**UIUC-CS412 "Introduction to Data Mining" (Summer 2016)**

# Midterm Exam

Monday, July 11, 2016
**75 minutes, 75 points**

Name: NetID:

| 1 [7] | 2 [15] | 3 [20] | 4 [10] | 5 [20] | Total [75] |
|---|---|---|---|---|---|
| | | | | | |

**Instructions**

(a) Make sure your exam has **10 pages**.

(b) Please keep your ID cards on the desk.

1. Short Answer Questions [7 points].

These questions should be answered in **not more than 1-2 lines**. Be as specific as possible.

(a) [2] What are the value ranges for the following measures? (No explanation required)

(i) Jaccard coefficent
**Answer:** $[0, 1]$

(ii) Correlation
**Answer:** $[-1, 1]$

(b) [2] What are the best distance measures for each of the following applications? (No explanation required)

(i) Find whether two text documents are similar
**Answer:** Cosine similarity using word vectors. Or even levenshtein distance.

(ii) Find the maximum difference between any attribute of two vectors
**Answer:** $L_{max}$ distance

(c) [1] For Multiway Array Aggregation cube implementation method it is important to choose the order of scanning data chunks.
**Answer:** True

(d) [2] Which method is better for computing iceberg cubes, Multiway Array Aggregation or BUC? Explain why. (**1-2 lines of explanation required**)
**Answer:** BUC uses top down approach which allows us to prune the search when the minimum support condition fails. Whereas Multiway aggregation is a full cube computation.

2. Knowing and preprocessing data [15 points].

Consider 4 data points (population) in a 2-D space: (2,0), (0,-1), (4,1), and (6,2). (For all sub-questions below, correct answers without explanations receive full points; and incorrect answers with explanations may receive partial credit.)

(a) [4] Calculate the covariance matrix.
**Answer:** In order to find the covariance matrix we first mean normalize the data by subtracting the mean of each dimension.

$$\begin{bmatrix} 2 & 0 \\ 0 & -1 \\ 4 & 1 \\ 6 & 2 \end{bmatrix} - \begin{bmatrix} 3 & 0.5 \\ 3 & 0.5 \\ 3 & 0.5 \\ 3 & 0.5 \end{bmatrix} = \begin{bmatrix} -1 & -0.5 \\ -3 & -1.5 \\ 1 & 0.5 \\ 3 & 1.5 \end{bmatrix}$$

You can find the covariance matrix using the formula $\frac{1}{N} * A^T * A$

$$\sum = \frac{1}{4} * \begin{bmatrix} -1 & -3 & 1 & 3 \\ -0.5 & -1.5 & 0.5 & 1.5 \end{bmatrix} * \begin{bmatrix} -1 & -0.5 \\ -3 & -1.5 \\ 1 & 0.5 \\ 3 & 1.5 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 2.5 \\ 2.5 & 1.25 \end{bmatrix}$$

(b) [4] Calculate the pearson correlation coefficient for the two dimensions.

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x * \sigma_y} = 1.0$$

It is also evident from the graph plot of these points since they all lie in the same straight line.

(c) [4] Calculate the first and the second principal components (two vectors), and indicate which is the first principal component.
**Answer:** To find the principal components you first need to solve the following equation:

$$det(\sum -\lambda I) = 0$$

$$\begin{vmatrix} 5-\lambda & 2.5 \\ 2.5 & 1.25-\lambda \end{vmatrix} = 0$$

which gives two eigen vectors : $\lambda_1 = 6.25$, $\lambda_2 = 0$

Also,

$$[\sum -\lambda I] * \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

and for principal components

$$\sqrt{{v_1}^2 + {v_2}^2} = 1$$

On substituting the values of $\lambda$ we get the first component as $(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}})$ and the second component as $(\frac{1}{\sqrt{5}}, \frac{-2}{\sqrt{5}})$

(d) [2] What are the coordinates of the 4 data points, *projected* to the 1-D space corresponding to the first principal component?
**Answer:** To find the projected 1D points you should use the mean normalized data points and perform matrix multiplication with the first eigen vector as shown below

$$\begin{bmatrix} -1 & -0.5 \\ -3 & -1.5 \\ 1 & 0.5 \\ 3 & 1.5 \end{bmatrix} * \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} \frac{-2.5}{\sqrt{5}} \\ \frac{-7.5}{\sqrt{5}} \\ \frac{2.5}{\sqrt{5}} \\ \frac{7.5}{\sqrt{5}} \end{bmatrix}$$

(e) [1] Is the projection of data points to the 1-D space in the previous sub-question a lossless or lossy compression? Briefly explain.
**Answer:** It is a fully lossless compression as all the data points lie on the same line as the principal eigen vector direction.

