COMP1942 Exploring and Visualizing Data (Spring Semester 2016)
Final Examination (Question Paper)
Date: 26 May, 2016 (Thu)
Time: 16:30-19:30
Duration: 3 hours

Student ID:_____ Student Name:_____

Seat No. :_____

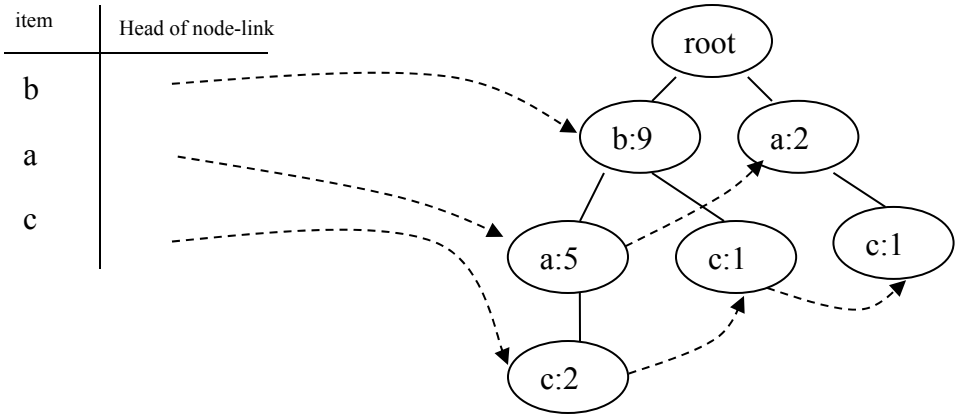
Instructions:
(1) Please answer **all** questions in Part A in the **answer sheet**.
(2) You can **optionally** answer the bonus question in Part B in the answer sheet. You can obtain additional marks for the bonus question if you answer it correctly.
(3) You can use a calculator.

Question Paper

Part A (Compulsory Short Questions)

Q1 (20 Marks)

(a) The following shows an FP-tree which is constructed from a set of transactions. Let the support threshold be 1. Please write down the corresponding transactions which are used to generate the FP-tree.



(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 itemsets 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.

Q2 (20 Marks)

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

	1	2	3	4	5	6	7
1	0						
2	10	0					
3	7	7	0				
4	30	23	21	0			
5	29	25	22	7	0		
6	38	34	31	10	11	0	
7	42	36	36	13	17	9	0

Please use the divisive (polythetic) approach to divide these seven points into two groups/clusters by using the group average linkage.

Q3 (20 Marks)

Consider Algorithm sequential k-means clustering.
When it reads a data point x, it will update the mean m of a cluster with the following operation.

$$m \leftarrow m + 1/n\ (x - m)$$

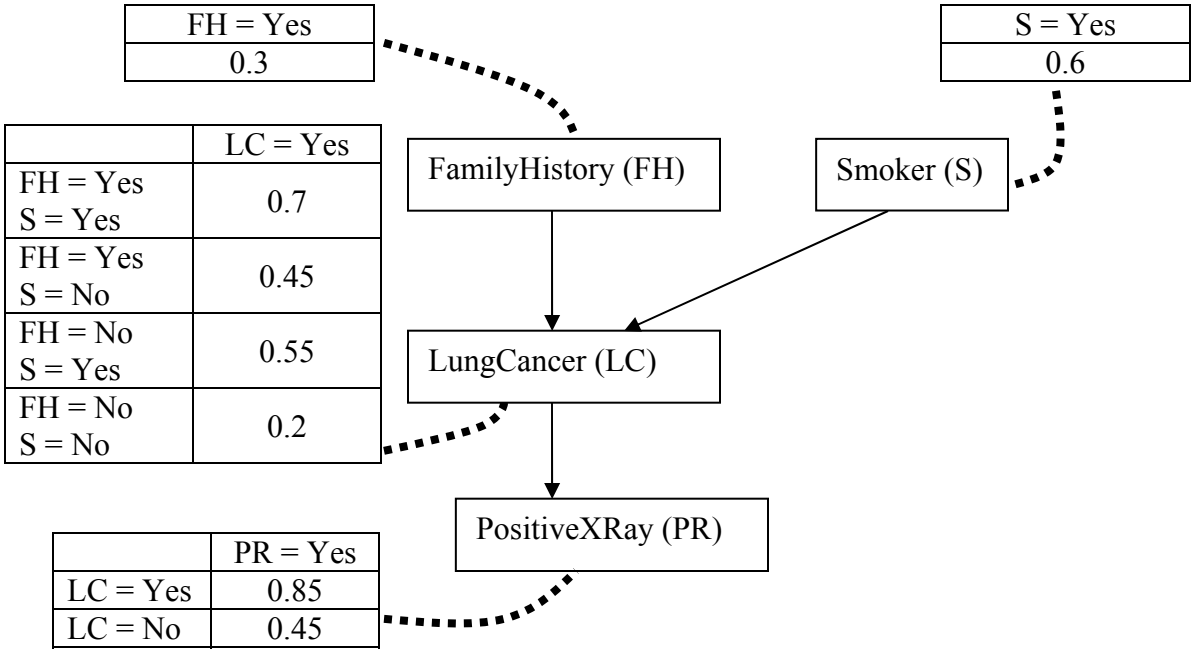
where n is the size of the cluster including the new data point x.

