Griffith UNIVERSITY

Student name: Perry, Sebastian Student ID: S5132483

Family name, Given name

School of Information and Communication Technology (ICT)

3804ICT Data Mining

Central Exams - Trimester 2, 2021

Reading time
10 minutes

Writing time
2 hours

Upload time
20 minutes

Examination conditions
This is an open book exam.

Writing is permitted during reading time.

No dictionaries permitted.

Scientific calculator permitted.

Instructions to students:

THIS EXAMINATION PAPER MUST NOT BE REMOVED FROM THE EXAM VENUE

Total Marks: [100 marks]

SECTION I: Short answer questions (20 marks)

Question 1. (2*5 marks) Briefly explain the following concepts (answer in 1-2 sentences).

(1) Data mining and its functions.

Data mining is the extraction of useful patterns of knowledge that can allow the ability of the following functions: to be capable of categorizing and discriminating, to associate and correlate, and to predict.

(2) Frequent patterns and strong association rules.

Frequent patterns are a set, or sequence of items which frequently appear in a given dataset. Strong association rules are rules which can predict the likelihood of something based of the presence of something else to the high accuracy due to knowledge gained from a dataset

(3) Two processes of supervised machine learning.

Two processes of supervised machine learning are: Bayes Classification Methods, and support vector machines

(4) Sequential data mining.

Is the mining of a dataset consisting of a sequence of events.

(5) Web usage data mining.

Is the automatic discovery and analysis of a users web history and otherwise recorded web usage data.

Question 2. (2*5 marks) Answer the following questions (answer in 1-3 sentences).

(1) Use the Min-max normalization and the z-score normalization to normalize the following group of data: (330,470,550,680,720,810,930).

*MinMax = ((xi – min)/(max - min)) * (new max – new min) + new min*

While new min = 0 and new max = 1

X	*330*	*470*	*550*	*680*	*720*	*810*	*930*
New X	*0.00*	*0.23*	*0.37*	*0.58*	*0.65*	*0.80*	*1.00*

(2) Briefly compare star schema, snowflake schema, and fact constellation schema.

Star schema (the most basic form which is used to create the other 2) consists of a single fact table connected to a set of dimension tables. The Snowflake schema is a refinement of the star schema, some of the dimension tables have been normalised into a set of smaller dimension tables. Finally, the Fact constellation schema instead has multiple fact tables sharing the sets of dimension tables.

(3) What are the differences between noise and outliers?

Noise is when a set of data instances may have a set of instances that have a high variance around a given point making it difficult to ascertain the origin. An outlier is when a data instance contains a value widely different to others which makes it distinctly alienated to the others in comparison.

(4) How many types of outliers are there? Please briefly describe them.
There are three different types of outliers:

Global outliers: were an instance is considered a global outlier as is it is outside the entirety of a given dataset. E.g. a person being very tall for a human being.

Contextual outliers: were an instance is considered a contextual outlier as it is unusual given the context, in a different context this value may not be unusual. E.g. a female being unusually tall for a female, but this height not being unusual for a male.

Collective outliers: were a subset of instances are considered collective outliers as they are, as a subset, unusual to the entirety of the dataset, but not in the context of nearby instances. E.g. a race of humans being taller on average when compared to others.

(5) Please briefly describe five types of Information Retrieval queries.
Five types of information retrieval queries are as follows:

Keyword queries – function with the user supplying a word, or list of words to try and find in a given document, selection of documents, or datasets. E.g. ctlr – f on a website.

Boolean queries – function with the user being able to make use of Boolean operators to find content logically true regarding the query. E.g. an example of this could be using tags, or filters.

Phrase Queries – function with the user supplying a string or sequence which will be used to find in a given document, selection of documents, or datasets were the sequence appears at least once. E.g. the usage of "" on google

Proximity queries – function by taking in a given instance and finding instances which are very similar in properties to the given instance. E.g. a suggestion to read a article which is similar to the current based on tags or keywords

Natural language queries – function in the same why people talking to each other do, with the computer inferring the context and meaning of a given question and returning the relevant answer. E.g. typing "what is the time" into google and receiving the current time in our location.