3. Data Warehousing and OLAP for Data Mining [20'].
Consider a base cuboid of 4 dimensions that contains 3 base cells:

$$(a_1, a_2, a_3, a_4)$$
$$(a_1, b_2, a_3, b_4)$$
$$(c_1, a_2, c_3, a_4)$$

where $a_i$, $b_i$, and $c_i$ are distinct for $i = 1, 2, 3, 4$. 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:**
$(1 + 1)^4$

(b) [8'] How many **nonempty aggregated** (i.e, *count* $\geq 1$ and non-base) cells does the complete cube contain? Briefly explain your answer.

**Answer:**
2 points for correct answer
2 points for removing base cases
2 points for accounting for duplicates
2 points for method / closeness to correct answer

There are many ways to solve this problem. Here is one of method. We first consider the first base cell and the second base cell. The dimensions can be rearranged for readability without affecting the final outcome. The first and third dimensions have two options, $a$ and $*$. The second and fourth dimensions have three options $a$, $b$, and $*$.

$$\begin{pmatrix} a_1 & a_3 \\ * & * \end{pmatrix} \begin{pmatrix} a_2 & a_4 \\ b_2 & b_4 \\ * & * \end{pmatrix}$$

The total number of possible combinations from the first and second base cell is $2^2 \times 3^2$. Now we need to subtract combinations that has *count* $= 0$. We know that the count will be zero when there is only one $b$ present in both the second and forth dimension. We subtract these two cases from the right matrix ($(a_2, b_4)$ and $(b_2, a_4)$). Therefore the number of nonempty aggregate cells is $2^2 \times (3^2 - 2)$.

Now we consider the first base cell and the third base cell. Similarly, the dimensions can be rearranged for readability without affecting the final outcome. The second and fourth dimensions have two options, $a$ and $*$. The first and third dimensions have three options $a$,

$c$, and $*$.

$$\begin{pmatrix} a_2 & a_4 \\ * & * \end{pmatrix} \begin{pmatrix} a_1 & a_3 \\ c_1 & c_3 \\ * & * \end{pmatrix}$$

The total number of nonempty aggregate cells from the first and third base cell is $2^2 \times (3^2 - 2)$.

Now we need to subtract the redundant cells from the above calculations. The above calculations both include combinations from only $a$ and $*$.

$$\begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ * & * & * & * \end{pmatrix}$$

We remove redundancy by subtracting the above combination of cells $2^4$.

We do not need to consider combination of the second and third base cell as the only combination that produces a nonempty aggregate cell is $(*, *, *, *)$ which is already covered. But we do have to remove the 3 base cases.

So the final answer is $2^2 \times (3^2 - 2) + 2^2 \times (3^2 - 2) - 2^4 - 3 = 37$

(c) [4'] List all the **nonempty aggregated closed** cells the full cube will contain.

**Answer:**
Three. $(*, *, *, *)$, $(a_1, *, a_3, *)$, and $(*, a_2, *, a_4)$

(d) [4'] List all the **nonempty aggregate** cells an iceberg cube will contain if the condition of the iceberg cube is $count \geq 2$.

**Answer:**
Seven. $(*, *, *, *)$, $(a_1, *, *, *)$, $(*, a_2, *, *)$, $(*, *, a_3, *)$, $(*, *, *, a_4)$, $(a_1, *, a_3, *)$, and $(*, a_2, *, a_4)$

4. BUC algorithm. [10'].
We have 3-D data array with 3 dimensions $A$, $B$, and $C$. The data contained in the array is as follows:

| | | |
|---|---|---|
| $(a_0, b_0, c_0) : 1$ | $(a_0, b_1, c_0) : 1$ | $(a_0, b_2, c_0) : 1$ |
| $(a_0, b_0, c_1) : 1$ | $(a_0, b_1, c_1) : 1$ | $(a_0, b_2, c_1) : 1$ |
| $(a_0, b_0, c_2) : 1$ | $(a_0, b_1, c_2) : 1$ | $(a_0, b_2, c_2) : 1$ |

(a) [5'] If we set *min_support* = *5* with the order of $C \rightarrow B \rightarrow A$, how many cells would be considered/computed?

