CS 412: Introduction to Data Mining	**Summer 2017**
Homework 2	
Handed Out: June 15, 2017	*Due: June* 27, 2017 *11:59 pm*

1 General Instructions

- This assignment is due at 11:59 PM on the due date. We will be using Compass (http://compass2g.illinois.edu) for collecting this assignment. Contact TAs if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The homework MUST be submitted in pdf format. Handwritten answers are not acceptable. Name your pdf file as YourNetid-HW2.pdf
- You need to explain the logic of your answer/result for every question. A result/answer without any explanation will not receive any points.
- It is OK to discuss the problems with the TAs and your classmates, however, it is NOT OK to work together or share code. Plagiarism is an academic violation to copy, to include text from other sources, including online sources, without proper citation. To get a better idea of what constitutes plagiarism, consult the CS Honor code (http://cs.illinois.edu/academics/honor-code) on academic integrity violations, including examples, and recommended penalties. There is a zero tolerance policy on academic integrity violations. Any student found to be violating this code will be subject to disciplinary action.
- Please use Piazza if you have questions about the homework. Also feel free to send TAs emails and come to office hours.

2 Question 1 (5 points)

1. (2') What do the eigenvectors of covariance matrix represent in PCA?
2. (3') Given the following covariance matrix A, decide if any of the following vectors are eigenvectors. Also, determine the eigenvalues corresponding to the eigenvectors.

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix}$$

(a)
$$a^T = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

(b)
$$b^T = \begin{bmatrix} 0 & 1 & 2 \end{bmatrix}$$

(c)
$$c^T = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

Solution:

1. (2') Principle components.
2. (3')
   (a) a is eigenvector with eigenvalue 2.
   (b) b is eigenvector with eigenvalue 11.
   (c) c is not eigenvector.

3 Question 2 (5 points)

Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010? In each dimension, there exists a hierarchy, which is presented in Figure 1 as a top down order.

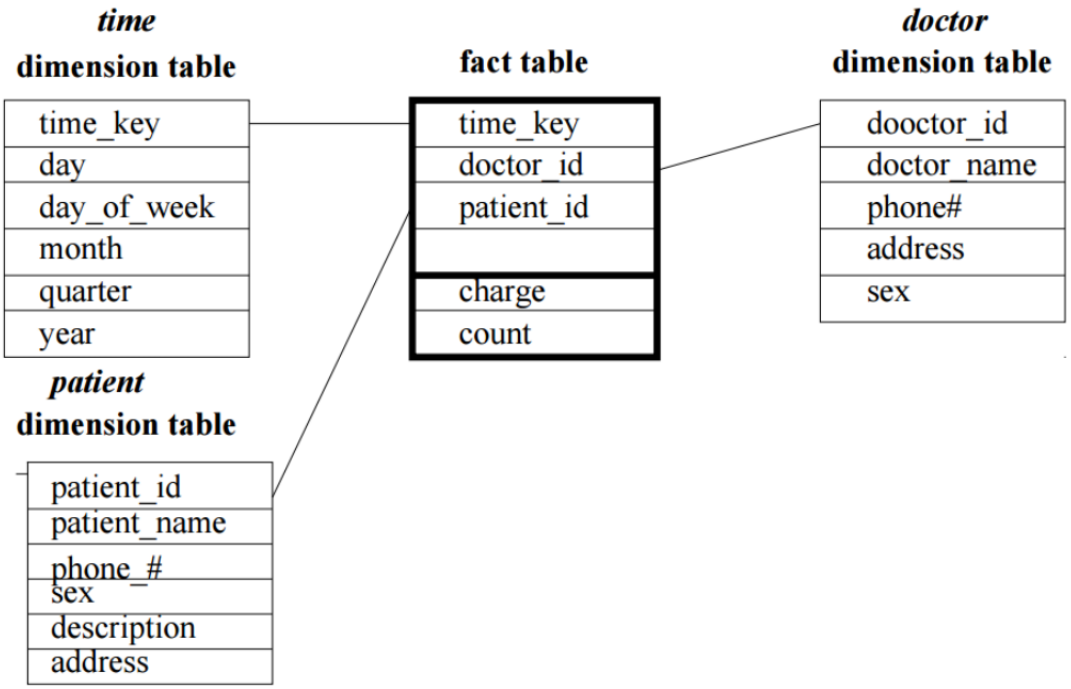

Figure 1: Star schema for data warehouse.

