UIUC-CS412 "Introduction to Data Mining" (Fall 2015)

Midterm Exam

Thursday, Oct. 29, 2015
75 minutes, 75 points

Name: NetID:

1 [4′]	2 [22′]	3 [15′]	4 [14′]	5 [17′]	6 [3′]	Total [75′]

1. Short Answer Questions [4'].

   (a) [$1'$] Is the function $std(\cdot)$ (standard deviation) distributive, algebraic or holistic? Briefly explain your answer.
   [Answer: $std(\cdot)$ is algebraic. It can be computed by an algebraic function with several arguments, each of which is obtained by applying a distributive aggregate function.]

   (b) [$1'$] Name two schemas for modeling data warehouses. [Answer: Star Schema, Snowflake Schema, Fact constellations or Galaxy Schema.]

   (c) [$1'$] In order to sample 50 university students to study a disease, should we use random sampling or stratified sampling? Briefly explain your answer. [Answer: stratified sampling because the data are extremely unbalanced: the number of students who got the disease is way more bigger than the rest. If using random sampling, she might not get any students who got the disease to study. If answering random sampling and explaining that we have few information about if the disease is popular or not, you will still have full point.]

   (d) [$1'$] Briefly explain how FP-Growth performs divide-and-conquer on F-list=f-m-p. [Answer: It separate the search space of all itemsets into itemsets with item 'p', itemsets with item 'm' but no item 'p' and itemsets with item 'f' but no items 'm' and 'p'.] [Comments: If you describe the whole fp-growth algorithm or part of it but do not talk about this procedure, you get 0 point. If your answer is relevant to this procedure but is not based on the F-list we give you, you get 0.5 point.]

2. Knowing and preprocessing data [22'].

   (a) [$2'$] Assume we have two datasets. One dataset records the midterm scores of students, and the other records their final scores. It turns out that the boxplots for these two datasets are the same. Will the histograms also be the same for the two datasets? Briefly explain. [Answer: No/Not necesarry to be the same. Data distribution may be different even though the 5 number statistics (boxplot) are the same.]

   (b) [$2'$] Given a 2-D dataset contaning samples of temperature and humidity measured at different places. Name a visualization method to help see the correlation between temperature and humidity, and briefly explain. [Answer: Scatterplot/Scatterplot Matrix. Since it is a 2-D dataset, using scatterplot with temperature and humidity on each axis would show whether these two dimensions are correlated.]

   (c) [$4'$] Two students are going to study the correlation between the height and the weight of students. Both of them plan to use covariance. However, Alice uses inches and pounds, while Bob uses centimeters and kilograms.

      (i) [$1'$] Will Alice and Bob get the same value for the covariance or not? [Answer: No, because covariance is sensitive with the scales of data. In particular, it does not have normalization component so that covariance will be changed when the scales of data points are changed]

(ii) [3′] If they will, briefly explain why. If not, either propose a modification to the covariance, or propose an alternative measure so that the value does not depend on the units they use. [ANSWER: We can use Pearson correlation coefficient. They should have the same result thanks to the normalization in the denominator. If you write the formula of correlation coefficient, you will see if multiplying a measure by k, k will be canceled out in both the numerator and denominator.]

(d) [14′] Consider 4 data points in a 2-D space: (1,1), (2,2), (3,3), and (4,4). (For all sub-questions below, correct answers without explanations receive full points; and incorrect answers with explanations may receive partial credit.)

(i) [3′] Calculate the covariance matrix. [ANSWER: First, need to mean-normalize data points: $(-1.5, -1.5), (-0.5, -0.5), (0.5, 0.5), (1.5, 1.5)$. Then the covariance matrix is $\begin{bmatrix} 1.25 & 1.25 \\ 1.25 & 1.25 \end{bmatrix}$ or $\begin{bmatrix} 5/3 & 5/3 \\ 5/3 & 5/3 \end{bmatrix}$ depending on if you divide the product with 4 or 3. Without normalization, which is wrong, you will get 1 points: $\begin{bmatrix} 7.5 & 7.5 \\ 7.5 & 7.5 \end{bmatrix}$

Detailed calculation:

$$1/4 * \begin{bmatrix} -1.5 & -0.5 & 0.5 & 1.5 \\ -1.5 & -0.5 & 0.5 & 1.5 \end{bmatrix} * \begin{bmatrix} -1.5 & -0.5 & 0.5 & 1.5 \\ -1.5 & -0.5 & 0.5 & 1.5 \end{bmatrix}^T = \begin{bmatrix} 1.25 & 1.25 \\ 1.25 & 1.25 \end{bmatrix}$$

