Queen Mary
University of London

***MSC* Examination by course unit**

Friday 19 May 2017 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:
Ioannis Patras
Anthony Constantinou

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, Linear Regression, etc).

[6 marks]

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

[6 marks]

[Q1 total 25 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:

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

i. Explain each of the symbols in the above equation, explaining what is given as input to the algorithm and what needs to be learned and/or determined.
ii. Briefly outline what this means in English.
iii. How should the value of λ be determined?

[12 marks]

c) Explain one of the greedy feature selection methods, that is, either forward or backward feature selection.

[6 marks]

[Q2 total 25 marks]

Question 3

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 regression problem in supervised learning, or neither.

i. Predicting whether a customer will default on their loan.

ii. Predicting stock market index.

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 bicycle factory given height of the population.

x. Prediction of fuel consumption for a flight.

[5 marks]

c) Describe how a decision tree is used for classification. How is overfitting avoided during training?

[7 marks]

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

[8 marks]

[Q3 total 25 marks]

Question 4

a) A doctor in Brazil can run a test for *Yellow fever*. The test has two possible outcomes: positive and negative. If *Yellow fever* is present, the test comes out positive 90% of the time. Among the population, *Yellow fever* is known to occur in 1% of all people, and on average 10% of people test positive for *Yellow fever*. Juan enters the clinic and tests positive for the disease. What is the probability that he really has Yellow fever? (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 assigned the task of deciding which pedestrian detection algorithm you will adopt in your self driving car 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 images of very small resolution images 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. The two algorithms are known to have $O_A(ND^2)$ and $O_B(ND)$ computational complexity respectively, for N data instances, and D dimensions. Which of these is more likely to be faster when installed on the car where a very high resolution camera will be used? 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]

[Q4 total 25 marks]

End of Paper