UNIVERSITY OF ABERDEEN SESSION 2017/2018

Examination in CS551G - Data Mining and Visualisation

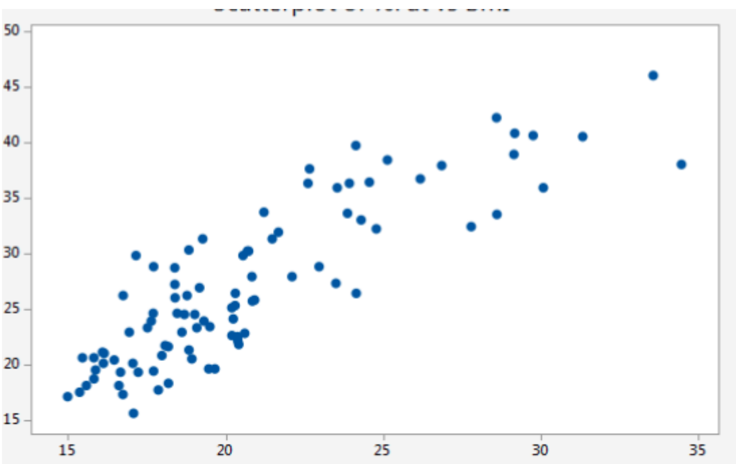
Date: 30 April 2018 Time: 3.00 pm – 5.00pm

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

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

Question 1:

a) Consider the following Scatter Plot:



Describe its strength, shape, direction, and outliers. [4]

b) Explain why the EM (Expectation-Maximization) clustering algorithm is considered as a generalised k-means algorithm. [5]

c) Consider the following samples:

1, 3, 2, 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]

PLEASE TURN OVER

d) Explain the differences between extractive and abstractive summarisation. [4]

e) Given the following **distance matrix (dissimilarity matrix)** for data points a~e, you use the agglomerative hierarchical clustering algorithm to cluster the data.

	a	b	c	d	e
a	0	0.1	0.90	0.35	0.80
b	0.1	0	0.30	0.40	0.50
c	0.90	0.30	0	0.60	0.70
d	0.35	0.40	0.60	0	0.20
e	0.80	0.50	0.70	0.20	0

Please **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 dis(i,j) to represent distance between *i* and *j*, where *i* and *j* are points or clusters. For instance, dis(a,b)=0.10 and dis(*ab*, d)=0.35, where *ab* is a cluster containing Points *a* and *b*.

Question 2:

a) **Multiple Choice:** To compute the likelihood of a sentence using a bigram model, you would: [2]

a. Calculate the conditional probability of each word given all preceding words in a sentence and multiply the resulting numbers

b. Calculate the conditional probability of each word given all preceding words in a sentence and add the resulting numbers

c. Calculate the conditional probability of each word in the sentence given the preceding word and multiply the resulting numbers

d. Calculate the conditional probability of each word in the sentence given the preceding word and add the resulting numbers

b) **Multiple Choice:** When training a language model, if we use an overly narrow corpus, the probabilities [2]

a. Don’t reflect the task

b. Reflect all possible wordings

c. Reflect intuition

d. Don’t generalize

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

Segment	PAA Value
1	-0.14
2	0.5
3	-0.96
4	0.96
5	0.56
6	-0.56

Compute the Symbolic Aggregate Approximation (SAX) representation for the above time series using the breakpoint information given below: [4]

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

d) Explain the differences between the continuous bag of words (CBOW) and skip-gram architectures for the word2vec model. [4]

e) Imagine you have two translated sentence pairs (S1, S2) for building a simple statistical machine translation model (i.e., IBM Model-1).
S1 English: a b French: x y
S2 English: a French: y

$$P(A,F|E) = \prod_{j=1}^{J} t(f_j|e_{a_j})$$

$$P(A|E,F) = \frac{P(A,F|E)}{\sum_A P(A,F|E)}$$

Here E denotes English, F denotes French, A denotes alignment models, P(A,F | E) is the probability of generating F though a particular alignment given E; P(A|E,F) is the probability of an alignment given E and F. In addition, assume the following uniform initial translation probabilities:
t(x|a)=1/2 t(y|a)=1/2 t(x|b)=1/2 t(y|b)=1/2

Based on the information given above
i. Write down all the possible word alignments given S1 and S2. [3]
ii. Compute $P(A,F|E)$. [3]
iii. Compute $P(A|E,F)$. [3]
iv. Update the translation probabilities $t(f_j|e_{a_j})$. [4]

END OF PAPER