]

(ii) [2′] Calculate the correlation coefficient for the two dimensions. [ANSWER: 1 because they are completely corrleated. You can also use formula to calculate, but I can see many of you forgot to square root when calculating standard deviation, which makes the correlation is not one. In that case, and when you show all the calculation, you will get one point
]

(iii) [3′] Calculate the first and the second principal components (two vectors), and indicate which is the first principal component. (*Note: drawing is not enough, and no calculation is needed*) [ANSWER: $(1/\sqrt{2}, 1/\sqrt{2}), (-1/\sqrt{2}, 1/\sqrt{2})$. The negation of them are also correct! Please note that you have to normalize the vectors to make them unit vectors because principal componenets must be orthonormal.]

(iv) [3′] What are the coordinates of the 4 data points, *projected* to the 1-D space corresponding to the first principal component? [ANSWER: By drawing the graph, we can easily guess they are: $-3\sqrt{2}/2, -\sqrt{2}/2, \sqrt{2}/2, 3\sqrt{2}/2$. You can also do matrix multiplication, and it is not too slow to do that as well, but you have to use the mean-normalized data instead of the raw one. If using the raw one, you will get only one point. Please note that data points after projecting to the 1-D space will have only one coordinate corresponding to the 1-D space. A few of you wrote the coordinates of points in the original 2-D space, which is incorrect. Detailed calculation:

$$\begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2}\end{bmatrix} * \begin{bmatrix}-1.5 & -0.5 & 0.5 & 1.5\\ -1.5 & -0.5 & 0.5 & 1.5\end{bmatrix} = \begin{bmatrix}-3\sqrt{2}/2 & -\sqrt{2}/2 & \sqrt{2}/2 & 3\sqrt{2}/2\end{bmatrix}$$

]

(v) [3′] Is the projection of data points to the 1-D space in the previous sub-question a lossless or lossy compression? Briefly explain. [ANSWER: Lossless because we can fully recover the original data points from them. Imagine that you want to know the original coordinates for $3\sqrt{2}/2$, you can calculate the coordinates after mean-normalization: $(3/2, 3/2)$, and then the raw coordinates will be $(4, 4)$. A shorter answer could be: as all the data points are on the same line, we only need one dimension to express the data, which is exactly what PCA does]

3. Data Warehousing and OLAP for Data Mining [15'].

Consider a base cuboid of 10 dimensions that contains 4 base cells:

$$\begin{gathered}(a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10})\\(b_1, a_2, b_3, a_4, b_5, a_6, b_7, a_8, a_9, a_{10})\\(c_1, a_2, c_3, a_4, c_5, a_6, c_7, a_8, a_9, a_{10})\\(d_1, a_2, d_3, a_4, d_5, a_6, d_7, a_8, a_9, a_{10})\end{gathered}$$

where a_i, b_i, c_i and d_i are distinct for $i = 1, 3, 5, 7$. There is *no* dimension with concept hierarchy. The measure of the cube is *count*. The count of each base cell is 1.

(a) [4'] How many cuboids are there in the full data cube?
[ANSWER: 2^{10}]

(b) [6'] How many *nonempty aggregated* (i.e, *count* ≥ 1 and non-base) cells does the complete cube contain? Briefly explain your answer.
[ANSWER: First, we consider those cells with at least one of the 1st,3rd,5th,7th dimensions not aggregated(i.e. not $*$), for example cell $(d_1, *, ..., *)$. For each base cell, the number of such cells is $(2^4 - 1) \times 2^6 = 960$. Since we have 4 such base cells, thus the total number of cells in this case is $4 \times 960 = 3840$. Second, we consider those cells with the 1st,3rd,5th,7th dimensions aggregated, for example cell $(*, a_2, *, a_4, *, a_6, *, a_8, ..)$, the number of such cells is $2^6 = 64$. Since all 4 base cells are same if we only consider the 2nd,4th,6th,8th,9th,10th dimensions, the total number of cells in this case is 64. Finally, the total number of non-based cells is $3840 + 64 - 4 = 3900$.]

