Homework 2
COMP530 Fall 2020 - Data Privacy and Security

Erhan Tezcan 0070881

05.12.2020

Answers

Answer 1: There are 4 equivalence classes (EC). The questions $1.b$ to $1.h$ will be answered in the following list for each EC, and the remaining answers will follow. All of the values here are obtained by running `q1.py`, which also outputs these answers prettily to the console.

(1) $EC_1 = \{r_1, r_4, r_5, r_9, r_{13}, r_{16}\}$. It is an EC of QIs Teacher, Male, $[50, 60)$. It has SIs Flu (x2), HIV (x1), Diabetes (x3).
- This EC is at most 6-anonymous.
- This EC is at most distinct 3-diverse.
- This EC is at most entropy 1.6688-diverse.
- This EC is at most recursive $(1, 2)$-diverse.
- This EC is at most recursive $(1.0, 2)$-diverse.
- This EC is at least 0.1470-close with variational distance.
- This EC is at least 0.0279-close with Kullback-Leibler divergence.

(2) $EC_2 = \{r_2, r_{11}, r_{14}\}$. It is an EC of QIs Teacher, Female, $[50, 60)$. It has SIs Flu (x1), HIV (x1), Diabetes (x1).
- This EC is at most 3-anonymous.
- This EC is at most distinct 3-diverse.
- This EC is at most entropy 1.4999-diverse.
- This EC is at most recursive $(1, 2)$-diverse.
- This EC is at most recursive $(0.5, 2)$-diverse.
- This EC is at least 0.0392-close with variational distance.
- This EC is at least 0.0015-close with Kullback-Leibler divergence.

(3) $EC_3 = \{r_3, r_6, r_8, r_{17}\}$. It is an EC of QIs Carpenter, Male, $[40, 50)$. It has SIs Flu (x1), HIV (x2), Diabetes (x1).
- This EC is at most 4-anonymous.
- This EC is at most distinct 3-diverse.
- This EC is at most entropy 1.6329-diverse.
- This EC is at most recursive $(1, 2)$-diverse.
- This EC is at most recursive $(1.0, 2)$-diverse.
- This EC is at least 0.2058-close with variational distance.
- This EC is at least 0.0379-close with Kullback-Leibler divergence.

(4) $EC_4 = \{r_7, r_{10}, r_{12}, r_{15}\}$. It is an EC of QIs Painter, Female, $[30, 40)$. It has SIs Flu (x2), HIV (x1), Diabetes (x1).
   - This EC is at most 4-anonymous.
   - This EC is at most distinct 3-diverse.
   - This EC is at most entropy 1.6329-diverse.
   - This EC is at most recursive $(1, 2)$-diverse.
   - This EC is at most recursive $(1.0, 2)$-diverse.
   - This EC is at least 0.1470-close with variational distance.
   - This EC is at least 0.0202-close with Kullback-Leibler divergence.

The questions $1.i$ to $1.k$ are answered in the following list:

- The table is k-anonymous for $k \leq 3$.
- The table is entropy l-diverse for $l \leq 1.4999$.
- The table is t-close for $t \geq 0.2058$ with variational distance.

Answer 2: We know that for an EC q^*, $p(q^*, s)$ is the fraction of records in q^* with $SA = s$. In probabilistic l-diversity, $\forall s \in SA$ we have:

$$p(q^*, s) \leq \frac{1}{l}$$

Let us denote $|q^*|$ as the number of records in EC q^*, and $|q^*_s|$ as the number of records in EC q^* for some sensitive attribute s.

$$\frac{|q^*_s|}{|q^*|} \leq \frac{1}{l}$$

Equivalently, $l \times |q^*_s| \leq |q^*|$.

(a) This is **true**.

Proof. Let EC q^* be a probabilistic l-diverse dataset for some set of sensitive attributes SA. Then, $\forall s \in SA$ we have

$$l \times |q^*_s| \leq |q^*|$$

If we sum these inequalities for all possible values of s we have:

$$\sum_{s \in SA} l \times |q^*_s| \leq \sum_{s \in SA} |q^*|$$

Dividing both sides by l yields:

$$\sum_{s \in SA} |q^*_s| \leq \frac{|SA| \times |q^*|}{l}$$

Notice that $\sum_{s \in SA} |q^*_s| = |q^*|$ because that is basically the number of records for all SA values in this EC. So we have:

$$|q^*| \leq \frac{|SA| \times |q^*|}{l}$$

The $|q^*|$ cancels out,

$$1 \leq \frac{|SA|}{l}$$

Finally:

$$l \leq |SA|$$

This means that you have at least l members in SA, which also means that you have at least l distinct significant attributes, which implies that this EC is l-diverse. □

(b) This is **false**.