(a) Please write down the steps for Algorithm sequential k-means clustering.

(b) Please prove that, with the above operation, the mean m is calculated correctly. That is, the mean m calculated is equal to the expected vector among all data points in the cluster.
(Hints: Let x_j be the j-th example in cluster i and $m_i(t)$ be the mean vector of cluster i containing t examples. Consider that x is the t-th example in cluster i. Note that $m_i(t) = \frac{x_1 + x_2 + ... + x_{t-1} + x_t}{t}$.)

Q4 (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 non-smoker
(3) his result of X-Ray is positive

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

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

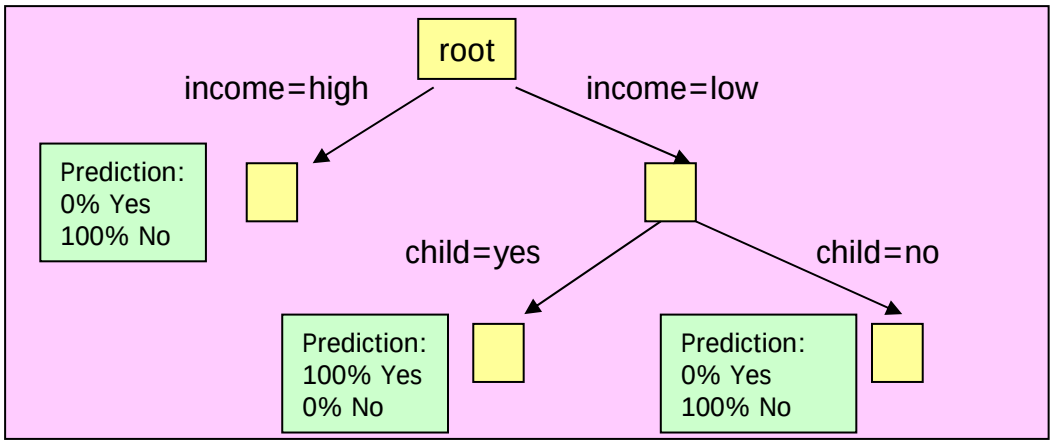
(a) Please use Bayesian Belief Network classifier with the use of Bayesian Belief Network to predict whether he is likely to have Lung Cancer.

(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?

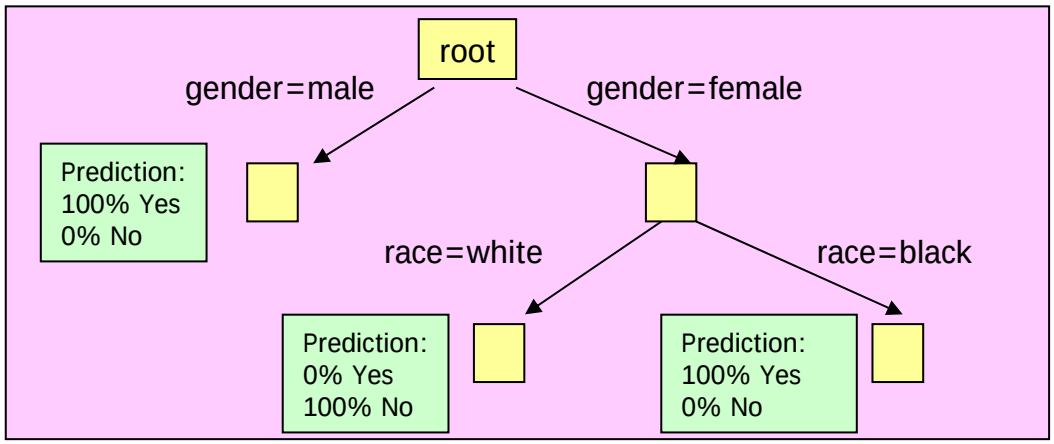
Q5 (20 Marks)

(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, Classifier 2 and Classifier 3.