Solution:

1. Roll up on time from day to year.

2. Slice for time = 2010.

3. Roll up on patient from individual patient to all.

4 Question 3 (15 points)

Assume a base cuboid of 10 dimensions contains only three base cells:

- (a1,d2,d3,d4,...,d10),
- (d1,b2,d3,d4,...,d10),
- (d1,d2,c3,d4,...,d10),

where $a1 \neq d1$, $b2 \neq d2$ and $c3 \neq d3$. The measure of the cube is *count*.

1. (5') How many nonempty cuboids will a full data cube contain?
2. (5') How many nonempty aggregate (i.e., nonbase) cells will a full cube contain?
3. (5') How many nonempty aggregate cells will an iceberg cube contain with the condition $count \geq 2$?

Solution:

1. (5') 2^{10}
2. (5') (i) Each cell generates $2^{10} - 1$ nonempty aggregated cells, thus in total we should have $3 * 2^{10} - 3$ cells with overlaps removed. (ii) We have $3 * 2^7$ cells overlapped once (thus count 2) and $1 * 2^7$ (which is $(*, *, *, d4, ..., d10)$) overlapped twice (thus count 3). Thus we should remove in total $1 * 3 * 2^7 + 2 * 1 * 2^7 = 5 * 2^7$ overlapped cells. (iii) Thus we have: $3 * 8 * 2^7 - 5 * 2^7 - 3 = 19 * 2^7 - 3$.
3. (5') (i) $(*, *, d3, d4, ..., d9, d10)$ has count 2 since it is generated by both cell 1 and cell 2; similarly, we have (ii) $(*, d2, *, d4, ..., d9, d10) : 2$, (iii) $(*, *, d3, d4, ..., d9, d10) : 2$; and (iv) $(*, *, *, d4, ..., d9, d10) : 3$. Therefore, we have $4 * 2^7 = 2^9$.

5 Question 4 (20 points)

This question aims to provide you a better understanding of data cube methods.

1. (10') We have a data array containing 3 dimensions A, B and C shown in Figure 2. The 3-D array is divided into 27 small chunks. Each dimension is divided into 3 equally sized partitions. The cardinality (size) of the dimensions A, B, and C is 900, 300, and 600. Since we divide each dimension into 3 parts with equal size, the sizes of the chunks on dimensions A, B, and C are 300, 100, and 200 respectively. Suppose we want to use **Multiway Array Aggregation Computation** to materialize the 2-D cuboids AB, AC and BC.

