Queen Mary
University of London

***MSC* Examination by course unit**

Friday, 01 May 2015 2:30 pm

ECS766P/U 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

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) Making reference to the performance on training and testing datasets, briefly explain the over and under-fitting pitfalls in supervised learning.

[7 marks]

b) List three factors that make over-fitting more likely in supervised learning.

[6 marks]

c) Briefly explain how over-fitting is controlled in one of the supervised learning methods that you have studied (e.g., KNN, Decision Tree, etc).

[6 marks]

d) Briefly outline the role of validation data in avoiding both over and under-fitting.

[6 marks]

Question 2

a) Compare and contrast the KNN and MaxEnt (aka Logistic Regression) approaches to classification.

[7 marks]

b) The task of learning a non-linear regression model is sometimes expressed as:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_i (y_i - \mathbf{w}^T f(\mathbf{x}_i))^2 + \lambda \mathbf{w}^T \mathbf{w}.$$

a. Briefly outline what this means in English.
b. What does the symbol λ (Lambda) mean?
c. How should the value of λ be determined?

[12 marks]

c) Outliers can adversely affect learning of regression models. Two categories of approaches to reduce the impact of outliers are based on anomaly detection and robust regression. Outline how these work.

[6 marks]

Question 3

a) Given a classification task with *N* training data, *M* testing data and *D* dimensions. The computational test-time complexity of Naïve Bayes is *O(MD)*.
   i. Give the computational test-time complexity of KNN and MaxEnt (aka Logistic regression) classifiers.
   ii. With reference to this complexity, what is the implication for the test time efficiency of these two methods if there is a large amount of training data?

[6 marks]

b) Consider training a 1-NN classifier with *N* training data, *M* validation data, and *D* dimensions. Suppose you are doing exhaustive feature selection to find out which dimensions are useful and which are irrelevant.
   i. How many combinations of features are tried in total?
   ii. How many times does the classifier run in total?

[6 marks]

c) Suppose you are working for a real-estate company, building models to predict house prices in London from data about each house, including square meters of living space and post-code. This will mean that the input attributes are of mixed data type.
   i. Number of square meters is a numeric data type. What data type is post-code?
   ii. Which of these data types is a problem for learning a regression model?
   iii. What can be done about it?

[6 marks]

d) Ensembles of supervised learning models often outperform individual models.
   i. How do model ensembles reduce the impact of overfitting?
   ii. Why is diversity important in an ensemble?
   iii. What are two ways to get diversity in an ensemble?

[7 marks]

Question 4

a) A doctor can run a test for the horrible disease *Examophobia*. The test has two possible outcomes: positive and negative. If Examophobia is present, the test comes out positive 80% of the time, and negative 20% of the time. Among the QMUL student population, Examophobia is known to occur in 10% of all students, and on average 20% of students test positive for Examophobia. A student Tom enters the clinic and tests positive for the disease. What is the probability Tom really has Examophobia? (You may wish to recall that Bayes theorem is: $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$.)

[7 marks]

b) You are in charge of making data mining technology purchasing decisions for your company. Two vendors A and B submit pre-trained methods for consideration. They provide equal accuracy. In terms of test-time computational efficiency, applying them on a small-scale sample of 0.1% your company's data reveals that the algorithm offered by vendor A takes 1 minute to run, and the algorithm offered by vendor algorithm B takes 2 minutes to run. However, the two algorithms are known to have $O_A(N^2D^2)$ and $O_B(ND)$ computational complexity respectively, for N data instances, and D dimensions. Which of these is more likely to be faster when analysing all of your company's data? Why?

[6 marks]

c) Learning models can be trained and applied on different computers. Deploying a learned model for test-time execution in an embedded system or mobile application often provides the constraint of limited memory.

i. Which supervised learning models that you have studied might this constraint rule out, and why?

ii. Suppose that a particular application needs a non-linear classifier, i.e., linear classifiers like MaxEnt (aka Logistic Regression) are ruled out. What other non-linear classifier might be suitable?

[6 marks]

d) Suppose you are applying k-means clustering for market-segmentation in ecommerce. A co-worker suggests that you should re-run the clustering algorithm multiple times. Is this a useful thing to do? Why?

[6 marks]

End of Paper