COMP1942 Exploring and Visualizing Data (Spring Semester 2013)
Final Examination (Question Paper)
Date: 21 May, 2013 (Tue)
Time: 12:30-15:30
Duration: 3 hours

Student ID:_____ Student Name:_____

Seat No. :_____

Instructions:
(1) Please answer **all** questions in Part A and Part B in the **answer sheet**.
(2) You can **optionally** answer the bonus question in Part C in the answer sheet. You can obtain additional marks for the bonus question if you answer it correctly.
(3) You can use a calculator.

Question Paper

Part A (Compulsory Short Questions)

Q1 (20 Marks)

(a) Suppose that we are given a dataset with some transactions in binary format, and the support threshold = 2. Finally, we obtain the set X of all frequent itemsets equal to

{ {P}, {Q}, {R}, {S},
{P, Q}, {P, R}, {P, S}, {Q, R}, {Q, S},
{P, Q, R}, {P, Q, S} }

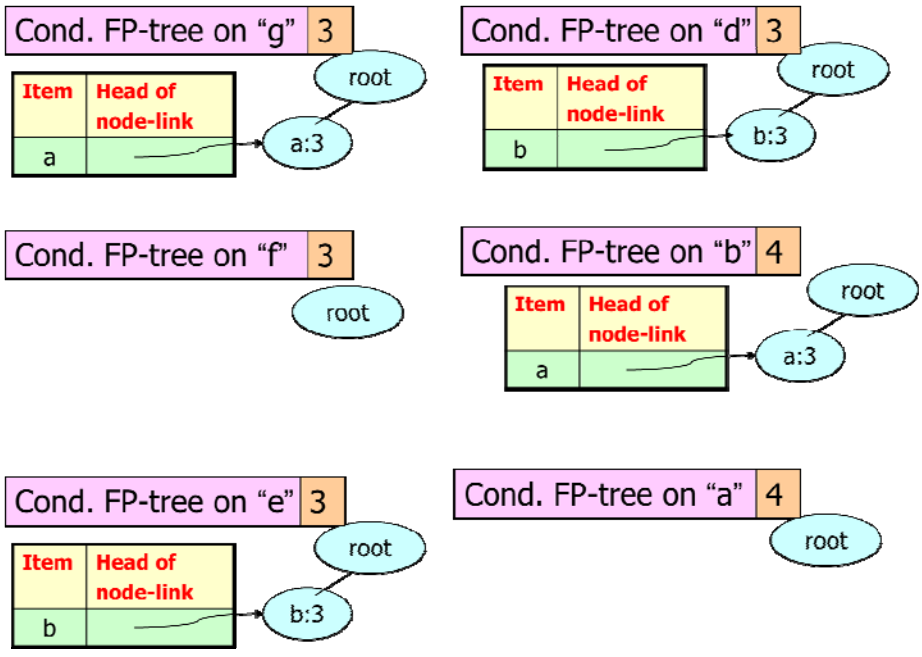
There are many possible datasets which have the same set X as the set of all frequent itemsets. Please give one possible dataset which has the minimum number of transactions in binary format. Assume that each transaction in this dataset contains items P, Q, R, S or T.

(b) In the Apriori algorithm, we know how to find some sets L_1, C_2, L_2, ….

(i) Is it always true that the number of itemsets in L_2 is smaller than or equal to the number of itemsts in C_2? If yes, please explain it. Otherwise, please give a counter example.

(ii) Is it always true that the number of itemsets in C_2 is larger than or equal to the number of itemsets in L_1? If yes, please explain it. Otherwise, please give a counter example.

(c) Consider a dataset containing items a, b, c, d, e, f and g. Let the support threshold = 3. During the execution of the FP-growth algorithm, we constructed the following conditional FP-trees.



(i) Please list out all frequent itemsets based on the above conditional FP-trees. For each frequent itemset, please also include its support.

(ii) We know that the above conditional FP-trees were constructed from an FP-tree. Is it always true that we can construct the FP-tree based on all conditional FP-trees constructed? If yes, please give the FP-tree based on the above conditional FP-trees. Otherwise, please explain it.

Q2 (20 Marks)

(a) In class, we learnt "Sequential K-means Clustering" and "Forgetful Sequential K-means Clustering". What is the scenario or application that "Forgetful Sequential K-means Clustering" is better used compared with "Sequential K-means Clustering"?

(b) 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 agglomerative approach to group these points with distance group average linkage. Draw the corresponding dendrogram for the clustering. You are required to specify the distance metric in the dendrogram.

