Midterm

- There are 15 problems in this exam, worth 45 points in total.
- You must not communicate with other students during this test.
- Books or notes are not allowed.
- Electronic devices except calculators are not allowed.
- You cannot use your mobile phone as calculator.
- The duration of this exam is 2 hrs.
- For each question, you will NOT get full credit if you only give out a final result. Necessary calculation steps and reasoning are required.
- Wait for instructions before turning this page.

1. Fill in your information:

Full Name: ________________________________

NetID: ________________________________

1/1. (2 points) What are the value ranges of the following correlation measures, respectively?

- χ^2
- Pearson correlation coefficient

Solution.

- χ^2: $[0, \infty\}$
- Pearson correlation coefficient: $[-1, 1]$

1/2. (2 points) What are the value ranges of the following correlation measures, respectively?

- χ^2
- Pearson correlation coefficient

Solution.

- χ^2: $[0, \infty\}$
- Pearson correlation coefficient: $[-1, 1]$

2/1. (3 points) The following table contains the medical record of two patients (John and Jane) on six lab tests. Calculate the similarity between John and Jane based on these records (P & N mean positive and negative test results, respectively). Write down the steps.

	test-1	test-2	test-3	test-4	test-5	test-6
John	P	N	P	N	N	N
Jane	P	N	P	N	P	N

Solution. these are assyimetric binary attributes, so we need to draw the contingency table and calcluate q, r, and s.

$$\frac{2}{2+0+1} = 0.66$$

2/2. (3 points) The following table contains the medical record of two patients (John and Jane) on six lab tests. Calculate the similarity between John and Jane based on these records (P & N mean positive and negative test results, respectively). Write down the steps.

	test-1	test-2	test-3	test-4	test-5	test-6
John	P	N	P	N	N	N
Jane	P	N	P	N	P	N

Solution. these are assyimetric binary attributes, so we need to draw the contingency table and calcluate q, r, and s. $\frac{2}{2+0+1} = 0.66$

3/1. (3 points) Suppose that the values for a given set of grades in a class are grouped into intervals. The intervals and corresponding frequencies are as follows. Compute an approximate median value for the data. Write down the steps.

grade	frequency
1-25	5
25-50	10
50-65	20
65-75	35
75-85	20
85-100	10

Solution. N = 100

$\frac{N}{2} = 50$

grade	frequency	cumulative frequency
1-25	5	5
25-50	10	15
50-65	20	35
65-75	35	70
75-85	20	90
85-100	10	100

median interval: [65-75]

median $= L1 + (\frac{\frac{N}{2} - (\Sigma freq)_l}{freq_{median}})width$

median $= 65 + (\frac{50 - 35}{20})10 = 72.5$

3/2. (3 points) Suppose that the values for a given set of grades in a class are grouped into intervals. The intervals and corresponding frequencies are as follows. Compute an approximate median value for the data. Write down the steps.

grade	frequency
1-25	5
25-50	10
50-65	20
65-75	35
75-85	20
85-100	10

Solution. N = 100

$\frac{N}{2} = 50$

grade	frequency	cumulative frequency
1-25	5	5
25-50	10	15
50-65	20	35
65-75	35	70
75-85	20	90
85-100	10	100

median interval: [65-75]

$$\text{median} = L1 + (\frac{\frac{N}{2} - (\Sigma freq)_l}{freq_{median}})width$$

$$\text{median} = 65 + (\frac{50 - 35}{20})10 = 72.5$$

4/1. (2 points) Suppose a cube has 20 dimensions: 15 dimensions with 1 level, 3 dimensions with 2 levels, and 2 dimensions with 3 levels (not including all). How many cuboids does this cube contain (including base and apex cuboids)? You don't need to calculate the final number.

Solution. $2^{15} \times 3^3 \times 4^2$

4/2. (2 points) Suppose a cube has 20 dimensions: 15 dimensions with 1 level, 3 dimensions with 2 levels, and 2 dimensions with 3 levels (not including all). How many cuboids does this cube contain (including base and apex cuboids)? You don't need to calculate the final number.