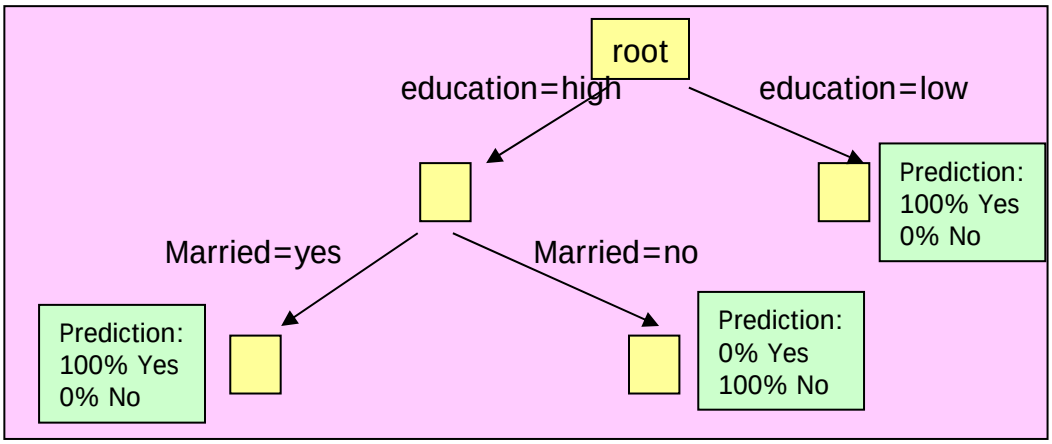
Classifier 1



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.

Q6 (20 Marks)

(a) In lecture notes, we know that a neural network containing only one neuron can only solve linearly separable problems.
Suppose that we want to solve a non-linearly separable problem.
In lecture notes, we know that we can solve this problem by using a multi-layer perceptron (i.e., a neural network which contains multiple layers and each layer contains some neurons).

Suppose that we still want to use the neural network containing only one neuron.
Is it possible to solve this non-linearly separable problem by some "additional" steps (e.g., data preprocessing)? If yes, please give a method to solve this problem using this neural network. If no, please give some reasons why this neural network cannot solve this problem by any additional steps.

(b) There are 10 data points in the dataset, namely data points 1, 2, …, 10. When we use the XLMiner software to perform "Hierarchical Clustering", we obtain the following result. In class, we learnt how to analyze the table in the result. Suppose that we want to find two clusters. Please give all data points in each of these two clusters.

Parameters/Options	
Draw Dendrogram	Yes
Selected Similarity Measure	Euclidean Distance
Selected Clustering Method	Single Linkage
Show Cluster Membership	Yes
# Clusters	2

Clustering Stages

Stage	Cluster 1	Cluster 2	Distance
1	8	9	2.236068
2	3	10	2.828427
3	3	7	2.828427
4	2	5	3.162278
5	3	8	3.605551
6	2	6	4.123106
7	2	4	5.385165
8	1	2	5.830952
9	1	3	80.05623

Q7 (20 Marks)

Consider a classification problem where the target attribute contains two possible values, “Yes” and “No”.

We are given a training dataset. We generate a classifier based on this dataset. We find that there are exactly ten tuples which target attribute values are predicted as “Yes”, and there are exactly eighteen tuples which target attribute values are predicted as “No”. We also know that the specificity of this classifier is 0.85 (or 85%) and the precision of this classifier is 0.70 (or 70%).

(a) Is it a must that we can find the accuracy of this classifier? If yes, please write down the accuracy of this classifier. Otherwise, please elaborate why we cannot find it.
(b) Is it a must that we can find the recall of this classifier? If yes, please write down the recall of this classifier. Otherwise, please elaborate why we cannot find it.
(c) Is it a must that we can find the f-measure of this classifier? If yes, please write down the f-measure of this classifier. Otherwise, please elaborate why we cannot find it.
(d) Is it a must that we can find the number of false negatives? If yes, please write down the number of false negatives. Otherwise, please elaborate why we cannot find it.
(e) Is it a must that we can find the decile-wise lift chart of this classifier? If yes, please write down the decile-wise lift chart of this classifier. Otherwise, please elaborate why we cannot find it.