Q3 (20 Marks)

(a) In class, we learnt that given three random variables, namely X, Y and Z, X is said to be conditionally independent of Y given Z if P(X | Y, Z) = P(X | Z).

Please state whether the following statement is true or not according to the above concept.
If yes, please give a proof. If no, please explain it.

"If X is conditionally independent of Y given Z, then P(X, Y | Z) = P(X | Z) x P(Y | Z)."

(b) One reason why we need to study subspace clustering is "curse of dimensionality". What is the meaning of "curse of dimensionality"?

(c) (i) Which classifier you learnt is similar to the way that brains process information?
(ii) What are the two advantages of the classifier in (c)(i)?

Q4 (20 Marks)

(a) We are given the following 4 data points: (6, 6), (8, 8), (5, 9), (9, 5). Use PCA to reduce from two dimensions to one dimension for each of these 4 data points. In this part, please show your steps.

(b) We are given the following 4 data points: (5, 5), (7, 7), (4, 8), (8, 4). Use PCA to reduce from two dimensions to one dimension for each of these 4 data points. In this part, please just write down the answer. You do not need to show your steps. Hints: You may use the answer of Part (a) for this part.

(c) We are given the following 4 data points: (18, 18), (24, 24), (15, 27), (27, 15). Use PCA to reduce from two dimensions to one dimension for each of these 4 data points. In this part, please just write down the answer. You do not need to show your steps. Hints: You may use the answer of Part (a) for this part.

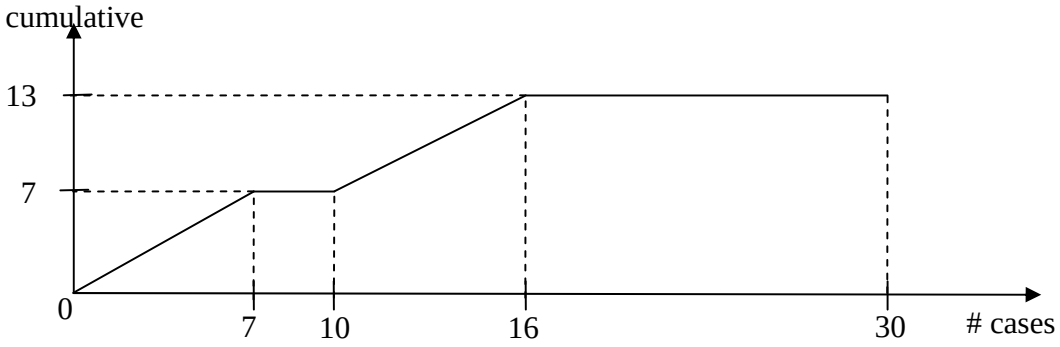
Q5 (20 Marks)

(a) Consider the following table where the first three columns correspond to the input attributes and the fourth column corresponds to the target attribute.

Race	Income	Child	Insurance
black	high	no	yes
white	high	yes	yes
white	low	yes	yes
white	low	yes	yes
black	low	no	no
black	low	no	no
black	low	no	no
white	high	no	no

In XLMiner, we want to find a decision tree (or classification tree) from the above table. But, we cannot directly use this table. Instead, we need to do a transformation process on this table so that we can use the transformed table to find a decision tree by XLMiner. What is this transformation process?

(b) This part is independent of Part (a). We are given the following lift chart based on a classifier.



(i) Is it possible to find the number of false positives? If yes, please write down the number. Otherwise, please explain it.

(ii)Is it possible to find the number of true positives? If yes, please write down the number. Otherwise, please explain it.