Solution. $2^{15} \times 3^3 \times 4^2$

5/1. (3 points) For the datacube shown in the following table, return the result of the following query: $< a_2, b_1, ?, *, ? : count >$.

Table 1: Datacube

A	B	C	D	E
a_1	b_1	c_1	d_1	e_1
a_1	b_2	c_1	d_2	e_1
a_1	b_2	c_1	d_1	e_2
a_2	b_1	c_1	d_1	e_2
a_2	b_1	c_1	d_1	e_3
a_2	b_1	c_1	d_2	e_2

Solution. $< a_2, b_1, c_1, *, e_2 : 2 >$, $< a_2, b_1, c_1, *, e_3 : 1 >$

5/2. (3 points) For the datacube shown in the following table, return the result of the following query: $< a_2, b_1, ?, *, ? : count >$.

Table 2: Datacube

A	B	C	D	E
a_1	b_1	c_1	d_1	e_1
a_1	b_2	c_1	d_2	e_1
a_1	b_2	c_1	d_1	e_2
a_2	b_1	c_1	d_1	e_2
a_2	b_1	c_1	d_1	e_3
a_2	b_1	c_1	d_2	e_2

Solution. $< a_2, b_1, c_1, *, e_2 : 2 >$, $< a_2, b_1, c_1, *, e_3 : 1 >$

6/1. (2 points) At what conditions that a cube is more efficient to compute using BUC than using multiway array aggregation?

Solution. BUC is more efficient for sparse and iceberg cubes. Key idea: pruning sparse high-dimensional cubes.

6/2. (2 points) At what conditions that a cube is more efficient to compute using BUC than using multiway array aggregation?

Solution. BUC is more efficient for sparse and iceberg cubes. Key idea: pruning sparse high-dimensional cubes.

7/1. (4 points) Below is the contingency table that summarizes a supermarket transaction data of hot dogs and hamburgers sell.

	$hotdogs$	$\overline{hotdogs}$	$\sum_{col}$
$hamburgers$	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
$\sum_{row}$	3000	2000	5000

A. Suppose that the association rule " hot dogs $\Rightarrow$ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50% is this association rule strong?

B. Calculate two measures of correlation: *maximum confidence* and *Kulczynski* on the above given data.
(Hint: $max\_conf(A, B) = max\{P(A|B), P(B|A)\}$, $\quad Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A))$.)

C. What is the main difference between *maximum confidence* and *Kulczynski* measure?

Solution.

A. Support $= \frac{2000}{5000} = 40\%$ and Confidence $= \frac{2000}{3000} = 66.7\%$. Hence the rule is strong.

B. $max\_conf(A, B) = max\{\frac{2000}{3000}, \frac{2000}{2500}\} = 0.8$, $Kulc(A, B) = 0.735$

C. Imbalance Ratio

7/2. (4 points) Below is the contingency table that summarizes a supermarket transaction data of hot dogs and hamburgers sell.

	$hotdogs$	$\overline{hotdogs}$	$\sum_{col}$
$hamburgers$	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
$\sum_{row}$	3000	2000	5000

A. Suppose that the association rule " hot dogs $\Rightarrow$ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50% is this association rule strong?

B. Calculate two measures of correlation: *maximum confidence* and *Kulczynski* on the above given data.
(Hint: $max\_conf(A, B) = max\{P(A|B), P(B|A)\}$, $\quad Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A))$.)

C. What is the main difference between *maximum confidence* and *Kulczynski* measure?

Solution.

A. Support $= \frac{2000}{5000} = 40\%$ and Confidence $= \frac{2000}{3000} = 66.7\%$. Hence the rule is strong.

B. $max\_conf(A, B) = max\{\frac{2000}{3000}, \frac{2000}{2500}\} = 0.8$, $Kulc(A, B) = 0.735$

C. Imbalance Ratio

8/1. (4 points) Suppose a WalMart manager is interested in only the frequent patterns (i.e., itemsets) that satisfy certain constraints. For the following cases, state the characteristics (i.e., categories) of every constraint in each case.

