing set, and builds a perceptron model from the training data. When she predicts on the testing data, the program she is using generates the following matrix.

		Truth	
		0	1
Prediction	0	126	23
	1	32	69

   a) How many examples did Jennifer have in her testing dataset? (1 mark)

   b) How many of those testing examples are diamonds worth over £10,000 ? (1 mark)

   c) Compute the error rate, the sensitivity, and the specificity of the perceptron. (6 marks)

   d) If Jennifer undervalues the diamond (i.e. predicts it is worth less than it actually is) then it costs the company an average of £1000 per diamond. If she overvalues (i.e. predicts it is worth more than it is) then it costs the company nothing. What is the cost of the classifier predictions above in the testing data? (2 marks)

   e) Jennifer's boss is not convinced by the analysis, and asks her to explain how the perceptron works. Jennifer is your friend, so she asks you (a machine learning expert) for help. Write a full description of: (i) how the perceptron makes its decisions, (ii) how the learning algorithm works, (iii) what types of problems it can and cannot solve, and (iv) what parameters in the model can be tuned by the user to make it work more effectively, in terms of its accuracy and computational complexity.

   (10 marks)

Section C

Answer *one* question only from this section.

3. The *Support Vector Machine* (SVM) is an effective and popular Machine Learning algorithm for classification.

   a) Describe the general principle of an SVM and its learning strategy. In your description, it is essential to explain the important concepts in an SVM, e.g., margin, support vectors and cost (or loss) function used for learning. Explain why the SVM outperforms different perceptron classifiers trained on the same data set in terms of generalisation for a given linearly separable classification task. (8 marks)

   b) Given a training data set $X = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$, its weight vector, $\mathbf{w}$, and bias, b, of the SVM trained on X have the following forms:

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n, \quad b = y_k - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \cdot \mathbf{x}_k$$

   where $\mathbf{x}_n$ is a support vector only if $\alpha_n \neq 0$ for $n = 1, \cdots, N$, $\mathbf{x}_n \cdot \mathbf{x}_k$ is the dot product between two vectors $\mathbf{x}_n$ and $\mathbf{x}_k$, and $\alpha_k \neq 0$. Hence, its optimal decision boundary has the following form:

$$\sum_{n=1}^{N} \alpha_n y_n (\mathbf{x}_n \cdot \mathbf{x} - \mathbf{x}_n \cdot \mathbf{x}_k) + y_k = 0.$$

   Explain implications of all above formulae in terms of SVM learning, its application and computational efficiency in its extension to kernel SVM.

   (6 marks)

   c) The original SVM can be extended to nonlinear SVM to solve a linearly non-separable classification problem. Describe the general principle of nonlinear SVM. A kernel function is always used in nonlinear SVM. Describe what a kernel function is and explain why the use of a kernel function in nonlinear SVM makes its training and testing very computationally efficient. (6 marks)

4. Clustering analysis is an unsupervised learning process and required by different real world applications. Also clustering analysis algorithms share some common issues with supervised learning algorithms in their applications.

   a) *Agglomerative* is a popular hierarchical clustering algorithm. Describe this algorithm in detail and give *one advantage* and *one disadvantage* of the *Agglomerative* algorithm. (6 marks)

   b) Clustering analysis may be an effective technique to facilitate establishing a web search engine. Suppose that an internet company asks you to apply the Agglomerative algorithm in their web search engine for efficient textual document retrieval. Describe your method in detail. It is essential to give main steps with some justification.

      (8 marks)

   c) *K-Means* and *K Nearest Neighbours* (*K-NN*) are two simple yet popular machine learning algorithms. Describe the main differences between two algorithms. Discuss common issues both algorithms may face and possible solutions when two algorithms are used in real applications. (6 marks)

END OF EXAMINATION