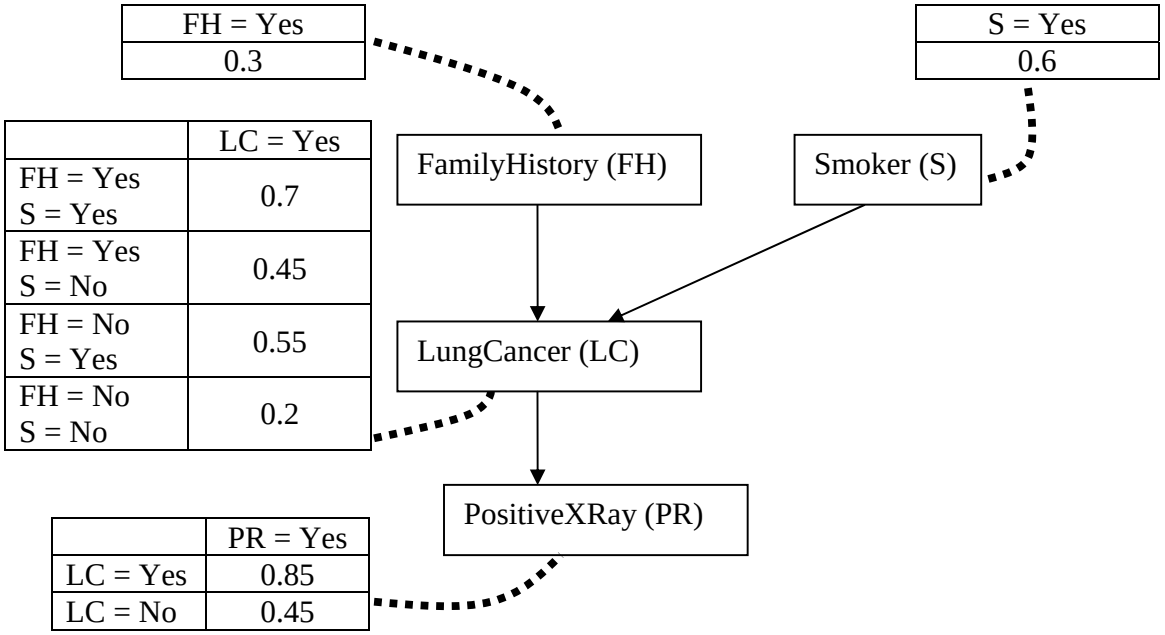
(iii) Is it possible to find the number of false negatives? If yes, please write down the number. Otherwise, please explain it.

(iv) Is it possible to find the number of true negatives? If yes, please write down the number. Otherwise, please explain it.

(v) Is it possible to find the decile-wise lift chart? If yes, please give the chart. Otherwise, please explain it.

Q6 (20 Marks)

We have the following Bayesian Belief Network.



	LC = Yes
FH = Yes S = Yes	0.7
FH = Yes S = No	0.45
FH = No S = Yes	0.55
FH = No S = No	0.2

	PR = Yes
LC = Yes	0.85
LC = No	0.45

Suppose that there is a new person. We know that

(1) he has his family history
(2) he is a smoker
(3) his result of X-Ray is negative

We would like to know whether he is likely to have Lung Cancer.

Family History	**Smoker**	**PositiveXRay**	**Lung Cancer**
Yes	Yes	No	?

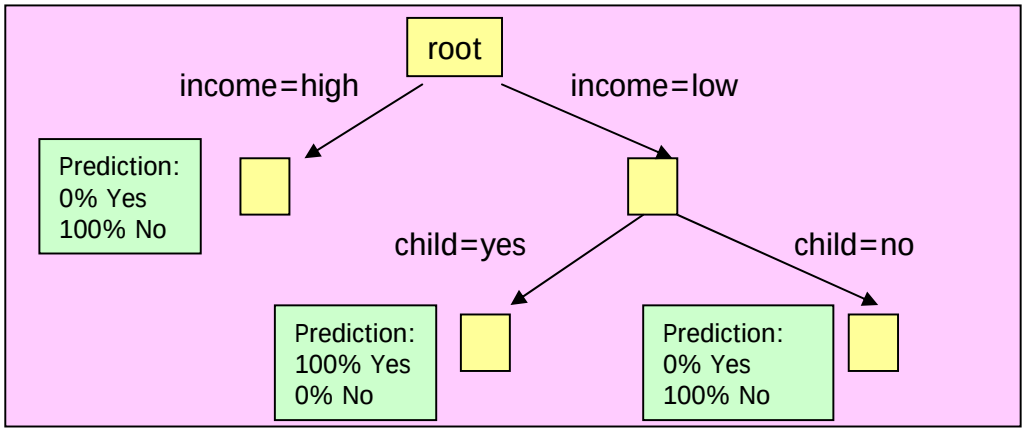
(a) Please use Bayesian Belief Network classifier with the use of Bayesian Belief Network to predict whether he is likely to have Lung Cancer.

(b) Although Bayesian Belief Network classifier does not have an independent assumption among all attributes (compared with the naïve Bayesian classifier), what are the disadvantages of this classifier?