A. The sum of the price of all the items (in each pattern) is between \$100 and \$200.

B. The average price of all the items in each pattern is less than \$10

Solution.

A. $C1 : sum(S.price) \geq \$100$ is montone. $C2 : sum(S.price) \leq \$200$ is antimontone.

B. $C : avg(S.price) < \$10$ is convertible anti-monotone constraint if the items in each transaction are sorted in price ascending order.

8/2. (4 points) Suppose a WalMart manager is interested in only the frequent patterns (i.e., itemsets) that satisfy certain constraints. For the following cases, state the characteristics (i.e., categories) of every constraint in each case.

A. The sum of the price of all the items (in each pattern) is between \$100 and \$200.

B. The average price of all the items in each pattern is less than \$10

Solution.

A. $C1 : sum(S.price) \geq \$100$ is montone. $C2 : sum(S.price) \leq \$200$ is antimontone.

B. $C : avg(S.price) < \$10$ is convertible anti-monotone constraint if the items in each transaction are sorted in price ascending order.

9/1. (4 points) Consider the following transaction database:

TransactionID	Items
T1	M,O,N,K,E,Y
T2	D,O,N,K,E,Y
T3	M,A,K,E
T4	M,U,C,K,Y
T5	C,O,O,K,I,E

Suppose that minimum support is set to 60% and minimum confidence to 80%.

A. List all frequent item sets of size 1

B. List all frequent item sets of size 2

C. Which among the above frequent item sets are closed.

Solution.

A. $L_1 = \{E : 4, K : 5, M : 3, O : 3, Y : 3\}$

B. $L_2 = \{EK : 4, EO : 3, KM : 3, KO : 3, KY : 3\}$

C. EK , KM, KY, K

9/2. (4 points) Consider the following transaction database:

TransactionID	Items
T1	M,O,N,K,E,Y
T2	D,O,N,K,E,Y
T3	M,A,K,E
T4	M,U,C,K,Y
T5	C,O,O,K,I,E

Suppose that minimum support is set to 60% and minimum confidence to 80%.

A. List all frequent item sets of size 1

B. List all frequent item sets of size 2

C. Which among the above frequent item sets are closed.

Solution.

A. $L_1 = \{E : 4, K : 5, M : 3, O : 3, Y : 3\}$

B. $L_2 = \{EK : 4, EO : 3, KM : 3, KO : 3, KY : 3\}$

C. EK , KM, KY, K

10/1. (3 points)

Identify all the core patterns for the following transaction dataset where the core ratio (τ) is 0.5.

transaction	# of transactions
(acb)	100
(ac)	100
(acf)	100

Solution.

(acb), (ab), (b), (acf), (af), (cf), (f)

10/2. (3 points) Identify all the core patterns for the following transaction dataset where the core ratio (τ) is 0.5.

transaction	# of transactions
(acd)	100
(ac)	100
(ace)	100

Solution.

(acd), (ad), (d), (ace), (ae), (ce), (e)

11/1. (3 points)

An enterprise is interested in obtaining large-sized patterns that are closed and maximal in its transaction dataset. It is ensured that the number of large sized patterns are very few in number. The enterprise implemented Apriori and FP growth pattern mining algorithms to discover the large-sized patterns, but the algorithm failed to discover the patterns in polynomial time. Explain with reason a condition under which such a problem can happen. Suggest a possible algorithm taught in class that can address this problem.

