UNIVERSITY OF ABERDEEN SESSION 2014/2015

Examination in CS4038 Data Mining and Visualisation

Date: 10 December 2014 Time: 12.00 noon – 2.00pm

Candidates are not permitted to leave the Examination Room during the first or last half hours of the examination.

---

*Calculators Allowed. Answer any **TWO** questions. Each question is worth 25 marks; the marks for each part of a question are shown in brackets.*

Question 1:

a) There are several **types of clusters** in clustering analysis. List at least **three** different types of clusters and give a brief description for each of them.

[3]

b) Explain what z-score is. Consider the following samples:

1, 2, 3, 1, 3

Please calculate the z-score for **each** of the above samples.

[5]

c) Explain why the EM (Expectation-Maximization) clustering algorithm is considered as a generalised k-means algorithm.

[5]

d) Consider a time series represented by Piecewise Aggregate Approximation (PAA) of six segments as shown below:

Segment	PAA Value
1	0.12
2	0.34
3	0.96
4	-0.23
5	-0.56
6	0.56

Compute the Symbolic Aggregate Approximation (SAX) representation for the above time series using the breakpoint information given below:

Alphabet	Breakpoint 1	Breakpoint 2
a	Negative Infinity	< -0.84
b	>= -0.84	< -0.25
c	>=-0.25	<0.25
d	>= 0.25	<0.84
e	>=0.84	<Positive Infinity

[4]

e) Consider the following six data objects (points) *a*~*f* in the two-dimensional Euclidean space (x_1 and x_2 are their coordinates):

Point	x_1	x_2
a	1	1
b	3	1
c	1	3
d	3	3
e	5	3
f	5	1

We are going to use the k-means algorithm to cluster the above data objects into **two** clusters.

i) When Objects *a* and *c* are selected as the initial cluster centers, give detailed steps of the algorithm when processing the above data and the value of SSE (Sum of Squared Error) after convergence.

[3]

ii) When Objects *a* and *e* are selected as the initial cluster centers, give detailed steps of the algorithm when processing the above data and the value of SSE (Sum of Squared Error) after convergence.

[3]

iii) What conclusion(s) can be drawn from i) and ii)?

[2]

Note:
Please use *dist(i, j)* to represent the distance between *i* and *j*, where *i* and *j* could be any points or cluster centers. Similarly, you can use $dist^2(i,j)$ to represent the squared distance between *i* and *j*.

Question 2.

(a) For data D and hypothesis H, say whether or not the following equations must always be true.

i. $\sum_h P(H = h|D = d) = 1$
ii. $\sum_h P(D = d|H = h) = 1$
iii. $\sum_h P(D = d|H = h)P(H = h) = 1$

[3]

(b) Explain the concept "class conditional independence assumption" used by the Naïve Bayesian Classifiers.

[4]

(c) A dataset collected in an electronics shop showing details of customers and whether or not they responded to a special offer to buy a new laptop is shown in the table below. This dataset has been used to build a decision tree to predict which customers will respond to future special offers.

ID	Age	Income	Student	Credit	Buys
1	< 31	High	No	Bad	No
2	< 31	High	No	Good	No
3	31−40	High	No	Bad	Yes
4	> 40	Med	No	Bad	Yes
5	> 40	Low	Yes	Bad	Yes
6	> 40	Low	Yes	Good	No
7	31−40	Low	Yes	Good	Yes
8	< 31	Med	No	Bad	No
9	< 31	Low	Yes	Good	Yes
10	> 40	Med	Yes	Bad	Yes
11	< 31	Med	Yes	Good	Yes
12	31−40	Med	No	Good	Yes
13	31−40	High	Yes	Bad	Yes
14	> 40	Med	No	Good	No

i. Draw a decision tree using *Age* as the root node. [5]
ii. A colleague has suggested that *Student* would be a better attribute to consider at the root node rather than *Age*. Show whether this is the case or not. [9]
iii. Yet another colleague has suggested that the *ID* attribute would be a very good variable to consider at the root node. Would you agree with this suggestion? [4]

Question 3:

a) Explain what Dynamic Time Warping is. Give a simple example showing that Dynamic Time Warping is better than Euclidean distance when calculating the distance between two time series data (you could draw two time series data and make comparisons).

[4]

b) Describe the main differences between supervised and unsupervised learning. List two supervised learning algorithms and two unsupervised learning algorithms.

[3]

c) In the context of Decision Tree Learning, define what is meant by the following terms:

i. Entropy
ii. Information gain

[4]

d) On a multiple-choice exam, there are 100 questions each with 4 possible answers. A student is certain of the correct answer to each question with probability 0.6 and guesses randomly among the four choices otherwise.

i. What is the probability that the student correctly answers question 1. [3]
ii. What is the probability that the student was certain of the answer to question 1 given that they got it correct? [3]

e) Given the following proximity matrix for data points a~d, use the agglomerative hierarchical clustering algorithm to cluster these data points.

	a	b	c	d
a	1.00	0.60	0.20	0.65
b	0.60	1.00	0.70	0.60
c	0.20	0.70	1.00	0.40
d	0.65	0.60	0.40	1.00

Please **draw dendrograms (tree diagrams)** for the algorithm using the following inter-cluster similarity measures: **MIN** (Single Link), **MAX** (Complete Linkage). Please also give detailed steps of your calculation.

[8]

Note: In the detailed steps, please use sim(i,j) to represent similarity between *i* and *j*, where *i* and *j* are points or clusters. For instance, sim(a,b)=0.60 and sim(*ab*, d)=0.65, where *ab* is a cluster containing Points *a* and *b*.

END OF PAPER