COMP 24111

Two hours

EXAM PAPER MUST NOT BE REMOVED FROM
THE EXAM ROOM AND MUST BE RETURNED

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Machine Learning and Optimisation

Date: Monday 23rd January 2017

Time: 09:45 - 11:45

Answer ALL Questions in Section A
Write your answers directly on the exam paper. Only answers written in the boxes on the exam paper will be marked.

Answer ALL Questions in Section B
Answer ALL Questions in Section C

Use SEPARATE answerbooks each for Sections B and C.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]

Section A

This Section contains Multiple Choice Questions and is therefore restricted

Section B

Answer this section in a new answer booklet.

1. The concept of entropy is central to many branches of computer science including machine learning.

   a) Calculate the entropy of the following collection of pets {cat, dog, rabbit, cat, cat, cat, dog, rabbit}. For convenience use the log of base 2: $\log_2(0.5) = -1, \log_2(0.25) = -2, \log_2(0.125) = -3$. Show all your working. (2 marks)

   b) Which four pets would you remove from the collection to minimise entropy and what is the final entropy? (1 mark)

   c) Using the original collection of pets {cat, dog, rabbit, cat, cat, cat, dog, rabbit}, if you had to choose between adding a cat, dog, rabbit, or mouse, which would increase the entropy most? (1 mark)

   d) Define information gain, and describe its role in the tree learning method ID3 assume binary slits. (3 marks)

   e) What is Kulback-leiber distance? (1 mark)

2. One of the most successful techniques in machine learning is that of Random Forests.

   a) State the algorithm for Random Forests. (4 marks)

   b) State two advantages of Random Forests over a simple decision tree. (2 marks)

   c) State one disadvantage of Random Forests over a simple decision tree. (1 mark)

Section C

Answer this section in a new answer booklet.

1. Clustering analysis is an unsupervised learning process where a variety of issues need to be addressed.

   a) Distance information underlying data is crucial for clustering analysis. To measure the distance between objects of nominal features, nominal features are first converted into binary features so that the distance metric for symmetric binary features can be applied. For a cohort of professionals, four nominal features are used to characterise them: Degree = $\{$*Bsc, MSc, Ph.D*$\}$, Income = $\{$*Low, Medium, High*$\}$, Gender =$\{$*Female, Male*$\}$ and Position = $\{$*Senior, Junior*$\}$. For two professionals, their feature vectors are as follows: P1 = (*BSc, Medium, Female, Junior*) and P2 = (*BSc, Low, Male, Junior*). Based on the procedure stated above, calculate their distance. It is essential to describe all the steps to achieve the distance in your answer.

   (7 marks)

   b) For cluster validation, there are two distinct clustering validity indexes: *internal* and *external* indexes. Describe the main difference between internal and external indexes. Give one exemplary applications for each of internal and external indexes, respectively, and explain how to apply an index to your exemplary application. It is essential to state the main steps in each of your exemplary applications.

   (8 marks)

END OF EXAMINATION