SECTION II: Problem solving questions (80 marks)
Question 3. (20 marks) Consider the following transaction data:

TID	Items bought
100	a, c, d, g, i, f, m, p
200	a, b, c, f, l, m, o
300	b, f, h, j, o, w
400	b, c, k, s, p
500	a, f, c, e, l, p, m, n

a. (10 marks) Given the transaction data above, please use FP-Growth algorithm to find out all the frequent k-itemsets (min_support_count = 3). **You need to list the steps involved.**

Where min_support_count = 3

Step 1 - frequency of 1-itemsets

Item	Sup count
C	4
F	4
A	3
m	3
P	3
B	3
L	2
O	2
D	1
G	1
I	1
H	1
J	1
W	1
K	1
S	1
E	1
N	1

Step 2 and 3 – all items with less frequency than support count, items are also sorted in descending order

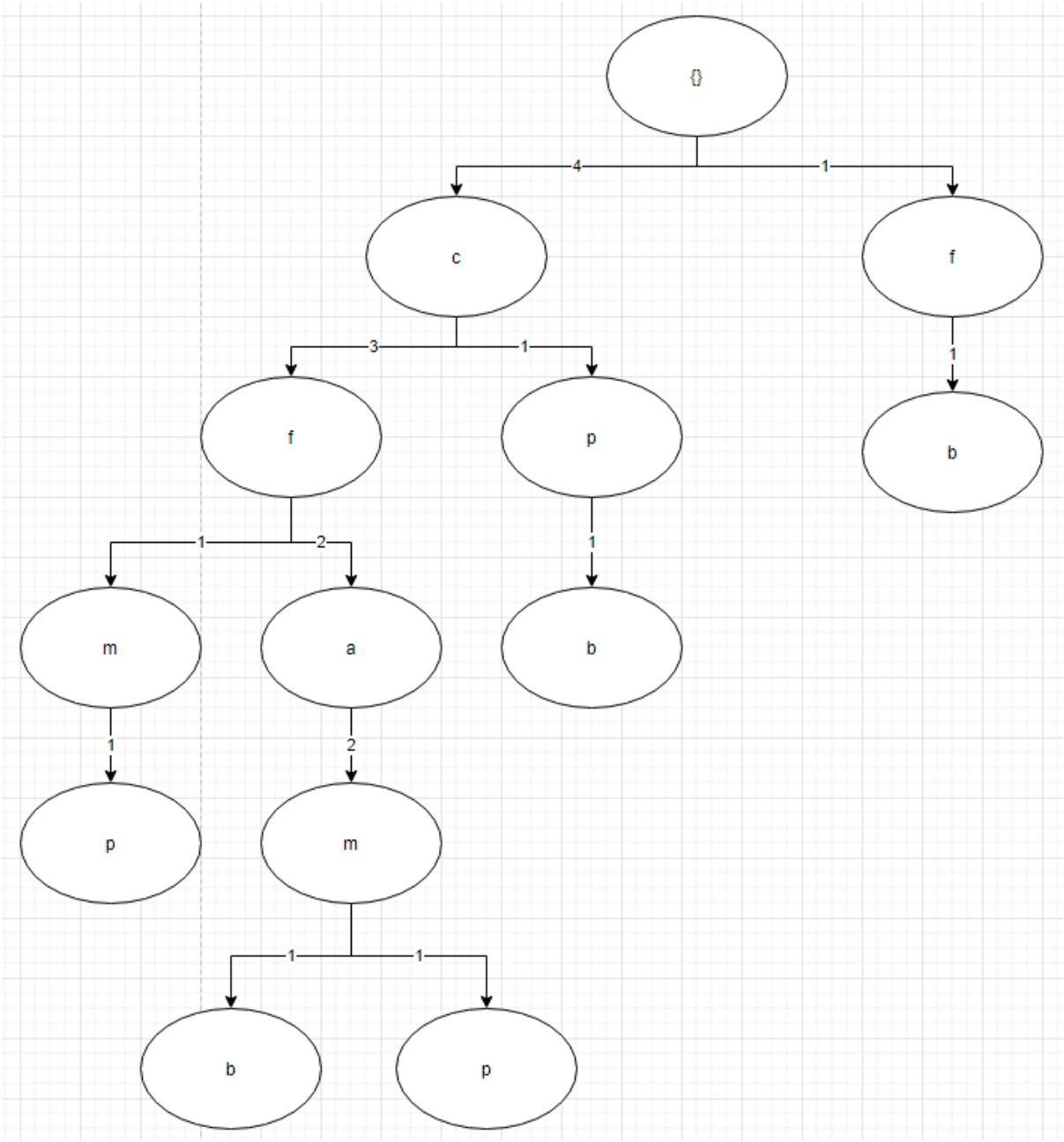
Item	Sup count
C	4
F	4
A	3
m	3
P	3
B	3

Step 4 - sort initial data based on what remains and the given order

TID	**Items bought**
100	c, f, m, p
200	c, f, a, m, b
300	f, b
400	c, p, b
500	c, f, a, m, p

Step 5 – construct tree

Step 5 – construct fp tree



Step 6 -

b. (10 marks) Based on the frequent k-itemsets, please find out all the strong association rules (min_confidence = 75%). **You need to list the steps involved.**