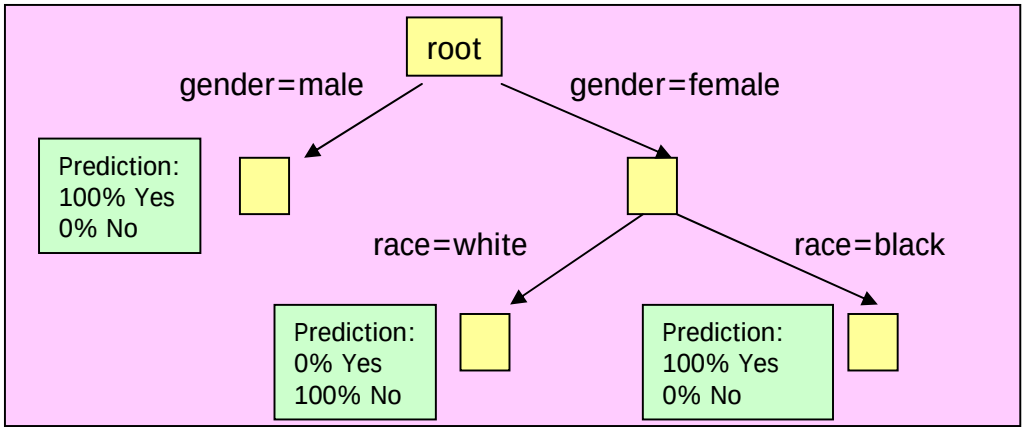
Q7 (20 Marks)

(a) We are given a table with six input attributes, namely Race, Gender, Education, Married, Income and Child, and one target attribute, namely Insurance. Based on this table, we construct three classifiers based on different criteria, namely Classifier 1 and Classifier 2 and Classifier 3.

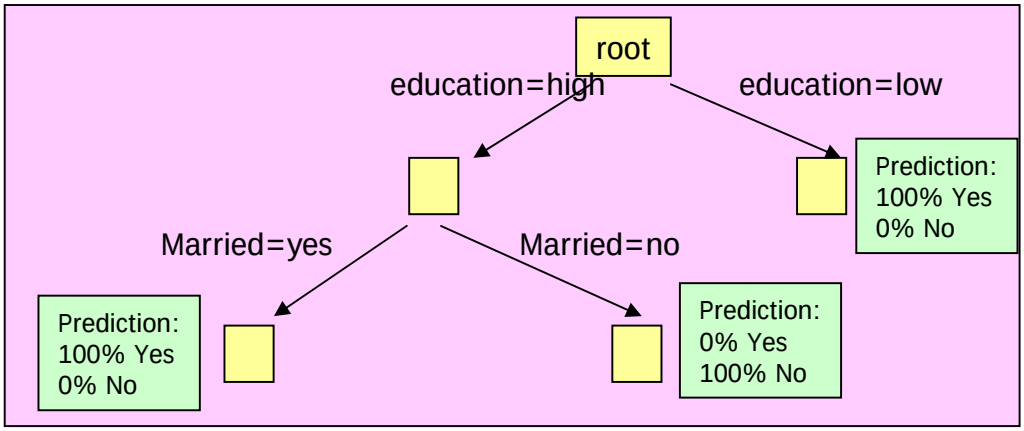
Classifier 1



Classifier 2



Classifier 3



Consider a group of classifiers called an "ensemble" studied in class. Suppose that we want to predict whether a male married customer with his "black" race who has high education and high income with a child will buy an insurance policy. What is the overall predicted result (i.e., whether the customer will buy an insurance policy)? Please elaborate how you obtain the overall predicted result.

(b) We are given the following records.

Record ID	Input Attribute 1	Input Attribute 2	Target Attribute
1	1	9	Yes
2	2	10	Yes
3	4	8	No
4	6	7	No
5	8	6	No
6	7	5	No
7	5	6	No
8	4	3	No
9	3	2	Yes
10	4	9	No
11	9	1	Yes
12	10	4	Yes
13	6	2	Yes
14	8	3	Yes

We want to predict the target attribute of the new record with the input attribute 1 equal to 5 and the input attribute 2 equal to 2. Suppose that we want to use a 3-nearest neighbor classifier and we adopt the Euclidean distance as a distance measurement between two given points. What is the target attribute of this record? Please write down the target attribute of this record and the record IDs of the corresponding 3 nearest neighbors.

