Candidates are admitted to the examination room ten minutes before the start of the examination. On admission to the examination room, you are permitted to acquaint yourself with the instructions below and to read the question paper.

Do 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 2012 **SE3DM11 2011/12 A 001**

1 Answer Book
Only CASIO fx-83ES or -83MS calculators permitted

UNIVERSITY OF READING

DATA MINING (SE3DM11)

One and a half hours

Answer **TWO** questions.

1. This question compares and contrasts a number of similarity and distance measures.

   (a) Provide the definition of EACH of the following proximity measures: Jaccard index, Hamming distance, Simple Matching Coefficient and Cosine similarity. The Hamming measure is a distance, while the other three measures are similarities: how can we convert one type of measure into the other? Which approach, Jaccard index or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the Cosine measure? Explain your answer. (12 marks)

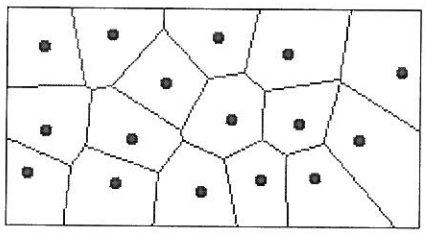
   (b) The Minkowski Distance is a generalization of the Euclidean Distance and defines a set of norms: L_1, L_2, …, L_r. Which norm is equivalent to the Hamming distance for binary data? The Jaccard index is a measure of the similarity which can be applied to sets and to binary data. Compute the Hamming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

   x = 0101010001
   y = 0100011000 (5 marks)

   (c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain 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? Explain your answer. (Note that two human beings share more than 99.9% of their genes.) (4 marks)

2. The Voronoi diagram (see figure below) for a set of K points $\{c_i\}$ in the plane is a partition of the plane into K regions, such that every point of the plane is assigned to the region containing the closest point c_i.



(a) What is the relation between the Voronoi diagram and the K-Means algorithm? (5 marks)

(b) Provide a description (in pseudocode) of the basic K-Means algorithm. (6 marks)

(c) What are the main advantages and disadvantages of K-Means for Cluster Analysis? (6 marks)

(d) How can we evaluate the clusters generated by K-Means? In general, in cluster validation, what are the internal indices for unsupervised settings? (8 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. (10 marks)

(b) Is a descriptive or a predictive data mining approach appropriate to this task? Provide ONE example of a descriptive data mining technique and ONE example of a predictive technique. Justify why one is appropriate to this task and the other is not. (5 marks)

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

(d) Use pseudo-code to describe the data mining algorithm selected in part (c) in detail. (5 marks)

(End of Question Paper)