

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 25th January 2016

Time: 14:00 - 16:00

---

**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]

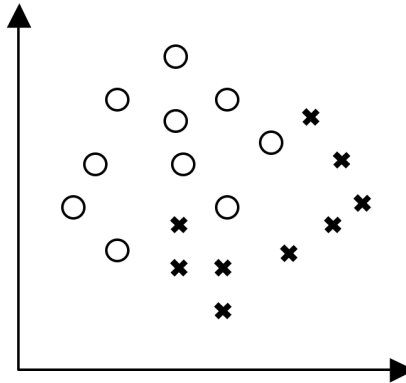
# *Section A*

*This Section contains Multiple Choice Questions and is therefore restricted*

**Section B**

**Answer this section in a new answer booklet.**

1. a) Imagine we apply a Perceptron classifier to the data below. Do you expect good performance? You must provide a *concise* reason for your answer. (1 mark)



- b) On the same data, we now apply a  $k$ -nearest neighbour classifier. Give pseudo-code for the testing phase of the KNN, which should work for any value of  $k$ . (2 marks)
- c) I have some historical weather records for Manchester over 365 days, and I encode this into a feature  $X$ . There were 100 days with rain, 250 days with clear sky, and 15 days of snow. Calculate the entropy of this variable, being sure to state which logarithm base you are using. Give the name of a Machine Learning algorithm that uses entropy. (3 marks)
- d) Give the names of THREE supervised learning models, for each stating (i) an associated model selection parameter, and (ii) what effect it has as you change it. The parameters you choose must be *unique to the model*, i.e. you cannot say “amount of training data” as this is common to all models. (6 marks)
- e) State the algorithm for Random Forests, and what makes it different from Bagging. (3 marks)

**Section C**

**Answer this section in a new answer booklet.**

1. Clustering analysis is an unsupervised learning process and often required by real world applications.
  - a) Describe the essential steps of *K-means* algorithm for clustering analysis.  
(3 marks)
  - b) For a given internal cluster validity index, describe how to apply it to a data set in order to decide the optimal number of clusters  $K$  in clustering analysis.  
(2 marks)
  - c) A new mobile phone service chain store would like to open 20 service centres in Greater Manchester. Each service centre should cover at least one shopping centre and 5,000 households of annual income over 75,000. Based on the clustering analysis knowledge you have learned from this course unit, design a scalable algorithm that decides locations of service centres by taking all the aforementioned constraints into consideration.  
(10 marks)