(c) [5'] How many *nonempty aggregated* cells does an iceberg cube contain, if the condition of the iceberg cube is *count* ≥ 4? Briefly explain.
[ANSWER: Only those cells with the 1st, 3rd, 5th and 7th dimensions aggregated(i.e. *) have count 4. And the number of such cells is $2^6 = 64$]

4. Data Cube Implementation [14'].
Suppose we use *Bottom-Up Computation* (BUC) to materialize the cube. We have a 3-D data array with 3 dimensions A, B, C. The data contained in the array are as follows:

$(a_0, b_0, c_0) : 1$	$(a_0, b_0, c_1) : 1$	$(a_0, b_0, c_2) : 1$	$(a_0, b_0, c_3) : 1$
$(a_0, b_1, c_0) : 2$	$(a_0, b_1, c_1) : 2$	$(a_0, b_1, c_2) : 2$	$(a_0, b_1, c_3) : 2$
$(a_0, b_2, c_0) : 2$	$(a_0, b_2, c_1) : 2$	$(a_0, b_2, c_2) : 2$	$(a_0, b_2, c_3) : 2$

(a) [3'] Draw the *trace trees* of expansion for the two exploration orders: $A \to B \to C$ and $C \to B \to A$.
[ANSWER: For $A \to B \to C$, see Figure 1. For $C \to B \to A$, see Figure 2.] **Note:** For the trace of C → B → A, if you reverse the order of letters in the cuboid name (e.g. you write ABC as CBA, AC as CA ...), you will be deducted 1 point since they represent different

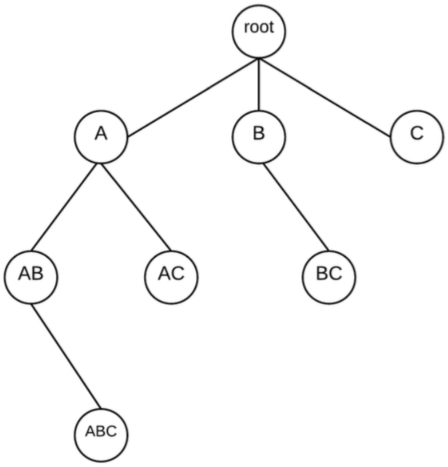

Figure 1: Trace Tree for the order of ABC

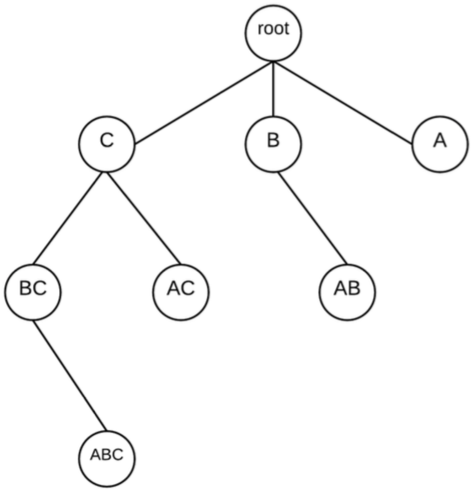

Figure 2: Trace Tree for the order of CBA

cuboids. For example, AB and BA are different cuboid since in cuboid AB, the first dimension is A, second dimension is B and the base cell is in the format (a_i, b_j). However, for cuboid BA, the first dimension is B, second dimension is A and the base cell would be (b_i, a_j).

(b) [7'] If we use exploration order $A \to B \to C$ with $min\_support = 6$, how many cells would be considered/computed? List each of them with its count and whether it is expansible in the BUC process.
If we follow the order of $A \to B \to C$, the cells with their information (count, expansibility) that need to be computed are listed as follows:

All $(*, *, *) : 20$ - expansion

A $(a_0, *, *) : 20$ - expansion

AB $(a_0, b_0, *) : 4$
AB $(a_0, b_1, *) : 8$ - expansion
AB $(a_0, b_2, *) : 8$ - expansion

ABC $(a_0, b_1, c_0) : 2$
ABC $(a_0, b_1, c_1) : 2$
ABC $(a_0, b_1, c_2) : 2$
ABC $(a_0, b_1, c_3) : 2$
ABC $(a_0, b_2, c_0) : 2$

ABC $(a_0, b_2, c_1) : 2$
ABC $(a_0, b_2, c_2) : 2$
ABC $(a_0, b_2, c_3) : 2$

AC $(a_0, *, c_0) : 5$
AC $(a_0, *, c_1) : 5$
AC $(a_0, *, c_2) : 5$
AC $(a_0, *, c_3) : 5$

B $(*, b_0, *) : 4$
B $(*, b_1, *) : 8$ - expansion
B $(*, b_2, *) : 8$ - expansion

BC $(*, b_1, c_0) : 2$
BC $(*, b_1, c_1) : 2$
BC $(*, b_1, c_2) : 2$
BC $(*, b_1, c_3) : 2$
BC $(*, b_2, c_0) : 2$
BC $(*, b_2, c_1) : 2$
BC $(*, b_2, c_2) : 2$
BC $(*, b_2, c_3) : 2$

C $(*, *, c_0) : 5$
C $(*, *, c_1) : 5$
C $(*, *, c_2) : 5$
C $(*, *, c_3) : 5$

Thus, there are totally 32 cells which would have to be computed.

