

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: Wednesday 24th January 2018

Time: 14:00 - 16:00

Please 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.

Please answer all questions in Sections B and C
Use SEPARATE answerbooks each for Section.

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
contains Multiple
Choice Questions
and is restricted*

Section B

Answer this section in a new answer booklet.

1. Find a least squares regression line for the three data points: (0, 0), (1, 1), (2, 2.3) step by step, by following the instructions below.
 - a) Write down the formulation of the linear model and state what model parameters need to be optimised. (2 marks)
 - b) Write down the sum-of-squares error function. (3 marks)
 - c) Derive the partial derivative of the error function with respect to each model variable. (3 marks)
 - d) Set the derivatives to zero and compute the optimal model parameters. (2 marks)

2. Explain the k-fold cross-validation procedure used in machine learning experiments. (5 marks)

Section C

Answer this section in a new answer booklet.

1. Clustering analysis is an unsupervised learning process, and the *K-means* is a commonly used clustering algorithm.
 - a) Consider a two dimensional dataset \mathcal{D} with the objects : $O_1(2, 2)$, $O_2(2, 4)$, $O_3(4, 2)$, $O_4(4, 4)$, $O_5(3, 6)$, $O_6(7, 6)$, $O_7(9, 6)$, $O_8(5, 10)$, $O_9(8, 10)$, $O_{10}(10, 10)$. The distance metric is the *Manhattan distance*. Show the results of the *K-means* algorithm at each step, assuming that you start with two clusters ($k = 2$) with centres $C_1 = (6, 6)$ and $C_2 = (9, 7)$. It is essential to give the membership of each cluster and cluster centres at each step. Stop after five iterations if it has not already converged. (10 marks)
 - b) The *K-means* is a typical partitioning clustering algorithm. Explain how the *K-means* partitions the data space where those data points in a data set reside after the *K-means* is applied to this data set for clustering analysis. (5 marks)