**Answer:**
8 cells are considered/computed

All $(*, *, *)$ : 18 - expansion

---

C $(*, *, c_0)$ : 3
C $(*, *, c_1)$ : 3
C $(*, *, c_2)$ : 3

---

B $(*, b_0, *)$ : 3
B $(*, b_1, *)$ : 3
B $(*, b_2, *)$ : 3

---

A $(a_0, *, *)$ : 9

(b) [5'] If we set *min_support = 5* with the order of $A \rightarrow B \rightarrow C$, how many cells would be considered/computed?

**Answer:**
14 cells are considered/computed

All $(*, *, *)$ : 9 - expansion

---

A $(a_0, *, *)$ : 9 - expansion

---

AB $(a_0, b_0, *)$ : 3
AB $(a_0, b_1, *)$ : 3
AB $(a_0, b_2, *)$ : 3

---

AC $(a_0, *, c_0)$ : 3
AC $(a_0, *, c_1)$ : 3
AC $(a_0, *, c_2)$ : 3

---

B $(*, b_0, *)$ : 3
B $(*, b_1, *)$ : 3
B $(*, b_2, *)$ : 3

---

C $(*, *, c_0)$ : 3
C $(*, *, c_1)$ : 3
C $(*, *, c_2)$ : 3

5. Frequent pattern and association mining [20'].
Based on the transactions in Table 1, answer the following questions.

| **Transaction Number** | **Items bought** | **Ordered frequent items** |
|---|---|---|
| 1 | p, f, n, s, a, j, k, m | |
| 2 | p, f, n, a, c, z | |
| 3 | f, s, a, q | |
| 4 | p, m, c, r | |
| 5 | p, s, k, r | |
| 6 | p, f, n, s, j, q, x | |
| 7 | p, f, n, u, o | |

Table 1: Transactions records

(a) [4'] Fill out the **Ordered frequent items** column in Table 1. *min_support = 3*

**Answer:**
1. (p, f, n, s, a)
2. (p, f, n, a)
3. (f, s, a)
4. (p)
5. (p, s)
6. (p, f, n, s)
7. (p, f, n)

(b) [4'] Draw a FP-tree based on the frequent item list from (a). If there is a tie, use the alphabetical order as a tie breaker.

**Answer:**

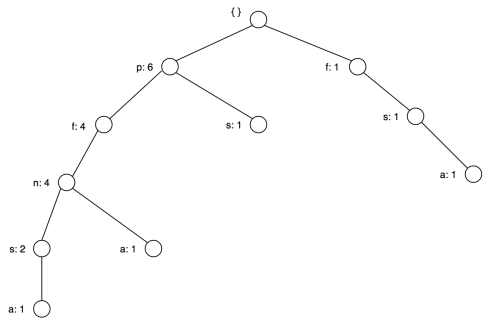


Figure 1: FP-tree of a transaction DB.

(c) [4'] Show $s$'s conditional (i.e., projected) databases.

**Answer:**
{p, f, n: 2}, {p: 1}, {f: 1},

(d) [4'] Show all the frequent $k$-itemsets for the **largest** $k$. *min_support = 3*
Hint: Look at the FP-tree

**Answer:**
k = 3. {p, f, n: 4}

(e) [4'] List **all** association rules with $min\_conf = 0.5$ for $\{p, f, n : 4\}$.

**Answer:**
$\{p, f, n\} \rightarrow a$
$\{p, f, n\} \rightarrow s$

Or

$\{p\} \rightarrow f$
$\{p\} \rightarrow n$
$\{f\} \rightarrow p$
$\{f\} \rightarrow n$
$\{n\} \rightarrow f$
$\{n\} \rightarrow p$
$\{p, f\} \rightarrow n$
$\{p, n\} \rightarrow f$
$\{f, n\} \rightarrow p$
$\{p\} \rightarrow \{f, n\}$
$\{f\} \rightarrow \{p, n\}$
$\{n\} \rightarrow \{p, f\}$