UIUC-CS412 "Introduction to Data Mining" (Spring 2016)

Midterm Exam, Version 1

Friday, Apr. 1, 2016
75 minutes, 90 points + 10 bonus points

Name: NetID:

1 [19]	2 [20]	3 [20]	4 [16]	5 [15]	6 [10] (Bonus)	Total

1. [19] Knowing and Preprocessing Data

   (a) [2'] Assume we have two sets of numerical data having the same **boxplot**. Will the **histograms** also be the same for these two datasets? Briefly explain your answer.
   ANSWER: **No/Not necessarily to be the same. Data distribution may be different even though the 5 number statistics (boxplot) are the same.**

   (b) [2'] In order to find out the hard disk space usage of a computer system, which visualization technique is better: Parallel Coordinates, Chernoff Faces, or Tree-map? Select only **one** and briefly explain your answer.
   ANSWER: **Tree-map. Since the hard disk space usage is a hierarchical data structure, we may need a hierarchical visualization technique. Parallel Coordinates is a geometric projection visualization technique, and Chernoff Faces is an icon-based visualization technique.**

   (c) [5'] You are interested in analyzing data from Data Inc., to identify principal components of two features, but Data Inc. is reluctant to release the data. Instead, the company releases the sample mean and the sample covariance matrices for the two features. The values are:

   $$\mu = [3.5,\ 6]$$

   $$\Lambda = \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix}$$

   As you compute the PCA from the sample covariance and mean data, you realize that Data Inc. has not supplied you with the correct data. How do you figure this out?

ANSWER: **The determinant of the matrix is negative. Hence the matrix cannot represent a covariance matrix since covariance matrixes are positive semi-definite – the determinant is always non-negative, or in other words, the eigenvalues are all non-negative.**

(d) [3'] Give an example when two variables x and y are **dependent** and yet the correlation $r_{x,y} = cov(x, y)/\sigma_x\sigma_y$ is 0.
ANSWER: **Consider two variables x and y, where $y = x^2$ and $x \in [-2, 2]$. Then the correlation is 0.**

(e) [4'] The approximate median is calculated using the formula:

$$L_1 + \left(\frac{n/2 - \sum_l f_l}{f_{median}}\right) \times (L_2 - L_1)$$

where L_1 is the lower interval limit of the bin where the median lies, $\sum_l f_l$ is the sum before the interval, and $(L_2 - L_1)$ is the width of the interval where the median lies.

i. [2'] What is the **minimum** approximation error? Use 2-3 sentences to explain when this case occurs.
ANSWER: **The minimum error is zero. There can be several cases, but the simplest case is when $\sum_{f_l} = n/2$, then the median is exactly given by L_1. The other case is when the data in the interval $L_2 - L_1$ are unique and are distributed evenly; for example, $L_2 - L_1 = f_{median}$.**

ii. [2'] What is the **maximum** approximation error? Use 2-3 sentences to explain when this case occurs.
ANSWER: **The maximum error is $L_2 - L_1$. This happens when $n/2 - \sum_l f_l = f_{median}$ and when all data in the interval are all equal to L_1. In other words, the median is L_1 and the formula gives L_2 as the result.**

(f) [3'] Assume that you have transformed X, a $n \times m$ data matrix into Y, a $n \times m$ matrix using the principal components of X. Why don't we benefit further by computing the principal components of Y?
ANSWER: **We use PCA to align the data which are represented in the canonical coordinate system, with the "natural" axis of the data. The "natural" axes of the data correspond to the directions with maximum variance. Once we transform X into Y, we have aligned the canonical coordinate system with the natural axes of the data in Y. Thus, further PCA decompositions of Y, will simply result in the canonical axes as the principal dimensions, where the data of Y have maximum variance.**

2. [20] Data Warehouse and OLAP

(a) [2'] Is the function $median(\cdot)$ distributive, algebraic or holistic? Briefly explain your answer.
ANSWER: **Holistic. Because there is no constant bound on the storage size needed to describe a sub-aggregate.**

(b) [2'] Name two schemas for modeling data warehouses.
ANSWER: **Star Schema, Snowflake Schema, Fact Constellations or Galaxy Schema.**

