

COMP1942 Exploring and Visualizing Data (Spring Semester 2024)

Homework 1

Deadline: 20 March, 2024 9:00am

(Please hand in during lecture.)

Full Mark: 100 Marks

Coupon Instructions:

1. You can use a coupon to waive any question you want and obtain full marks for this question.
2. You can waive at most one question in each assignment.
3. You can also answer the question you will waive. We will also mark it but will give full marks to this question.
4. The coupon is non-transferrable. That is, the coupon with a unique ID can be used only by the student who obtained it in class.
5. Please staple the coupon to the submitted assignment.
6. Please write down the question no. you want to waive on the coupon.

Q1 [20 Marks]

(a) Given the following transactions in a membership-only warehouse club retail store called “Sam's Club” and the support threshold = 2. Assume that the item names are sorted in lexicographic ordering (or alphabetical ordering).

AngusBeefBurger	CrabRoeNoodle	LimeJuice	MulberryJuice	RoastDuck	ScallopCongee
0	1	0	0	1	1
0	1	0	0	1	1
1	1	0	0	0	1
1	1	1	1	0	0
1	0	1	0	1	0
1	0	0	0	0	1
0	0	1	0	1	0

Follow the steps of the Apriori Algorithm and deduce L_1 , C_2 , L_2 , C_3 , L_3 , C_4 , ... until all the large itemsets are discovered. Please show the steps and list all large itemsets. You cannot use XLMiner in this part.

(b) Find all association rules where the support threshold is 2 and the confidence threshold is 90%. Please use XLMiner to find all such association rules. You just need to list all association rules for this part. You do not need to submit any softcopy related to XLMiner.

Q2 [20 Marks]

(a) Given the following transactions and the support threshold = 2.

TID	Items bought
1	b, c, d, p
2	f, j, q
3	c, i
4	a, d
5	c, m
6	b, d, f
7	a, d, f
8	a, l
9	c, g
10	c, k
11	f, n, o
12	e, f
13	f, h
14	b, d

Follow the steps of the FP-growth algorithm to find all frequent itemsets. Please show the steps and list all frequent itemsets. You cannot use XLMiner in this part.

(b) Find all association rules where the support threshold is 2 and the confidence threshold is 60%. Please use XLMiner to find all such association rules. You just need to list all association rules for this part. You do not need to submit any softcopy related to XLMiner.

Q3 [20 Marks]

Consider the following eight two-dimensional data points:

$x_1:(17, 12)$, $x_2:(5, 12)$, $x_3:(17, 14)$, $x_4:(5, 16)$, $x_5:(20, 15)$, $x_6:(3, 9)$, $x_7:(12, 3)$, $x_8:(12, 32)$

Consider algorithm k-means.

(a) Please answer the following questions. You are required to show the information about each final cluster (including the mean of the cluster and all data points in this cluster) as the output of the algorithm. You should use XLMiner to find the following answers. Please use “No. of iterations = 50, No. of starts = 10, seed = 54321 without rescaling/normalizing the input data” in XLMiner. You do not need to submit any softcopy related to XLMiner.

(i) If $k = 2$, what is the output of the algorithm?

(ii) If $k = 3$, what is the output of the algorithm?

(b) What are the advantages and the disadvantages of algorithm k-means?

Q4 [20 Marks]

Consider eight data points.

The following matrix shows the pairwise distances between any two points.

	1	2	3	4	5	6	7	8
1	0							
2	11	0						
3	5	13	0					
4	12	2	14	0				
5	7	17	1	18	0			
6	13	4	15	5	20	0		
7	9	15	12	16	15	19	0	
8	11	20	12	21	17	22	30	0

Please use the divisive (polythetic) approach to divide these eight points into two groups/clusters by using distance complete linkage.

Please write down all data points for each cluster and write down the distance between the two clusters.

You cannot use XLMiner in this question.

Q5 [20 Marks]

The following shows a list of customers with attributes “HasMacBook”, “Income” and “Age”. We also indicate whether they bought “Apple Vision Pro” or not in the last column. The first column “No.” is just for you to refer the record number only and you do not need to use this column for generating the classifier.

You cannot use XLMiner in this question.

No.	HasMacBook	Income	Age	Buy_AppleVisionPro
1	no	high	old	yes
2	no	high	middle	yes
3	yes	low	old	yes
4	yes	medium	old	yes
5	no	high	old	no
6	no	medium	young	no
7	yes	low	young	no
8	yes	high	young	no

- (a) We want to train a C4.5 decision tree classifier to predict whether a new customer will buy “Apple Vision Pro” or not. We define the value of attribute “Buy_ApplyVisionPro” to be the *label* of a record.
- Please find a C4.5 decision tree according to the above example. In the decision tree, whenever we process (1) a node containing at least 80% records with the same label or (2) a node containing at most 2 records, we stop to process this node for splitting.
 - Consider an old customer with MacBook who has high income. Please predict whether this customer will buy “Apple Vision Pro”.
- (b) What is the difference between the C4.5 decision tree and the ID3 decision tree? Why is there a difference?