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 2014 SE3DM11 2013/14 A 001 & SE3DM11 2012/13 A 201

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. Given a classification problem and a dataset, where each record has several attributes and a class label, a learning algorithm can be applied to the data in order to determine a classification model. The model is then used to classify previously unseen data (data without a class label) to predict the class label.

   (a) What is the general approach to learn a classification model in the form of a decision tree? What are the three main design choices in a specific decision tree induction algorithm? Provide the pseudocode of the Hunt's algorithm.

   (8 marks)

   (b) How do you measure the performance of a Decision Tree? How is the error indicated when it is computed, respectively, in the test set and in the training set?

   (3 marks)

   (c) Consider the set of records (training set) with three binary features (x, y, z) and a class label shown in Table Q1. Compare the two decision trees shown in Figure 1 by computing the two estimates of the generalization error based on the re-substitution error:

   - the optimistic estimate and
   - the pessimistic estimate with penalty term of 0.5.

   (4 marks)

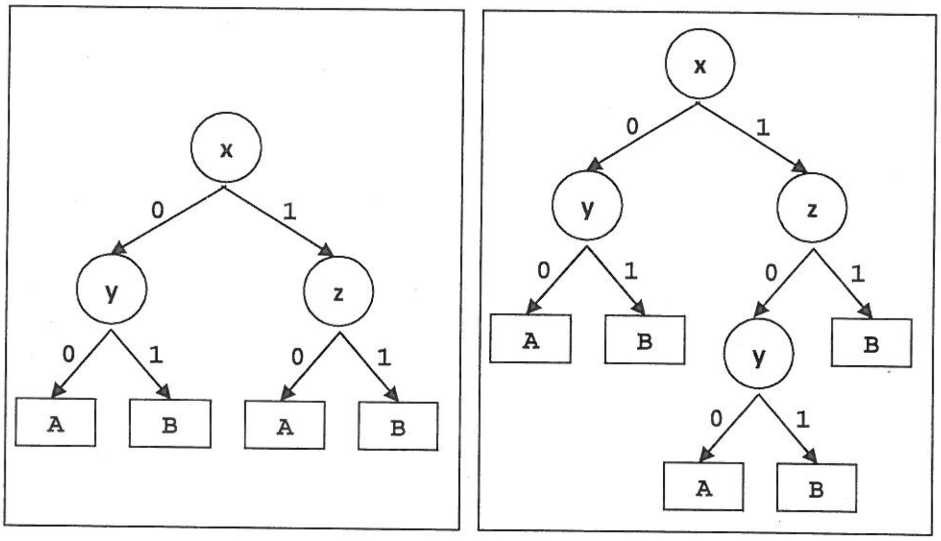
   (d) What is the meaning of the penalty term in estimating the generalisation error? For which value of the penalty term would the decision tree in *Figure Q1.a* have a smaller pessimistic estimate of the generalisation error than the one in *Figure Q1.b*?

   (5 marks)

(please see the table and figure shown on the next page...)

Instance	x	y	z	class
1	0	0	0	A
2	0	0	1	A
3	0	1	0	A
4	0	1	1	B
5	0	1	0	B
6	1	0	0	A
7	1	1	0	B
8	1	0	1	A
9	1	1	0	B
10	1	1	0	B

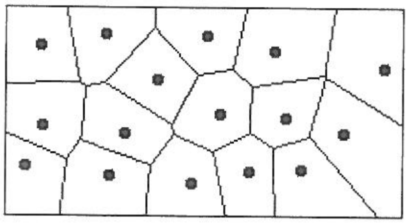
Table Q1: training data



(1.a) (1.b)

Figure Q1: Decision Trees

2. The Voronoi diagram (see figure) 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 algorithm K-Means? (5 marks)

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

(c) Discuss at least two advantages and two disadvantages of the K-Means algorithm for cluster analysis? (4 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? (5 marks)

3. A training provider ITMedia offers a variety of courses on computer skills related training to companies. ITMedia has a large database of historic data including details of:

- courses their customers (companies) have ordered,
- the employees of their customers that attended the courses,
- when the orders were placed and when the training was taken.

Using this example answer, the following sections:

(a) Give an overview of Association Rule Mining for Market Basket Analysis. How can it be applied for the above case?

(6 marks)

(b) Provide the formal definition of *support* and *confidence*. Compare and contrast them.

(6 marks)

(c) Provide details of the data attributes and structure that would be needed for generating Association Rules for the given case and discuss alternative database layouts.

(4 marks)

(d) Which extended Association Rules (e.g. generalised, quantitative, interval data, maximal, or sequential) would be suitable for the given scenario? Provide a description of the chosen extended technique.

(4 marks)

(End of Question Paper)