(c) [16'] Consider a base cuboid of 10 dimensions that contains 4 base cells:

$$(a_1, a_2, a_3, a_4, a_5, \ldots, a_{10})$$
$$(b_1, a_2, a_3, a_4, a_5, \ldots, a_{10})$$
$$(c_1, c_2, a_3, a_4, a_5, \ldots, a_{10})$$
$$(d_1, d_2, d_3, a_4, a_5, \ldots, a_{10})$$

where a_i, b_i, c_i and d_i are distinct for $i = 1, 2, 3$. There is no dimension with concept hierarchy. The measure of the cube is count. The count of each base cell is 1.

i. [4'] How many cuboids are there in the full data cube?
ANSWER: $2^{10} = 1024$**, since we have 10 dimensions.**

ii. [6'] How many **nonempty aggregated** cells does the complete cube contain? Briefly explain your answer.
ANSWER: **3196. Firstly, we consider those cells with the 1st dimension not aggregated, for example cell** $(D_1, D_2, D_3, *, ..., *)$**. The number of such cells is** $2^9 * 4 = 2048$**. Secondly, we consider those cells with the 1st dimension aggregated but 2nd dimension not aggregated, for example cell** $(*, D_2, D_3, *, ..., *)$**. The number of such cells is** $2^8 * 3 = 768$**. Thirdly, we consider those cells with the 1st and 2nd dimension aggregated but 3rd dimension not aggregated, for example cell** $(*, *, D_3, *, ..., *)$**. The number of such cells is** $2^7 * 2 = 256$**. Finally, we consider those cells with the 1st, 2nd, and 3rd dimension aggregated, for example cell** $(*, *, *, *, ..., *)$**. The number of such cells is** $2^7 * 1 = 128$**. Sum these 4 values and subtract 4 (representing 4 base cells), we get:**$2048 + 768 + 256 + 128 - 4 = 3196$**.**

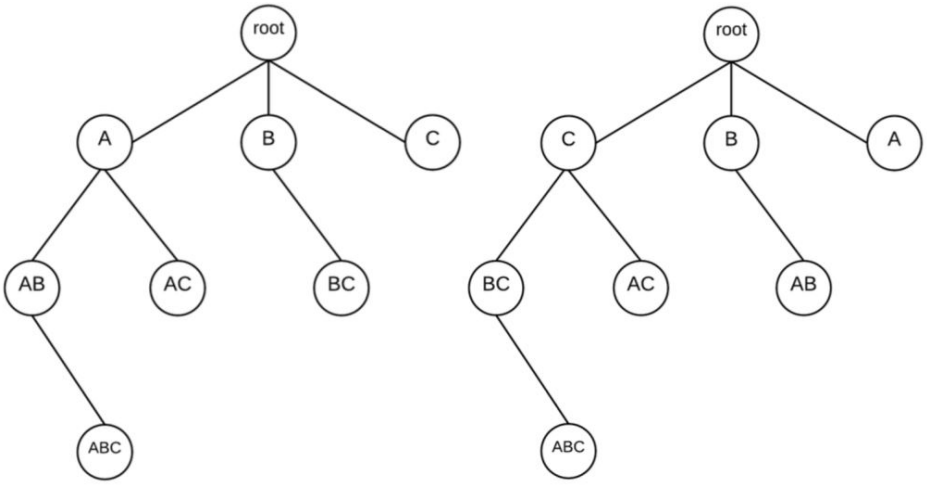
iii. [6'] How many **nonempty aggregated** cells does an iceberg cube contain, if the condition of the iceberg cube is $count \geq 3$? Briefly explain your answer.
ANSWER: **256. Firstly, we note that we must have 1st and 2nd dimension aggregated. Then, if the 3rd dimension is aggregated, we will get** 2^7 **cells whose support is 4. If the 3rd dimension is not aggregated, we will get** 2^7 **cells whose support is 3 and** 2^7 **cells whose support is 1 (these should be neglected). Sum these 2 values, we get:**$2^7 + 2^7 = 2^8 = 256$**.**

3. [20] Data Cube Implementation

(a) [15'] 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:

$(a_0, b_0, c_0) : 1$	$(a_0, b_0, c_1) : 1$	$(a_0, b_0, c_2) : 1$	$(a_0, b_0, c_3) : 1$
$(a_0, b_1, c_0) : 4$	$(a_0, b_1, c_1) : 3$	$(a_0, b_1, c_2) : 2$	$(a_0, b_1, c_3) : 1$
$(a_0, b_2, c_0) : 2$	$(a_0, b_2, c_1) : 2$	$(a_0, b_2, c_2) : 2$	$(a_0, b_2, c_3) : 2$

i. [4'] Draw the trace trees of expansion for the two exploration orders: $A \to B \to C$ and $C \to B \to A$.
ANSWER: **See figures below. Note: For the trace tree of $C \to B \to A$, 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 cuboids. For example, AB and BA are different cuboids 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).

