UNIVERSITY OF ABERDEEN SESSION 2015/2016

Examination in CS5012 - Data Mining and Visualisation

Date: 10 December 2015 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) Explain the three important features that a data mining task must have. Given these three features and the following tasks, which of them are data mining tasks and which are not. Explain your reasons.

i) Use a computer program to extract social media data and find out what are the most trending topics.
ii) Calculate the total number of girls in a class.
iii) Divide the customers of an online music shop into several groups according to their income.
iv) Divide the customers of a supermarket into different groups based on their gender, age, income, credit history, and shopping preferences.

[4]

b) Statistically we define four levels of measurement for attribute values of data: Nominal, Ordinal, Interval, and Ratio. Consider the following attributes, please point out which levels of measurement they belong to:

i) Postal Code
ii) Temperature in Celsius
iii) Temperature in Kelvin
iv) Street numbers
v) Age
vi) Volume of Liquid in Liters

[3]

c) Explain what types of relationships you can see from a Scatter Plot.

[2]

d) The following is a normalised time series:

(-0.1, 0.4, 0.3, 1.1, -0.1, 0.2, 2.1, -1.2, 0.6)

If w=3 (the number of portions divided on the above time series), compute the PAA (Piecewise Aggregate Approximation) of the above time series.

[5]

e) Consider the following samples:

2, 1, 3, 1, 2, 3

Please calculate the **mean**, **standard deviation**, and **median** for the above samples. Then calculate the **z-score** (standard score) for **each** of the above samples.

[4]

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

Point	x_1	x_2
a	2	1
b	2	2
c	4	1
d	4	2
e	7	1
f	7	2

Use the k-means algorithm to cluster the above data objects into **three** clusters.

i) When Objects *b*, *d*, and *f* 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*, b, 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)?

[1]

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*.

PLEASE TURN OVER

Question 2:

	$X = x_1$	$X = x_2$
$Y = y_1$	0.02	0.30
$Y = y_2$	0.14	0.32
$Y = y_3$	0.10	0.12

(a) Given the joint distribution for X and Y listed in the above table calculate the following:

i. $P(X = x_1)$

[1]

ii. $P(Y = y_2)$

[1]

iii. $P(Y = y_2|X = x_1)$

[2]

(b) Distinguish between classification learning and regression learning, and give an example of each of them.

[3]

(c) Hard margin SVM cannot deal with data which is not linearly separable. In this case, explain what solution(s) can be used?

[3]

(d) Suppose we generate a training set from a decision tree and then apply decision-tree learning to the training set. Is it the case that the learning algorithm will eventually return the correct tree as the training set size goes to infinity? Why or why not?

[3]

X	Y	Class
T	T	+
T	F	-
T	F	+
T	T	+
F	T	-

(e) The above provides a classification for a data set of X Y pairs

i. Calculate the entropy for this classification.

[1]

ii. Calculate the entropy for X = T and X = F

[2.5]

iii. Calculate the information gain Y

[2.5]

(f) The table below shows a confusion matrix.

	Predicted +	**Predicted -**
True +	100	40
True -	60	300

i. Calculate the following measures: overall accuracy; precision; recall; F-Measure.

[4]

ii. Explain under what circumstances using accuracy as measure will cause issues.

[2]

Question 3:

a) Let O_1 and O_2 be two objects from the universe of possible objects. The distance (dissimilarity) is denoted by $D(O_1,O_2)$. Please explain what is Triangle Inequality. Then give one example illustrating how we can use Triangle Inequality to save computational time during data mining.

[5]

b) After you yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and that the test is 99% accurate (i.e., the probability of testing positive given that you do have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people of your age.

i. What are the values for P(test|disease) and P(¬test|¬disease) (Note: ¬is logical negation. Here 'test' denotes the testing is positive, and so '¬test' denotes the testing is negative.) ?

[2]

ii. Calculate P(disease) and P(¬disease).

[2]

iii. What are the chances that you actually have the disease?

[5]

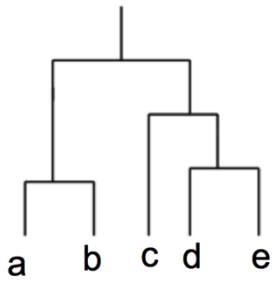
c) Is latent Dirichlet allocation (LDA) a supervised or unsupervised learning algorithm? Explain why.

[3]

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

	a	b	c	d	e
a	1.00	0.80	0.10	0.65	0.20
b	0.80	1.00	0.71	0.60	0.50
c	0.10	0.71	1.00	0.70	0.30
d	0.65	0.60	0.70	1.00	0.90
e	0.20	0.50	0.30	0.90	1.00

If the clustering result is shown as follows:



What scheme has been used in order to obtain the above result? Possible schemes are MIN, MAX, and AVERAGE. Please also give your **explanation**.

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)*=0.90 and *sim(ab, d)*=0.65, where *ab* is a cluster containing Points *a* and *b*.

[8]

END OF PAPER