Solution. Incrementally increasing the size of the patterns runs into the problem of **exponentially large number of mid-sized patterns**. This hinders the discovery of large sized patterns in the dataset. [2 points]

Pattern Fusion, or any colossal pattern mining algorithm description. Even approximation pattern mining algos are ok. [1 point]

11/2. (3 points)

An enterprise is interested in obtaining large-sized patterns that are closed and maximal in its transaction dataset. It is ensured that the number of large sized patterns are very few in number. The enterprise implemented Apriori and FP growth pattern mining algorithms to discover the large-sized patterns, but the algorithm failed to discover the patterns in polynomial time. Explain with reason a condition under which such a problem can happen. Suggest a possible algorithm taught in class that can address this problem.

Solution. Incrementally increasing the size of the patterns runs into the problem of **exponentially large number of mid-sized patterns**. This hinders the discovery of large sized patterns in the dataset. [2 points]

Pattern Fusion, or any colossal pattern mining algorithm description. Even approximation pattern mining algos are ok. [1 point]

12/1. (3 points)

Show the FP tree construction for the following transaction dataset. Also generate an ordered list of frequent items based on the transaction database when the minimum support is 0.5.

transaction id	# items
0	a, b
1	b, c, d
2	a, c, d, e
3	a, d, e
4	a, b, c
5	b, c, e

Solution.
(a:4, b:4, c:4, d:3, e:3) or fractional form.

12/2. (3 points)

Show the FP tree construction for the following transaction dataset. Also generate an ordered list of frequent items based on the transaction database when the minimum support is 0.5.

transaction id	# items
0	a, b
1	b, c, d
2	a, c, d, e
3	a, d, e
4	a, b, c
5	b, c, e

Solution.
(a:4, b:4, c:4, d:3, e:3) or fractional form.

13/1. (2 points) Suppose the median of n observations $x_1, x_2......x_n$ as

$$Median = \begin{cases} x_{(n+1)/2} & \text{: n is odd} \\ x_{((n+1)/2)+1} & \text{: n is even} \end{cases}$$

Is this measure a distributive, algebraic or holistic measure? Justify your answer.

Solution. Holistic

13/2. (2 points) Suppose the median of n observations $x_1, x_2......x_n$ as

$$Median = \begin{cases} x_{(n+1)/2} & \text{: n is odd} \\ x_{((n+1)/2)+1} & \text{: n is even} \end{cases}$$

Is this measure a distributive, algebraic or holistic measure? Justify your answer.

Solution. Holistic

14/1. (4 points) We have a data array containing 3 dimensions A, B and C. The 3-D array is divided into small chunks. Each dimension is divided into 3 equally sized partitions. See Figure 1. For example, dimension A is divided into a_0, a_1 and a_2, and dimension B is divided into b_0, b_1 and b_2. There are totally 27 chunks and each chunk is represented by a sub-cube $a_i b_j c_k$. The cardinality (size) of the dimensions A, B, and C is 750, 300, and 450. Since we divide each dimension into 3 parts with equal size, the sizes of the chunks on dimensions A, B, and C are 250, 100, and 150 respectively. Now we want to use **Multiway Array Aggregation** Computation to materialize cubes. The base cuboid ABC is computed as a 3-D array. We want to materialize the 2-D cuboids AB, AC and BC.

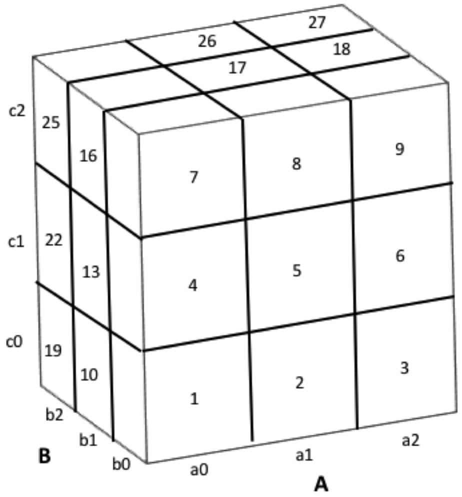

Figure 1: Figure 1: A 3-D array with dimensions A, B and C. This array is divided into 27 smaller chunks.