ii. [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.
ANSWER: **See process below.**

All $(*, *, *) : 22$ - expansion

A $(a_0, *, *) : 22$ - expansion

B $(*, b_0, *) : 4$ - pruning

B $(*, b_1, *) : 10$ - expansion

B $(*, b_2, *)$: 8 - expansion

C $(*, *, c_0)$: 7
C $(*, *, c_1)$: 6
C $(*, *, c_2)$: 5
C $(*, *, c_3)$: 4

AB $(a_0, b_0, *)$: 4 - pruning
AB $(a_0, b_1, *)$: 10 - expansion
AB $(a_0, b_2, *)$: 8 - expansion

AC $(a_0, *, c_0)$: 7
AC $(a_0, *, c_1)$: 6
AC $(a_0, *, c_2)$: 5
AC $(a_1, *, c_3)$: 4

BC $(*, b_1, c_0)$: 4
BC $(*, b_1, c_1)$: 3
BC $(*, b_1, c_2)$: 2
BC $(*, b_1, c_3)$: 1
BC $(*, b_2, c_0)$: 2
BC $(*, b_2, c_1)$: 2
BC $(*, b_2, c_2)$: 2
BC $(*, b_2, c_3)$: 2

ABC (a_0, b_1, c_0) : 4
ABC (a_0, b_1, c_1) : 3
ABC (a_0, b_1, c_2) : 2
ABC (a_0, b_1, c_3) : 1
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

iii. [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.

1. [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.

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

(b) [5'] John is a data analyst working on his company sales database and is using the `Apriori` algorithm. He decides to cut down the number of dimensions he is working with by merging the `cost` and `sale price` dimensions by defining a new measure called `profit`, where profit = sale price−cost. When he runs the `Apriori` algorithm, something is not right; what could be the problem?
ANSWER: **Unlike measures like `count`, the measure `profit` is not guaranteed to be non-negative, since companies will often sell products at a loss to drive sales or to bring customers to the store. Thus the measure `profit` will not satisfy the `Apriori` property and cannot be used to efficiently create an iceberg cube.**

4. [16] Frequent Pattern and Association Mining

(a) [12'] A database with 250 transactions has its FP-tree shown as follows. Let $min\_sup = 0.15$ and $min\_conf = 0.4$.

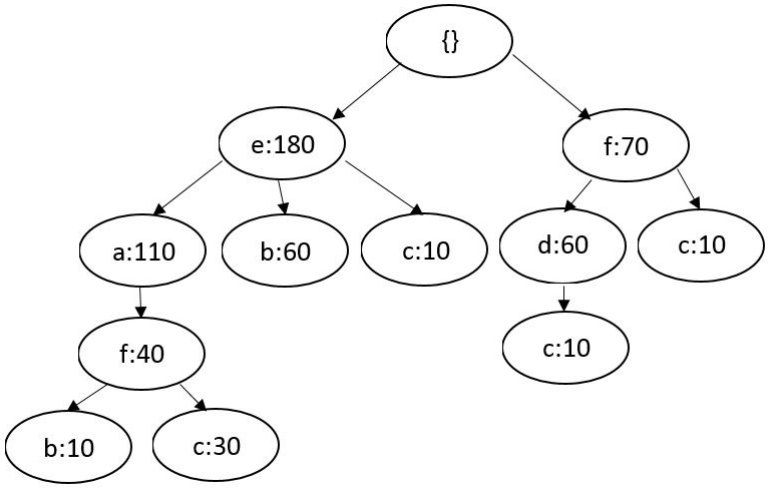
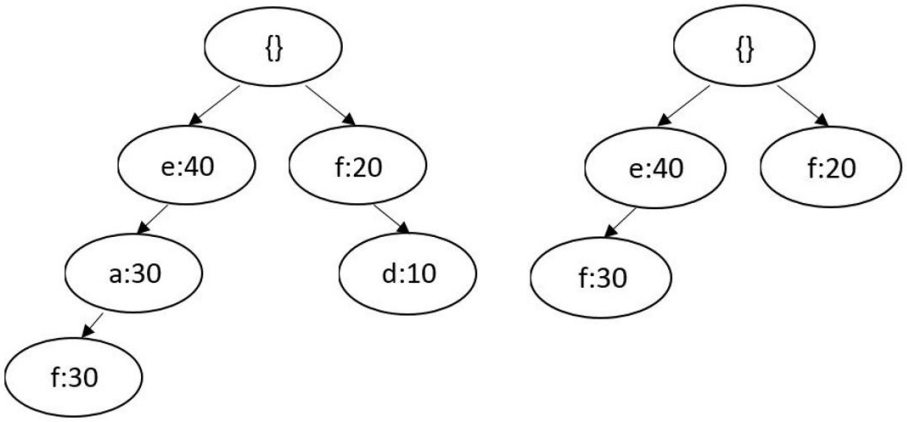