Note: You need to list all cells with required information for 8 cuboids in total. Cells for (All + A + AB) worth 1 point and cells for each of other cuboids worth 1 point. The count for how many cells which would be computed worth 1 point.

(c) [4'] For the following tasks, which cube implementation method is better? Choose from Multiway Array Aggregation Computation and Bottom-Up Computation (BUC), and briefly explain.

(i) [2'] Fully materializing a small data cube with 2 dimensions. [Answer: Multiway Array Aggregation. Multiway array aggregation has the least overhead in time and space when d is small. If you say the reason to use MAAC is just that we want to fully materialize the cube, you will be deducted 1 point since BUC could also be used to materialze full cubes. You should emphasize that for low dimensional data, MAAC should be used to materialze full cube.]

(ii) [2'] Computing a large iceberg cube of 9 dimensions. [Answer: BUC. BUC is suitable for processing data with relatively higher dimesions and we could do pruning to reduce the unnecessary cost when we compute the iceberg cube.]

5. Frequent pattern and association rule mining [17'].

A database with 200 transactions has its FP-tree shown in Figure 4. Assume $min\_sup = 40\%$ and $min\_conf = 60\%$.

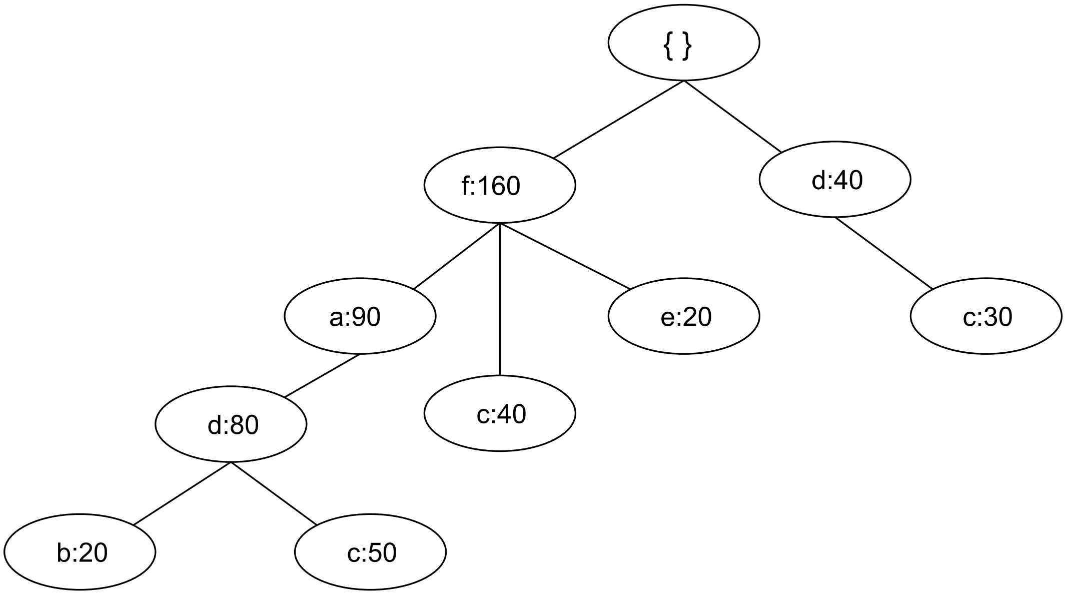

Figure 3: FP-tree of a transaction database.

(a) [4′] Show c's conditional (i.e, projected) pattern base;

[Answer: f:40, fad:50, d:30;]

[Comments: 1 point for each pattern; 1 point is deducted if the count is missing.]

(b) [4′] Draw c's conditional FP-tree;

[Answer: Figure 2]

[Comments: 2 points for tree structure (1 point is deducted if one node is missing or the order of two nodes is wrong (usually a and d)); 1 point is deducted if c appears in the conditional fp-tree; 1 point is deducted if the count is missing or wrong. We allow the node 'a' to be missing because it can be pruned due to min_sup. However, the two nodes of 'd' can not be pruned, because they need to be considered when we continue the conditioning on suffix 'dc'.]