(c) We know how to compute the impurity measurement of an attribute A under the ID3 decision tree, denoted by Imp-ID3(A). We also know how to compute the impurity measurement of an attribute A under the CART decision tree, denoted by Imp-CART(A). Consider two attributes A and B. Is it always true that if Imp-CART(A) > Imp-CART(B), then Imp-ID3(A) > Imp-ID3(B)? If yes, please show that it is true. Otherwise, please give a counter example showing that this is not true and then explain it.

Q8 (20 Marks)

Consider a table T: (part, supplier, customer, price) where "part" is an attribute for parts, "supplier" is an attribute for suppliers, "customer" is an attribute for customers and "price" is an attribute for prices. A record (p, s, c, x) means that the part p, supplied by supplier s and bought by customer c, has its price x. Suppose that the total size of this table is 10GB. We materialize this table.

(a) Consider the following six queries, namely Q1, Q2, Q3, Q4, Q5 and Q6.

Q1: We want to find the total price (or the sum of the prices) for each combination of part and customer.
Q2: We want to find the total price (or the sum of the prices) for each part.

Q3: We want to find the total number of records in T for each combination of part and customer.
Q4: We want to find the total number of records in T for each part.

Q5: We want to find the average price for each combination of part and customer.
Q6: We want to find the average price for each part.

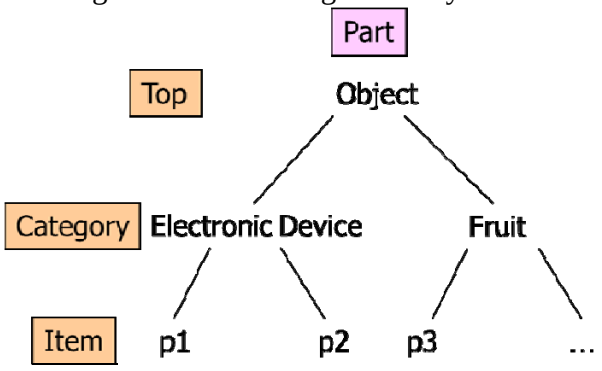
Suppose that we materialize the answers of Q1, Q3 and Q5. Each of these answers occupies 1GB storage.

We know that we can find the answer of Q2 from the answer of Q1 only in class.

(i) Is it a must that we can find the answer of Q4 from the answer of Q3 only? If yes, please explain it. Otherwise, please give what kinds of additional information (in addition to the answer of Q3) we can use with the minimum overall access cost (in terms of the total size of all materialized views accessed) and explain it.

(ii) Is it a must that we can find the answer of Q6 from the answer of Q5 only? If yes, please explain it. Otherwise, please give what kinds of additional information (in addition to the answer of Q5) we can use with the minimum overall access cost (in terms of the total size of all materialized views accessed) and explain it.

(b) We are given the following hierarchy for attribute “part”.



Consider the following two queries, namely Q7 and Q8.

Q7: We want to find the total price (or the sum of the prices) for each category in level category
Q8: We want to find the total price (or the sum of the prices) for each item in level item

(i) Suppose that a user changes his/her query from Q7 to Q8. What is the name of this query change?
(ii) Suppose that a user changes his/her query from Q8 to Q7. What is the name of this query change?

(c) We are given a table containing 100 records where the number of records with the target attribute equal to "Yes" is 30. Consider a classifier which predicts the target attribute value of only 20 records (in the table) as "Yes" (regardless of the original target attribute value).

(i) If we know that the precision of this classifier is 60%, is it always true that we can compute its recall? If yes, please write down its recall. Otherwise, please explain it.

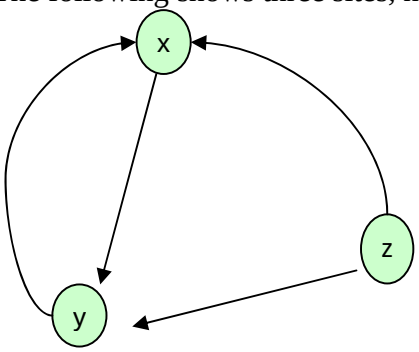
(ii) If we know that the recall of this classifier is 50%, is it always true that we can compute its accuracy? If yes, please write down its accuracy. Otherwise, please explain it.

Q9 (20 Marks)

(a) In the PageRank algorithm, we need to update a ranking vector r by $M \cdot r$ iteratively where M is the stochastic matrix. Suppose that r_n is the ranking vector after the update and r_0 is the ranking vector before the update. Prove that the sum of the values in r_n is equal to the sum of the values in r_0. For simplicity, you can assume that there are only three sites in this PageRank algorithm. In the proof, please use the following notations.

$$r_0 = \begin{pmatrix} r_{0,1} \\ r_{0,2} \\ r_{0,3} \end{pmatrix}, \qquad r_n = \begin{pmatrix} r_{n,1} \\ r_{n,2} \\ r_{n,3} \end{pmatrix}, \qquad M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}$$

(b) The following shows three sites, namely x, y and z, with their linkage.