Question 4. (10 marks) Consider the following seasonal data:

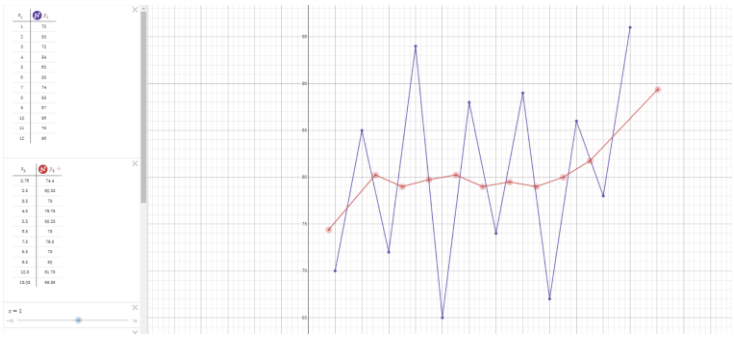
Year	Q1	Q2	Q3	Q4
2019	67	86	78	96
2018	65	88	74	89
2017	70	85	72	94

Please calculate the seasonal index and the deseasonalized data for different quarters and years, fill in the below table. **You need to list the steps involved.**

Step 1 - the 4 moving point average was calculated

Quarter	Actual	4 point moving average
2017Q1	70	
2017Q2	85	
2017Q3	72	80.25
2017Q4	94	79
2018Q1	65	79.75
2018Q2	88	80.25
2018Q3	74	79
2018Q4	89	79.5
2019Q1	67	79
2019Q2	86	80
2019Q3	78	81.75
2019Q4	96	

Step 2 – these points were both plotted to graphing software



Step 3 – the intersects on the averages line with the points were recorded as the seasonal index

Quarter	Actual	Seasonal Index	Deseasonalized Data
2017Q1	70	78.9	-8.9
2017Q2	85	78.5	+6.5
2017Q3	72	79.6	-7.6
2017Q4	94	79.3	+14.7
2018Q1	65	80	-15
2018Q2	88	79.6	+8.4
2018Q3	74	79.2	-5.2
2018Q4	89	79.2	+93.8
2019Q1	67	79.5	-12.5
2019Q2	86	80.9	+5.1
2019Q3	78	81.5	-3.5
2019Q4	96	81.4	+14.6

