Queen Mary
University of London

ECS766P

Data Mining MAIN Examination. ***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: Dr. Ioannis Patras, Dr. Anthony Constantinou

Question 1

a) What is the difference between nominal, ordinal, and continuous variables? You may use examples.

[3 marks]

b) What is the difference between supervised and unsupervised learning?

[4 marks]

c) Explain the difference between the Exhaustive and the Gradient Descent search methods.

[4 marks]

d) What is model underfitting and model overfitting? List and briefly explain three important factors that could contribute to model ovefitting, and three measures you could take to limit the risk of overfitting.

[10 marks]

e) In your own words, explain what is the ‘curse of dimensionality’ and why it is important.

[4 marks]

Question 2

a) Explain why correlation is not causation.

[2 marks]

b) Indicate whether each of the following statements is TRUE or FALSE, and explain why:

i. Adding more independent variables to a multiple non-linear regression, or increasing polynomial order, is guaranteed to improve the performance on the training set.
ii. Model complexity increases with the number of variables taken into consideration to learn the model.
iii. Complex models with insufficient data samples are at risk of adjusting to specific data patterns that do not represent the generic pattern.
iv. If the prediction error on the test dataset is considerably lower than the prediction error on the training dataset, then it becomes likely that the model has overfitted the test dataset.
v. When using PCA to perform dimensionality reduction, it is often suggested to preserve enough dimensions so that they can explain 1% of the variance in the dataset.

[5 marks]

c) Select four factors to compare the K-Nearest Neighbour and Naïve Bayes approaches to classification.

[8 marks]

d) What is the purpose of cross-validation? Describe the process of cross-validation.

[6 marks]

e) List and discuss two differences between Filtering and Embedded methods to feature selection.

[4 marks]

Question 3

a) 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?

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

[10 marks]

c) This question is about anomaly detection.

(i) What are some properties of applications for which anomaly detection is more suitable than supervised learning? What are the properties of applications for which supervised learning is more suitable?
(ii) Why might you want to use the multivariate Gaussian distribution in an anomaly detection application in preference to multiple univariate Gaussians?

[6 marks]

Question 4

a) A challenge in applying data mining algorithms in so called “Big Data” situations is that a vast volume of data can be stored on disk, but does not simultaneously fit into the computer’s main memory.
   (i) Why is this a problem for many traditional supervised learning algorithms?
   (ii) Name one supervised learning algorithm which is reasonable to apply in this situation and one algorithm that is not suitable. Explain why.

[8 marks]

b) Suppose that after analysing the university student records database, you discover that 20% of QMUL students typically get a distinction. 80% of the students who got a distinction regularly attended lectures. Overall, 50% of students regularly attend lectures. What is a probability that a new student George gets a distinction given that he regularly attends lectures? (You may wish to recall that Bayes theorem states that p(A|B)=p(B|A)*p(A)/p(B).)

[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]

End of Paper