Consider the HITS algorithm.

(i) When we compute the authority weights of all three sites, we perform an iterative step with the following formula.

$a' = X\, a$

where a' is the vector containing the updated authority weights of all three sites (i.e., x, y and z (in this ordering)) in an iteration, a is the vector containing all the previous authority weights of all three sites in an iteration, and X is a matrix.

What is the content of X?

(ii) When we compute the hub weights of all three sites, we perform an iterative step with the following formula.

$h' = Y\, h$

where h' is the vector containing the updated hub weights of all three sites (i.e., x, y and z (in this ordering)) in an iteration, h is the vector containing all the previous hub weights of all three sites in an iteration, and Y is a matrix.

What is the content of Y?

Part B (Compulsory Multiple-Choice (MC) Questions)

In this part, there are 4 multiple-choice questions, namely Q10, Q11, Q12 and Q13. The total scores in this part are 20 scores. Each question weighs 5 scores.

Q10. [Removed]

A. Statements (1) and (2) only
B. Statements (1) and (3) only
C. Statements (2) and (3) only
D. Statements (1), (2) and (3)
E. None of the above choices

Q11. [Removed]

A. Statement (1) only
B. Statement (2) only
C. Statement (3) only
D. Statements (1), (2) and (3)
E. None of the above choices

Q12. [Removed]

A. Statements (1) and (2) only
B. Statements (1) and (3) only
C. Statements (2) and (3) only
D. Statements (1), (2) and (3)
E. None of the above choices

Q13. [Removed]

A. Statement (1) only
B. Statement (2) only
C. Statement (3) only
D. Statements (1), (2) and (3)
E. None of the above choices

Part C (Bonus Question)

Note: The following bonus question is an **OPTIONAL** question. You can decide whether you will answer it or not.

Q14 (20 Additional Marks)

We are given four items, namely A, B, C and D. Their corresponding unit profits are p_A, p_B, p_C and p_D.

The following shows five transactions with these items. Each row corresponds to a transaction where a non-negative integer shown in the row corresponds to the total number of occurrences of the correspondence item present in the transaction.

A	B	C	D
0	0	3	2
3	4	0	0
0	0	1	3
1	0	3	5
6	0	0	0

The frequency of an itemset in a row is defined to be the minimum of the number of occurrences of all items in the itemset. For example, itemset {C, D} in the first row has frequency = 2. But, itemset {C, D} in the third row has frequency = 1.

The frequency of an itemset in the dataset is defined to be the sum of the frequencies of the itemset in all rows in the dataset. For example, itemset {C, D} has frequency = 2+0+1+3+0 = 6.

Define a function f on an itemset s. This function will be specified later. One example of this function is $f(s) = \sum_{i \in s} p_i$. In this example, if s = {C, D}, then $f(s) = p_C + p_D$.

The profit of an itemset s in the dataset is defined to be the product of the frequency of this itemset in the dataset and f(s).

For example, itemset {C, D} has profit = $6 \cdot f(\{C, D\})$

(a) Assume that we adopt function f such that $f(s) = (\sum_{i \in s} p_i)/|s|$ where |s| denotes the no. of items in s.
Suppose that we know that $p_A = 10$, $p_B = 10$, $p_C = 10$ and $p_D = 10$.
We want to find all itemsets with profit at least 50.
Can the Apriori Algorithm be adapted to find these itemsets?
If yes, please write down the pseudo-code and illustrate it with the above example.
If no, please explain why the Apriori Algorithm cannot be adapted. In this case, please also design an algorithm, write down the pseudo-code and illustrate it with the above example.

(b) Assume that we adopt function f such that $f(s) = \sum_{i \in s} p_i$.
Suppose that we know that $p_A = 5$, $p_B = 10$, $p_C = 6$ and $p_D = 4$.
We want to find all itemsets with profit at least 50.
Can the Apriori Algorithm be adapted to find these itemsets?
If yes, please write down the pseudo-code and illustrate it with the above example.
If no, please explain why the Apriori Algorithm cannot be adapted. In this case, please also design an algorithm, write down the pseudo-code and illustrate it with the above example.

End of Paper