UNIVERSITY OF ABERDEEN SESSION 2013/2014

Examination in CS4038 - Data Mining and Visualisation

22 January 2014 3.00pm – 5.00pm

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 what Data Mining is. Identify which of the following tasks are data mining tasks and which are not. Explain your reasons.

i) Predict whether a certain area is at risk of flooding based on previous historical weather and river levels data.
ii) Calculate the total number of female employees in a company.
iii) Divide the customers of a bank into three groups according to their income.
iv) Divide the customers of a bank into different groups based on their gender, age, income, credit history, and investment preferences.

[4]

b) Statistically we define four levels of measurement for attribute values of data: Nominal, Ordinal, Interval, and Ratio. Classify the following attribute values into these four levels of measurement:

[3]

i. Employee ID number
ii. Temperature in Celsius
iii. Temperature in Kelvin
iv. Gender: {male, female}
v. Quality of food: {good, better, best}

c) Briefly explain what you understand by the term 'Exploratory Data Analysis'.

[3]

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

[4]

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

Segment	PAA Value
1	0.72
2	0.25
3	-0.12
4	0.35
5	-0.78
6	-0.80

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	< -0.67
b	>= -0.67	< 0
c	>= 0	< 0.67
d	>= 0.67	Positive Infinity

[3]

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

	a	b	c	d	e
a	1.00	0.90	0.10	0.65	0.20
b	0.90	1.00	0.70	0.60	0.50
c	0.10	0.70	1.00	0.40	0.30
d	0.65	0.60	0.40	1.00	0.80
e	0.20	0.50	0.30	0.80	1.00

Draw dendrograms (tree diagrams) for the algorithm using the following inter-cluster similarity measures: **MIN** (Single Link), **MAX** (Complete Linkage), and **Group Average**. 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)*=0.90 and *sim(ab, d)*=0.65, where *ab* is a cluster containing Points *a* and *b*.

Question 2:

(a) Prove that $P(A|B)P(B) = P(B|A)P(A)$

[2]

(b) Explain the Naive Bayes assumption that lets us simplify the expression $P(X_1 = v_1, \ldots, X_d = v_d|C = c)P(C = c)$.

[4]

(c)

	docID	Words in document	Class label
Training set	1	apple mac iPad apple	E
	2	apple iPhone mac	E
	3	apple pear orange pear	F
Test set	4	pear apple pear mac	?

Data for parameter estimation, where E = Electronics and F = Fruit

Based on the data given in the above table:

i. Calculate the prior probability of a document occurring in each class, i.e. P(E) and P(F)

[2]

ii. Write down the vocabulary of the training set as well as the vocabulary size.

[2]

iii. For each word in the training set vocabulary, calculate the conditional word probability given a class label, i.e. P(apple|E), P(apple|F), etc.

[8]

iv. Build a multinomial Naïve Bayes classifier to determine the class label of test document *with* DocID 4 and state its predicted label.

[7]

Question 3:

(a) There are many **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 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 and make comparisons).

[3]

(c) Explain the intrinsic difference between classification and clustering. Give one example of classification and clustering, respectively.

[2]

(d) Consider the following binary classification problem as shown in the figures below.

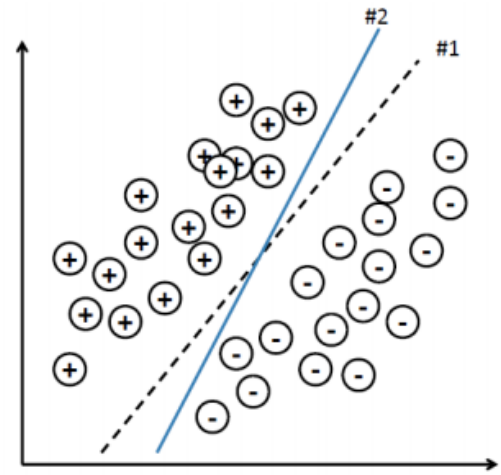

Figure 1 (a)

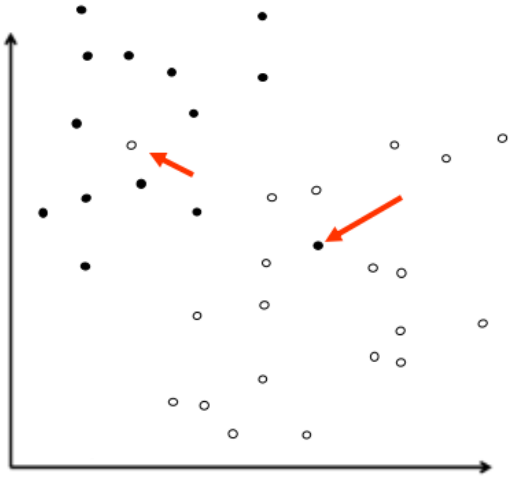

Figure 1 (b)

i. In Figure 1(a), we provide two possible linear decision boundaries.

a). Specify whether a support vector machine (SVM) will produce boundary #1, #2, or possibly both. Provide explanations to your answer.

[3]

b). Explain whether removing non-support vector instances in the training set will have impact on SVM's classification performance.

[3]

ii. Is the data in Figure 1(b) linearly separable? If yes, explain why; if not, explain what solution(s) can be used.

[4]

(e) Consider the following four data objects *a*, *b*, *c*, and *d*:

Object	x1	x2
a	2	2
b	8	6
c	6	8
d	2	4

Use the k-means algorithm to cluster the above data objects into two clusters. Assume Euclidean distance is used to measure the dissimilarity between data points.

i) When Objects *a* and *b* are first 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 *b* and *c* are first 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.

In ii), when calculating the distance between *a* and the two initial centers *b* and *c*, we get *dist(a,b)=dist(a,c)*. We require that *a* should be assigned to *b* for ease of marking.

END OF PAPER