On admission to the examination room, you should acquaint yourself with the instructions below. You must listen carefully to all instructions given by the invigilators. You may read the question paper, but must not write anything until the invigilator informs you that you may start the examination.

You will be given five minutes at the end of the examination to complete the front of any answer books used.

May/June 2015 **SE3DM11 2014/15 A 001**

1 Answer Book
Any calculator (including programmable calculator) permitted

UNIVERSITY OF READING

DATA MINING (SE3DM11)

One and a half hours

Answer any **TWO** out of THREE questions.

EACH Question is 20 marks

1. Compare and contrast some similarity and distance measures.

(a) Provide the definition of the following proximity measures: Jaccard index, Hamming distance, Simple Matching Coefficient and cosine similarity. How can you convert a distance into a similarity?

Which approach between the Jaccard index and Hamming distance is more similar to the Simple Matching Coefficient, and which approach is more similar to the Cosine measure? Justify your answer.

(9 marks)

(b) The Minkowski Distance is a generalization of the Euclidean Distance and defines a set of norms: L_1, L_2, …, L_r. Which norm is equivalent to the Hamming distance for binary data? The Jaccard index is a measure of the similarity which can be applied to sets and to binary data. Compute the Hamming distance and the Jaccard index between the following two binary vectors.

x = 0101010001 y = 0100011001

(3 marks)

(c) Suppose that you are investigating how similar two organisms of different species are in terms of the genes they share. Which measure between Hamming and Jaccard would be more appropriate for comparing the genetic makeup of two organisms? Justify your answer. (Assume that each organism is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

(4 marks)

(d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g. two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Justify your answer. (Note that two human beings share more than 99.9% of their genes.)

(4 marks)

2. Given a set of objects each defined by a set of features, Clustering is the assignment of the objects into subsets (clusters) so that objects in the same cluster are similar and objects in different clusters are not, according to a given proximity measure defined over the set of features.

(a) What is the difference between Partitional and Hierarchical Clustering? Compare and contrast them. (5 marks)

(b) K-Means is a Partitional Clustering algorithm. What is its Hierarchical variant? Provide the pseudocode for both. (10 marks)

(c) How can we validate the quality of the set of clusters resulting from a Partitional Clustering method? (5 marks)

3. A company wants to introduce a new product and needs to evaluate the potential market for a few alternatives. They decide to leverage their vast warehouse of customer data to identify the customers most likely to be interested in each new potential product. The company hires you to design and to develop a computer program to identify the customers with affinity for the new products. You are required to propose a knowledge discovery and data mining project that can identify customer pools for the potential new products.

(a) Give an overview of the knowledge discovery process you intend to apply. List the different stages and provide a brief description of each stage.

(8 marks)

(b) Is a descriptive or a predictive data mining approach appropriate to this task? Provide an example of a descriptive data mining technique and an example of a predictive technique. Justify why one is appropriate to this task and the other is not.

(3 marks)

(c) Identify a specific data mining algorithm that could be applied and discuss its advantages and disadvantages. (5 marks)

(d) Describe the selected data mining algorithm in details by providing some pseudo-code. (4 marks)

(End of Question Paper)