Q8 (20 Marks)

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.

Q9 (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) Consider a classification problem for the table with two input attributes, namely A_1 and A_2, and one target attribute Y, containing 200 records.

(i) In the support vector machine, we learnt that we want to maximize the margin in a classification problem. We learnt that the margin is equal to

$$\frac{2}{\sqrt{w_1^2 + w_2^2}}$$

where w_1 and w_2 are two variables to be found. In class, we learnt that we need to re-write the objective function as $w_1^2 + w_2^2$ and then we want to minimize this objective function. Why do we need to re-write this objective function?

(ii) In the support vector machine, how many constraints are there in form of $Y(w_1A_1 + w_2A_2 + b) \geq 1$ where w_1, w_2 and b are three variables to be found?

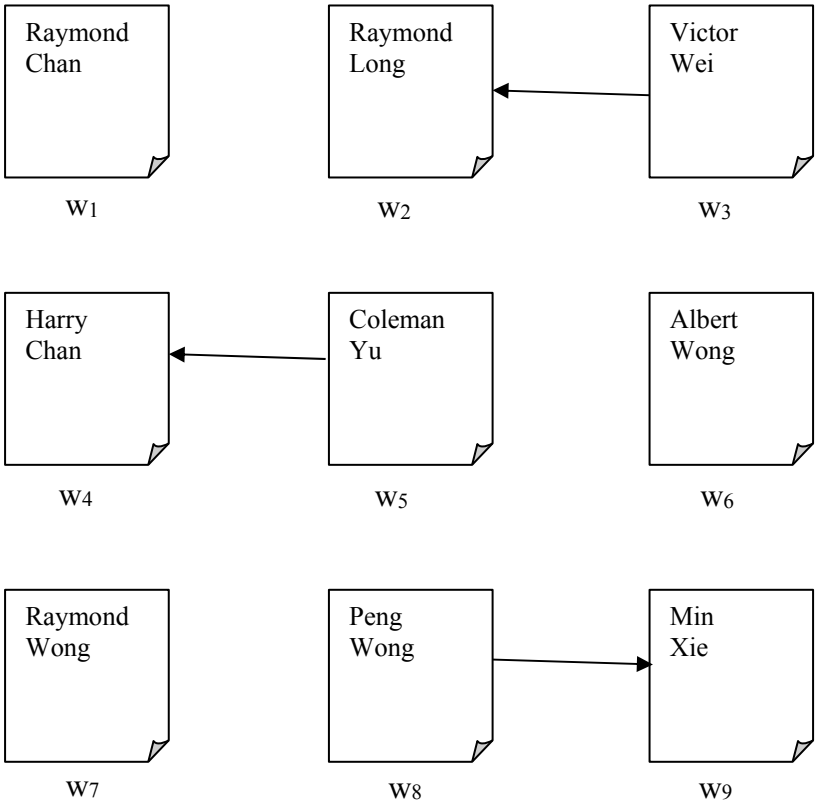
Q10 (20 Marks)

We are given the following adjacency matrix according to four sites, namely a, b, c and d.

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

(a) Is it possible to find the corresponding stochastic matrix? If yes, write down the stochastic matrix. Otherwise, please explain it.

(b) We are given the following 9 webpages, namely w_1, w_2, …, w_9.



The query terms typed by the user are "Raymond" and "Wong".

(i) What is the root set in this query? Please list the webpages in this set.

(ii) What is the base set in this query? Please list the webpage in this set.

Part B (Bonus Question)

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

Q11 (20 Additional Marks)

We are given a sequence of data points (where each data point is a *single* item) which come in real time.
Let s be the support threshold (in fraction) for finding frequent items (e.g., 20%). That is, consider a sequence of n data points, if the frequency of an item over this sequence is at least sn, then this item is a frequent item.

Consider the first problem of finding frequent *items* on the sequence from the first data point to the k-th data point where k is a positive integer.