Question 5. (20 marks) Consider the following sequence data:

ID	Sequences
1	<{AB}{CD}{E}{C}>
2	<{A}{B}{CD}{E}>
3	<{B}{A}{BD}{E}>
4	<{CDE}{CE}>
5	<{B}{A}{BC}{AD}>

Please use Generalised Sequential Pattern (GSP) algorithm to find out all the frequent sequential patterns (min_support = 50%). **You need to list the steps involved.**

Step 1 get all singleton sequences

Item	**Sup**
A	**4**
B	**4**
C	**4**
D	**5**
E	**4**

	a	b	c	d	e
a		ab	ac	ad	ea
b			bc	bd	be
c				cd	ce
d					de
e					

Question 6. (10 marks) Assume that there is a document collection D, which has 10 documents. Given a query q, the IR system returns 10 documents in an ascending order of rankings (relevance values).

Please calculate the corresponding precision, recall, and F-score at different ranking positions. *Note: There are 5 documents actually relevant (+) to the query q.* **You need to list the steps involved.**

Rank i	+/-	P(i)	R(i)	F(i)
1	+			
2	+			
3	-			
4	-			
5	+			
6	-			
7	-			
8	+			
9	+			
10	-			

Question 7. (20 marks) Given a game rating dataset as below, please use Item-Item Collaborative Filtering by using Pearson similarity (the number of the nearest neighbour is 3, i.e. |N|=3) to estimate the ratings of "Battlefield I" and "Monster Hunter" by User 6 (rating: 1-5). **You need to list the steps involved.**

	Battlefield I	COD	Fortnight	Witcher 3	Nier Automata	Monster Hunter	Uncharted	FIFA 18	NBA 18
User 1	5	4	2	5	5	3	4	1	1
User 2	4	3	5	3	1	1	1	1	1
User 3	1	1	1	5	5	1	1	5	4
User 4	2	1	1	3	4	5	5	3	5
User 5	1	2	3	2	5	5	3	1	1
User 6	?	2	4	4	2	?	3	4	5
User 7	2	3	3	4	4	3	5	2	3
User 8	4	5	2	4	3	4	3	4	5
User 9	5	4	4	3	4	3	5	2	3
User 10	1	2	5	4	2	4	5	1	2

Step 1 normalize the ratings by subtracting row mean

	Battlefield I	COD	Fortnight	Witcher 3	Nier Automata	Monster Hunter	Uncharted	FIFA 18	NBA 18
User 1	1.66666667	0.666667	-1.33333	1.666667	1.666667	-0.33333	0.666667	-2.33333	-2.33333
User 2	1.77777778	0.777778	2.777778	0.777778	-1.22222	-1.22222	-1.22222	-1.22222	-1.22222
User 3	-1.66666667	-1.66667	-1.66667	2.333333	2.333333	-1.66667	-1.66667	2.333333	1.333333
User 4	-1.22222222	-2.22222	-2.22222	-0.22222	0.777778	1.777778	1.777778	-0.22222	1.777778
User 5	-1.55555556	-0.55556	0.444444	-0.55556	2.444444	2.444444	0.444444	-1.55556	-1.55556
User 6		-1.42857	0.571429	0.571429	-1.42857		-0.42857	0.571429	1.571429
User 7	-1.22222222	-0.22222	-0.22222	0.777778	0.777778	-0.22222	1.777778	-1.22222	-0.22222
User 8	0.22222222	1.222222	-1.77778	0.222222	-0.77778	0.222222	-0.77778	0.222222	1.222222
User 9	1.33333333	0.333333	0.333333	-0.66667	0.333333	-0.66667	1.333333	-1.66667	-0.66667
User 10	-1.88888889	-0.88889	2.111111	1.111111	-0.88889	1.111111	2.111111	-1.88889	-0.88889

END OF EXAM