If we scan the chunk in the order $1, 2, 3, ...27$ when materializing the 2-D cuboids AB, AC and BC, to avoid reading a 3-D chunk into memory repeatedly, what is the minimum memory requirement to hold all the related 2-D planes?

Solution. Space requirement = AC + AB + BC = 750*450 + 100*750 + 100*150 = 427,500

14/2. (4 points) We have a data array containing 3 dimensions A, B and C. The 3-D array is divided into small chunks. Each dimension is divided into 3 equally sized partitions. See Figure 1. For example, dimension A is divided into a_0, a_1 and a_2, and dimension B is divided into b_0, b_1 and b_2. There are totally 27 chunks and each chunk is represented by a sub-cube $a_i b_j c_k$. The cardinality (size) of the dimensions A, B, and C is 750, 300, and 450. Since we divide each dimension into 3 parts with equal size, the sizes of the chunks on dimensions A, B, and C are 250, 100, and 150 respectively. Now we want to use **Multiway Array Aggregation** Computation to materialize cubes. The base cuboid ABC is computed as a 3-D array. We want to materialize the 2-D cuboids AB, AC and BC.

If we scan the chunk in the order $1, 2, 3, ...27$ when materializing the 2-D cuboids AB, AC and BC, to avoid reading a 3-D chunk into memory repeatedly, what is the minimum memory requirement to hold all the related 2-D planes?

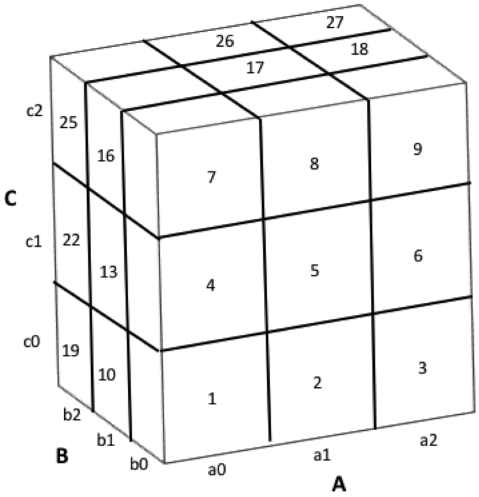

Figure 2: Figure 1: A 3-D array with dimensions A, B and C. This array is divided into 27 smaller chunks.

Solution. Space requirement = AC + AB + BC = 750*450 + 100*750 + 100*150 = 427,500

15/1. (3 points) Suppose a market shopping data warehouse consists of four dimensions: customer, date, product, and store, and two measures: count, and avg sales, where avg sales stores the real sales in dollar at the lowest level but the corresponding average sales at other levels.
Draw a **snowflake schema** diagram (sketch it, do not have to mark every possible level, and make your implicit assumptions on the levels of a dimension when you draw it).

Solution. There could be many different answers in the design. One possible answer could be as follows. Dimensions: Customer(name, address, birth of date, ...), Date (day, day of week, month, quarter, year), Product(product name, brand, type, supplier (...), ...), Store (store name, store type, address, ...) Measure: total = count(*), avg sales = avg(sales), ... The dimension tables should be linked to the fact table.

15/2. (3 points) Suppose a market shopping data warehouse consists of four dimensions: customer, date, product, and store, and two measures: count, and avg sales, where avg sales stores the real sales in dollar at the lowest level but the corresponding average sales at other levels.
Draw a **snowflake schema** diagram (sketch it, do not have to mark every possible level, and make your implicit assumptions on the levels of a dimension when you draw it).

Solution. There could be many different answers in the design. One possible answer could be as follows. Dimensions: Customer(name, address, birth of date, ...), Date (day, day of week, month, quarter, year), Product(product name, brand, type, supplier (...), ...), Store (store name, store type, address, ...) Measure: total = count(*), avg sales = avg(sales), ... The dimension tables should be linked to the fact table.