We are given an algorithm called SpaceSaving. Given a sequence S of data points, the *summary* stored in the Space-Saving algorithm for this set is denoted by SpaceSaving(S). Let X = SpaceSaving(S). X contains two components. The first component denoted by X.E contains a list of entries each in the form of (e, f, Δ) where e is the item number, f is the frequency of the item (recorded after this entry is created in the summary) and Δ is the maximum possible error in f. The second component denoted by X.p is equal to the variable p_e used in algorithm Space-Saving (which is the greatest possible frequency error (i.e., the greatest possible difference between f and f_o where f is the frequency stored in the entry of an item and f_o is the (actual) frequency of the item over the sequence S)).

For each item stored in the summary X in the form of (e, f, Δ), the estimated frequency of this item is defined to be f +Δ. For each item not stored in the summary X, the estimated frequency of this item is defined to be X.p. For each item, the *relative error* in the estimated frequency of this item for algorithm SpaceSaving is defined to be $(f' - f_o)/f_o$ where f' is the estimated frequency of this item and f_o is the (actual) frequency of this item over the sequence S.

Let Y be the greatest possible number of entries stored in the memory used by algorithm SpaceSaving. It is shown that the greatest relative error in the estimated frequency of an item for algorithm SpaceSaving is equal to 1/Y.

We want to make use of the result from the first problem to address the second problem to be described next.

Consider the second problem of finding frequent items over a sliding window (i.e., finding frequent items on the sequence from the k-th data point to the k'-th data point where k and k' are two positive integers and k < k'). Assume that we use the *batch-based approach* (which will be elaborated next) for this purpose. Let B be the batch size. The first B-th data points form the first batch. The next B-th data points form the second batch. We can also form other batches for the remaining data points. Let B_i be the i-th batch. Whenever we reach the boundary of the batch (i.e., whenever we finish reading the last data point in the batch and are ready to read the first data point in the next batch), we want to return all frequent items over the 4 recent batches. Define N = 4B.

Suppose that we want to re-use the Space-Saving algorithm.

We have the following algorithm for this purpose.

- Let X_1 be the summary stored in the oldest batch among the 4 recent batches.
- Let X_2 be the summary stored in the 2nd oldest batch among the 4 recent batches.
- Let X_3 be the summary stored in the 3rd oldest batch among the 4 recent batches.
- Let X_4 be the summary stored in the 4th oldest batch among the 4 recent batches.
- Assume the whole memory M is divided into two parts, namely M_1 and M_2. Memory M_1 can store all data points in a single batch while memory M_2 is used to store X_1, X_2, X_3 and X_4.
- Suppose that M_2 is divided into four equal parts such that each part stores a summary X_i where i = 1, 2, 3 and 4.
- For the first four batches,
  - $X_1 \leftarrow$ SpaceSaving(B_1)
  - $X_2 \leftarrow$ SpaceSaving(B_2)
  - $X_3 \leftarrow$ SpaceSaving(B_3)
  - $X_4 \leftarrow$ SpaceSaving(B_4)
- For the each of the other remaining batches, says B_i, where i = 5, 6, …
  - Discard the content of X_1
  - $X_1 \leftarrow X_2$
  - $X_2 \leftarrow X_3$
  - $X_3 \leftarrow X_4$
  - $X_4 \leftarrow$ SpaceSaving(B_i)
- We output the frequency of each item e such that $g(e) \geq sN$ where g(e) is defined below and is regarded as the estimated frequency of e.

  Given an item e, g(e) is defined to be $\sum_{i=1}^{4} h(e, X_i)$

  Given an item e and a summary X, h(e, X) is equal to $f + \Delta$ if there exists an entry (e, f, Δ) for e in X.E. It is equal to X.p otherwise.

Note that the above symbol "←" means that the content at the right hand side of this symbol is assigned to the content at the left hand side of this symbol.

Suppose that the memory size for M_2 is 4096 bytes. (Note: You can regard that "byte" is a storage unit in the memory). Consider a summary X. Each entry in the form of (e, f, Δ) which is stored in the first component of X (i.e., X.E) occupies 12 bytes. The second component of X (i.e., X.p) occupies 4 bytes.

For each item, what is the greatest *relative error* in the estimated frequency of this item for the above algorithm? Please show your steps. Given an item e in an entry, the *relative error* in the estimated frequency of this item is defined to be $(g(e) - f_o)/f_o$ where f_o is the (actual) frequency of this item over the sequence of the 4 recent batches.

End of Paper