Queen Mary
University of London

Main Examination Period 2019

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.

Some questions need to be answered on the question book – please return both the question and the answer book.

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: Dr. Ioannis Patras, Dr. Jesus Requena Carrion

Question 1

a) In this question we will study two regression models for a dataset containing pairs (x, y), where x is the feature and y the response. The first model is defined by the linear equation $f_1(x) = x + 5$, and the second model by the quadratic equation $f_2(x) = x^2 + 1$. We will assume that the parameters of both models have been trained by using the training dataset shown in Figure 1.

(i) For a dataset consists of N samples (x_1, y_1), (x_2, y_2) … (x_N, y_N), define the Mean Square Error (MSE) of a regression model $y = f(x)$.

(ii) Given the training set shown in Figure 1, calculate the training MSE of both models.

(iii) Given the validation set shown in Figure 1, calculate the MSE of both models.

(iv) Which model would you choose to represent your data? Why would you choose it?

[15 marks]

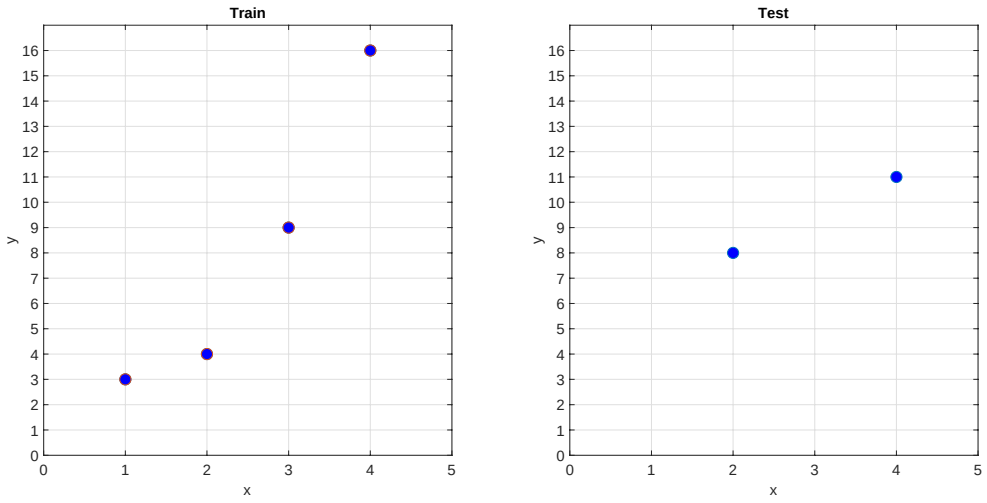

Figure 1: Training set (left) and validation set (right).

b) Explain the concept of overfitting and how it relates to the complexity of a model and the number of training samples in a dataset.

[6 marks]

c) Explain how changing the ratio of data samples assigned to the training and validation sets can affect the quality of a trained model and the estimation of its predicted performance.

[4 marks]

Question 2

a) Consider the dataset shown in Figure 2. This dataset consists of samples belonging to two classes **x** and **o** and each sample is described by two features X_A and X_B.

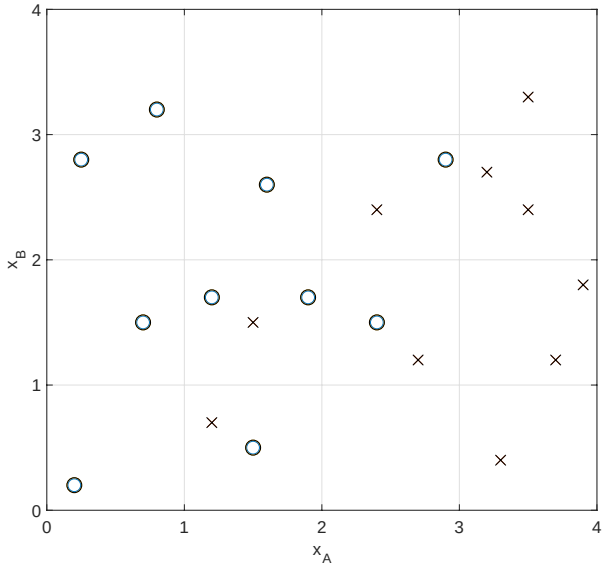

Figure 2

In this question we will study a family of linear classifiers defined by the following equation:

$$label = \begin{cases} \mathbf{x}, \text{ if } X_A \geq C \\ \mathbf{o}, \text{ if } X_A < C \end{cases}$$

where the value of *C* defines different classifiers.

(i) What is a linear classifier?

(ii) Draw on Figure 2 the boundaries of 3 classifiers defined respectively by the values $C = 1$, $C = 2$ and $C = 3$.

(iii) Define the notions of *accuracy* and *error rate* and calculate the error rate of the above classifiers using the dataset shown in Figure 2.

[10 marks]

b) In this part, we use the dataset in Figure 2 to explore the notion of confusion matrix. Let us consider the 3 linear classifiers defined in part a) by the values $C = 1$, $C = 2$ and $C = 3$, respectively.

(i) Using the dataset in Figure 2, calculate the confusion matrices of each classifier, where cells represent the number of samples.

(ii) Using the dataset in Figure 2, obtain the confusion matrices of each classifier, where cells represent rates.

(iii) Compare this family of linear classifiers with a second family of classifiers described by the equation:

$$label = \begin{cases} \mathbf{x}, \text{ if } X_B \geq C \\ \mathbf{o}, \text{ if } X_B < C \end{cases}$$

where C is a constant. Which family of classifiers do you expect to achieve a better performance? Why?

[15 marks]

Question 3

a) An ensemble model is the (weighted) sum of the decisions of individual experts/models.

(i) Why is it beneficial to increase the diversity of the experts in an ensemble?
(ii) How do boosting methods differ from methods that increase the diversity?
(iii) Describe the steps of a boosting algorithm.

[9 marks]

b) This question is about Gaussian Mixture Model (GMM) and K-means clustering.

(i) Contrast the assumptions made explicitly or implicitly by GMM versus K-means clustering algorithms.

(ii) Explain what is meant by the fact that both GMM and K-means only converge to local minima of their objective function, rather than global minimum.

(iii) In relation to the issue of convergence to local minima (cf. part (ii)), outline a simple procedure that can improve both the performance of K-means and GMM as well as the repeatability of their results.

[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 quantifying its performance using accuracy, while another co-worker suggests quantifying its performance using Receiver Operating Characteristic (ROC) curves. Which evaluation metric is indeed preferable and why?

[6 marks]

Question 4

a) In some learning-from-data problems, obtaining labels for data is costly. E.g., finding out if a financial transaction is fraudulent may require hours of an analyst's time. These situations typically result in more unlabelled data than labelled data.
   (i) Both semi-supervised and active learning aim to deal with the sparse labels problem, but there are a few important differences in how they work. What are they?
   (ii) Describe the steps of an active learning method. Give some criteria that could be useful for selecting which data samples will be labelled next.

[8 marks]

b) A test for twins in early pregnancy has probability of 0.9 of being correct, if indeed twins are present. 5% of pregnancies overall result in twins, and 10% 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: $p(B|A)= p(A|B)p(B)/p(A)$.)

[7 marks]

c) This question is about Hierarchical Clustering.

   (i) Describe the Agglomerative algorithm.

   (ii) Why one might want to use a hierarchical clustering algorithm, instead of K-Means? Give a possible application.

[10 marks]