Figure 1: FP tree of a transaction DB

i. [4'] Show c's conditional (i.e, projected) pattern base.
ANSWER: $eaf : 30, e : 10, fd : 10, f : 10$**.**

ii. [4'] Draw c's conditional FP-tree.
ANSWER: **See figures below. Note that node** a **and** d **can be omitted**



because they can be pruned due to min support.

iii. [4'] Compute the support and confidence of the following rules and decide if they are association rules or not accordingly.

1. [2'] $f \rightarrow c$.
   ANSWER: **Support:** $50/250 = 0.2$**; Confidence:** $50/110 = 0.4545$**, Yes.**

2. [2'] $ac \rightarrow e$.
   ANSWER: **Support:** $30/250 = 0.12$**; Confidence:** $30/30 = 1$**, No.**

(b) [2'] Why is the running time of `FP-Growth` and `Apriori` comparable for relative support thresholds of say even 2%, on large datasets?
ANSWER: **For large datasets, say 10M records, a relative support threshold of** 2% **is enormous! The number of itemsets satisfying this threshold will drop dramatically, reducing the number of itemsets that a level based algorithms which uses** `Apriori` **has to use.**

(c) [2'] What is the key difference between the $Lift(A, B)$ measure and the $cosine(A, B)$ measure, where A, B are two itemsets?
ANSWER: $Lift(A, B)$ **is not null-invariant, while** $cosine(A, B)$ **measure is null-invariant.**

5. [15] Advanced Frequent Pattern and Association Mining

(a) [9'] The following table lists some pattern space pruning constraints (S represents an itemset and v is a value). Please determine the type of these constraints and complete the following table by filling out each space in the table with **yes**, **no**, or **convertible**.

Constraint	Anti-monotone	Monotone	Succinct
$min(S) \geq v$	**Yes**	**No**	**Yes**
$sum(S) \geq v (a \in S, a \geq 0)$	**No**	**Yes**	**No**
$avg(S)\, \theta\, v, \theta \in \{=, \leq, \geq\}$	**Convertible**	**Convertible**	**No**

(b) [3'] What is pattern succinctness and how is it different from data succinctness ? Show one example of each.
ANSWER: **pattern succinctness refers the ability to trivially enumerate all the itemsets using just one data element – in other words, a single item can cause the pattern to be true; thus the pattern** $\min \leq v$ **is pattern succinct, since all one has to do is to enumerate all items** S **of value less than** v**; all the itemsets that contain any element from** S **will satisfy the pattern.**
Data succinctness refers to the ability of an item to determine the set of interest; for example, if the product dimension is "TV", then the only itemsets of interest are those that contain "TV".

(c) [3'] How is the sequential pattern mining problem different from older association rule mining problems?
ANSWER: **In sequential pattern mining of itemsets, the order of the itemsets matters, whereas in the older association mining rule problem, the order of items in the itemsets does not mater. In the sequential itemset mining, the order *within* an itemset in the sequence does not mater.**

6. [10] Bonus Question

The Mean Absolute Deviation (MAD) measure of central tendency is considered to be a more robust measure than standard deviation. Prove that $\mathbb{E}(MAD(x)) \leq \sigma$, where σ is the population standard deviation and where $\mathbb{E}$ is the expectation operator. Hint: $f(\mathbb{E}(X)) \leq \mathbb{E}(f(x))$, when f is a convex function.

ANSWER: **Use $f(x) = x^2$ and the result follows trivially.**

$$\begin{aligned}(\mathbb{E}(|X-\mu|))^2 &\leq \mathbb{E}(|X-\mu|)^2 \\ &\leq \mathbb{E}(X-\mu)^2 \\ &\leq \sigma^2 \\ \therefore \mathbb{E}(|X-\mu|) &\leq \sigma\end{aligned}$$