MSC **Examination by course unit**

Tuesday, 17 May, 2016, 2:30 pm

ECS766P Data Mining

Duration: 2 hours 30 minutes

YOU ARE NOT PERMITTED TO READ THE CONTENTS OF THIS QUESTION PAPER UNTIL INSTRUCTED TO DO SO BY AN INVIGILATOR

Answer ALL FOUR Questions.

Cross out any answers that you do not wish to be marked.

Calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

Complete all rough workings in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately. It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

EXAM PAPERS MUST NOT BE REMOVED FROM THE EXAM ROOM

Examiners:

Tim Hospedales

Tassos Tombros

Question 1

a) When assessing a supervised learning method, the resulting accuracy of prediction on test data is obviously very important. List five additional factors that are also useful to assess the learning procedure to decide its suitability for a particular application.

[5 marks]

b) Label the following applications according to their suitability for being treated as a **classification** or **regressio**n problem in supervised learning, or **neither**.

i. Predicting whether a customer will defect to a competitor.

ii. Predicting tomorrow’s temperature.

iii. Expected profit on a financial transaction.

iv. Predictive text on a mobile phone.

v. Person recognition at an access gate.

vi. Compressing the number of bytes required to store an image.

vii. Steering system on a self driving car.

viii. Road pedestrian detector in a self driving car.

ix. Deciding suitable sizes for a line of t-shirts given height and weight of the population.

x. Airline flight price prediction.

[5 marks]

c) Linear regression and logistic regression (aka MaxEnt classification) are two methods in supervised learning. Briefly outline their similarities and differences.

[7 marks]

d) Explain the role of the objective function in macine learning. Illustrate your answer by describing the objective function of your favourite learning algorithm.

[8 marks]

Question 2

a) Suppose we are doing 1-NN classification of a query data point $\mathbf{x}^q$=[0,1,0,2,3] against the database below in order to find out its class y^q.

i. Suppose each number takes 2 bytes to store. How much memory does a 1-NN classification model for the database below require for storage?

ii. Fill in the Euclidean $\sqrt{\sum_d (x_d^q - x_d)^2}$ and Hamming $\sum_d I(x_d^q \neq x_d)$ distances of the query point $\mathbf{x}^q$ to each point in the database.

iii. What class would the query point $\mathbf{x}^q$ be classified as under Euclidean distance? What about hamming distance?

iv. How could you determine automatically which distance measure is more suitable?

Training Database		Euclidean distance to $\mathbf{x}^q$= [0,1,0,2,3]	Hamming distance to $\mathbf{x}^q$= [0,1,0,2,3].
Data, **X**	Label, y		
[0,2,1,3,4]	1		
[0,-1,0,0,2]	2		
[0,1.5,0.5,2.5,3.5]	1		
[0,1,0,2,6]	2		
[0,3,1,4,3]	3		

[2+4+2+2 = 10 marks]

b) Suppose we want to improve the performance of a given KNN model by building a model ensemble.

i. What is the computational complexity of KNN at testing time, given N training points of D dimensions and M testing points?

ii. How does this change if we build an ensemble of R KNN models?

[5 marks]

Question 2 continued

c) Consider a bank that has to choose between two classifiers A and B that would be used to make lending decisions to potential customers. Assume there are two possible types of customers: Good and Bad customers that would repay or default on their loans respectively. The classifier should predict the type of customer from their credit history, and social network profile. Suppose that:

- Making a loan to a good customer earns the bank £200 in interest.
- Making a loan to a bad customer costs the bank a £400 write-off.
- Rejecting a loan applicant costs the bank £10 in paperwork no matter what type of customer it was.

The bank evaluates the two classifiers on their historic dataset of 10 loan applications and outcomes. This results in the following confusion matrices for the two classifiers:

	Classifier A	Actual Customer		Classifier B	Actual Customer	
		Good	Bad		Good	Bad
Predicted Customer	Good (Accept)	3	4	Good (Accept)	2	0
	Bad (Reject)	1	2	Bad (Reject)	2	6

i. What is the expected value per customer for Classifier A?
ii. What is the expected value per customer for Classifier B?
iii. Which classifer should the bank choose?

[4+4+2=10 marks]

Question 3

a) Feature engineering is a manual data preparation step that can help a downstream supervised learning process perform better. Briefly explain the concept of feature engineering and give two examples of situations where it can provide a benefit.

[8 marks]

b) Missing data is a common challenge for data mining in practice.

i. Give two typical causes for missing data in practice.

ii. Outline two different strategies for dealing with missing data, as well as their relative merits.

iii. Which classifier that you have learned about is naturally robust to missing data? Briefly outline why.

[2+4+4=10 marks]

c) Your data mining company is building a classifier that could be sold to many different customers with different applications. One co-worker suggests thoroughly evaluating your product and advertising its accuracy to your customers. Another co-worker suggests quantifying its performance using receiver operating characteristic (ROC) curves. Which evaluation metric is indeed preferable and why?

[7 marks]

Question 4

a) Sentiment analysis systems are used to analyse if a piece of text expresses a positive or negative opinion. One application of this is to analyse the opinions expressed in reviews on ecommerce websites. Before positive/negative sentiment can be predicted, it is common practice to transform text strings such as reviews into fixed length bag of words vectors. Consider a dataset consisting of three reviews R1="I love this book", R2="This was no good", R3="This book is best book ever".

i. Why is the bag of words representation used in this situation?

ii. How many unique words are there in this dataset?

iii. Draw a matrix illustrating the dictionary and bag of words vector for each ecommerce product review string.

iv. Comment on what this example reveals about the limitation of bag of words methods.

[2+1+6+2=11 marks]

b) A test for twins in early pregnancy has probability of 0.9 of being correct, whether or not twins are present. 5% of pregnancies overall result in twins, and 14% of tests overall are positive. If a particular test is positive, what is the probability that twins are indeed present? Hint: You may wish to recall Bayes theorem: $p(B|A) = p(A|B)p(B)/p(A)$.

[5 marks]

c) You are monitoring the size of corn cobs harvested from a farm near a nuclear power plant to avoid putting mutant corn into the food supply. Your initial set of corn cobs are 2, 3 and 4 inches long. A corn cob is to be declared an abnomal mutant if the probability of its length under a Gaussian anomaly detection algorithm is under 0.01. Use the initial set of corn to fit a Gaussian for corn length. You may wish to recall the formulae for the mean $\mu = \frac{1}{N}\sum_i x_i$ and variance $\sigma^2 = \frac{1}{N-1}\sum_i (x_i - \mu)^2$ of a set of numbers, as well as the probability of a number under a Gaussian distribution $p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{1}{2\sigma^2}(x-\mu)^2$.

i. What is the probability of a 1.5 inch corn cob? Is it a mutant?

ii. What is the probability of a 6 inch corn cob? Is it a mutant?

iii. Suppose we want to improve the accuracy of this method by accounting for both length and weight of corn cobs. What would be the issue with using one univariate Gaussian for length and one univariate Gaussian for weight, and how could it be alleviated?

[3+3+3=9 marks]

End of Paper