(c) [3′] List *all* frequent itemsets based on c's conditional FP-tree;

[Answer: c, fc, dc]

[Comments: 1 point for each; 1 point is deducted if anything more than these three is presented, even they are indeed frequent patterns. This is because we explicitly ask you to base on c's conditional FP-tree, and all other patterns should be generated based on other conditional FP-trees. If you generate more than these three itemsets, it basically means you do not fully understand the divide-and-conquer philosophy and implementation of FP-Growth.]

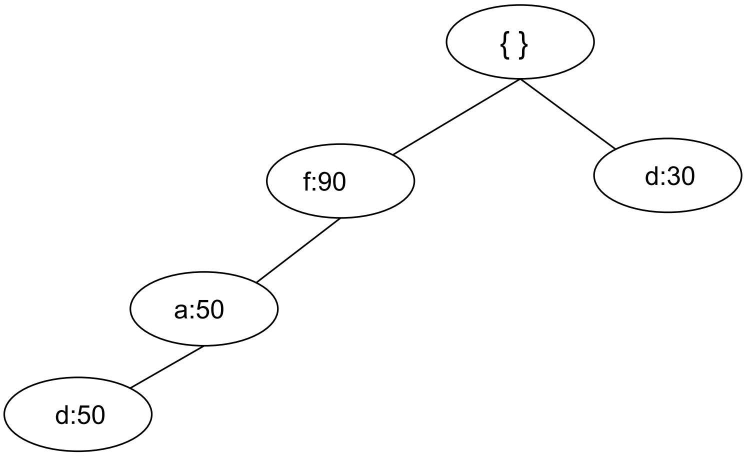

Figure 4: FP-tree of a transaction database.

(d) [2′] Compute the support and confidence of the following rules and decide if they are association rules or not accordingly.

(i) $d \to c$. [ANSWER: 0.4, 667, yes.]

(ii) $cf \to d$. [ANSWER: 0.25, 0.556, no.]

[COMMENTS: 0.5 point for each of the four figures computed above.]

(e) [4′] Answer the following questions:

(i) Based on your computations, if we want to compute *all frequent 2-itemsets* that include item c using FP-Growth, how many times do we need to scan the whole transaction database? [ANSWER: Twice. (Once.)]

[COMMENTS: 1 point. Explanation is not required. However, we provide it here: 1 for generating F-list and 1 for constructing the initial FP-tree. All following operations are done on the FP-trees and conditional pattern bases. This is one of the major contributions of FP-Growth - to compress the actual database into an FP-tree and thus avoid frequent scanning on the database. However, if you answered 'once', we still give you 1 point, because it is easy to miss the 1 time for generating F-list, and 1 time and 2 times actually make no much differences in the long run.]

(ii) To compute the same thing, will Apriori algorithm require more time to scan the whole transaction database? [ANSWER: No. (Yes.)]

[COMMENTS: 1 point. Explanation is not required. Apriori will also scan the whole database 2 times for computing all frequent 2-itemsets - 1 for generating L1 from C1 and 1 for generating L2 from C2. However, if you answered 'yes' for this question and 'once' for the previous one, we still give you 1 point.]

(iii) From this observation, if database scanning is the most time-consuming operation, which algorithm is more efficient? [ANSWER: The same. (FP-Growth.)]

[Comments: 2 points. Explanation is not required. Since we are only considering frequent 2-itemsets, for which the two algorithms need the same times of database scanning and we are focusing on the time of database scanning, the efficiency of the two algorithm should be generally the same. However, if you answered 'FP-Growth' and the answers of your previous two questions are correct, we still give you 1 point. This is because from the observation above, it is not hard to conclude that FP-Growth in the long run will be more efficient - it will require no more scanning on the database to generate any frequent k-itemsets later, but Apriori will need one more scanning for each k. Therefore, we deduct 1 point if your answer for this question is not consistent with the previous 2 (e.g. you answered something more than 'twice' for question 1 and 'no' for question 2 but 'fp-growth' here) or your answer is Apriori but consistent with your (wrong) observation. We deduct 2 points only if your answer is Apriori and also inconsistent with your previous answers, which is rare but does happen.]

6. Opinion [3'].

   (a) I □ like □ dislike the exams in this style.

   (b) In general, the exam questions are □ too hard □ too easy □ just right.

   (c) I □ have plenty of time □ have just enough time □ do not have enough time to finish the exam questions.