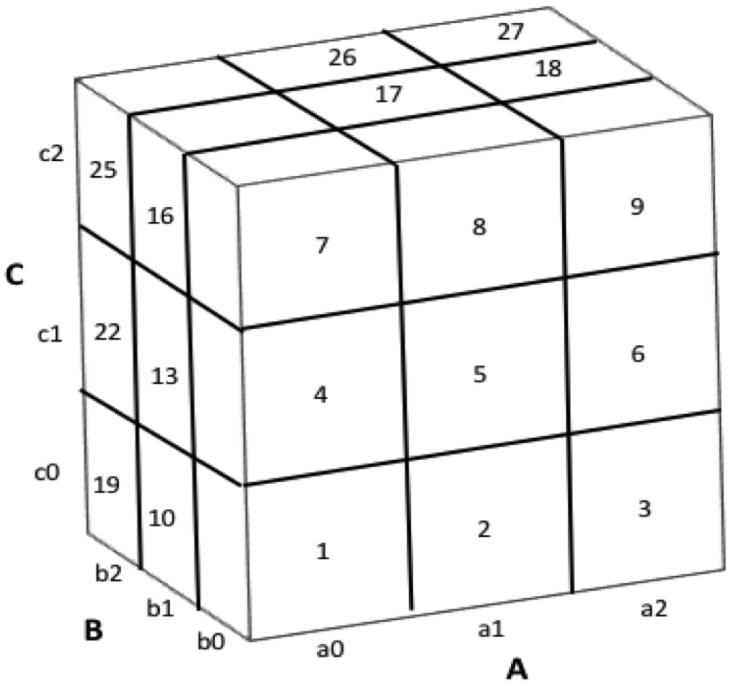

Figure 2: 3-D array of a data cube containing dimensions A, B, C and is divided into 27 small chunks.

(a) (5') What is the ordering of chunk scanning that achieves the maximum computation efficiency, i.e. requires the least memory units?

(b) (5') Following the ordering in part (a), what is the minimum memory requirement for holding all the 2-D planes?

2. Suppose we use **Bottom-Up Computation (BUC)** to materialize a data array containing 3 dimensions A, B and C. The data contained in the array is as follows:

$$\begin{array}{lll} (a_0; b_0; c_0) : 1 & (a_0; b_0; c_1) : 1 & (a_0; b_0; c_2) : 1 \\ (a_0; b_1; c_0) : 2 & (a_0; b_1; c_1) : 1 & (a_0; b_1; c_2) : 1 \\ (a_1; b_2; c_0) : 1 & (a_1; b_2; c_1) : 2 & (a_1; b_2; c_2) : 1 \\ (a_1; b_3; c_0) : 1 & (a_1; b_3; c_1) : 1 & (a_1; b_3; c_2) : 3 \end{array}$$

Now suppose we construct an iceberg cube for dimension A, B, C with the order: C, B, A and the $\boldsymbol{min\_sup} = 4$.

(a) (5') Draw the trace tree of expansion with regard to the given exploration order.

(b) (5') How many cells would be computed given the exploration order and minimum support? Please give detailed explanation.

Solution:

1. (a) $1, 10, 19, 4, 13, 22, 7, 16, 25, 2, 11, 20, 5, 14, 23, 8, 17, 26, 3, 12, 21, 6, 15, 24, 9, 18, 27$.

   (b) $300 \times 200\ (AC) + 300 \times 300\ (AB) + 300 \times 600\ (BC) = 330000$.

2. (a) See Figure 3.

   (b) $1\ (all) + 3\ (C) + 12\ (CB) + 6\ (CA) + 4\ (B) + 3\ (BA) + 2\ (A) = 31$.

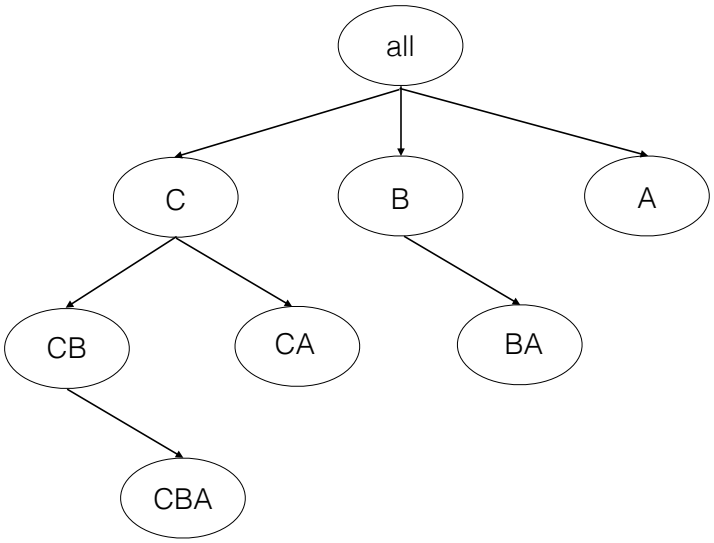

Figure 3: Solution for Question 3: 2(a).