Proof. We can show this by giving a counter example. Imagine a 3-diverse EC q^* where the set of sensitive attributes $SA = \{s_1, s_2, s_3\}$ and $|q^*_{s_1}| = 8$, $|q^*_{s_2}| = 1$, $|q^*_{s_3}| = 1$, and $|q^*| = 10$. For s_1 we have:

$$\frac{|q^*_{s_1}|}{|q^*|} = \frac{8}{10} \leq \frac{1}{3}$$

We see that $24/30 \leq 10/30$ is false, so this is wrong. Even though the table is distinct 3-diverse it is not probabilistic 3-diverse. □

(c) This is **true**.

Proof. In a single generalization step $D \to D'$, new ECs are created by merging existing ECs. Let us look at such case, where two ECs, namely q^L and q^R, are merged into q^N. Suppose that the table is probabilistic l-diverse with set of sensitive attributes SA. This implies both q^L and q^R are probabilistic l-diverse. We will see what the diversity might be for q^N.

Similar to what we have shown in the proof of first part of this question, realize that

$$|q^L| \leq \frac{|SA| \times |q^L|}{l} \tag{1}$$

and

$$|q^R| \leq \frac{|SA| \times |q^R|}{l} \tag{2}$$

When these two are merged into q^N, for probabilistic l-diversity we require:

$$|q^N| \leq \frac{|SA| \times |q^N|}{l}$$

Notice that $|q^N| = |q^L| + |q^R|$. Thus,

$$|q^L| + |q^R| \leq \frac{|SA| \times |q^L|}{l} + \frac{|SA| \times |q^R|}{l}$$

We can see that this is infact a true inequality, because it is actually sum of inequalities in equations (1) and (2). This shows that generalization can not result in a dataset that has less probabilistic l-diversity, so we can say that generalization is monotonic with respect to probabilistic l-diversity. □

Answer 3:

(a) The generalized table is given in Table 1.

	Job	Nationality	Gender	Disease
r_1	Professional	American	★	Flu
r_2	Professional	American	★	Gastritis
r_3	Professional	American	★	HIV
r_4	★	Western Europe	Male	Flu
r_5	★	Western Europe	Male	Flu
r_6	Artist	★	Female	Asthma
r_7	Artist	★	Female	Gastritis

TABLE 1. 2-anonymous table with $EC_1 = \{r_1, r_2, r_3\}$, $EC_2 = \{r_4, r_5\}$ and $EC_3 = \{r_6, r_7\}$.

(b) The distortion metrics of the table is $MD(T) = \sum_{r \in R} MD(r)$, and the distortion metric of a record $MD(r)$ is the number of generalizations over it's values, where going up a level in DGH corresponds to 1.

$MD(r_1) = 1 + 2 + 1 = 4$
$MD(r_2) = 1 + 2 + 1 = 4$
$MD(r_3) = 1 + 2 + 1 = 4$
$MD(r_4) = 2 + 1 + 0 = 3$
$MD(r_5) = 2 + 1 + 0 = 3$
$MD(r_6) = 1 + 3 + 0 = 4$
$MD(r_7) = 1 + 2 + 0 = 3$

Therefore, $MD(T) = 4 + 4 + 4 + 3 + 3 + 3 + 4 + 3 = 28$.

(c) The loss metric of the table is $LM(T) = \sum_{r \in R} M(r)$ and the loss metric of a record $LM(r)$ is given by $LM(r) = \sum_{v \in r} w_a \times LM(v)$ where v is a value and a is the corresponding attribute (e.g. $v =$ American and $a =$ Nationality). Then, the loss metric of a value v is given by:

$$LM(v) = \frac{\#\text{ descendant leaves of } v - 1}{\#\text{ leaves in DGH} - 1}$$

With this in mind, let us note that the number of leaves in DGHs: $DGH_{job} = 5$, $DGH_{nationality} = 16$ and $DGH_{gender} = 2$. Also note that our weights for attributes are: $w_{job} = 0.4$, $w_{nationality} = 0.4$, $w_{gender} = 0.2$.

$LM(r_1) = \frac{0.4}{5-1}(3-1) + \frac{0.4}{16-1}(5-1) + \frac{0.2}{2-1}(2-1) = 0.2 + \frac{1.6}{15} + 0.2 = \frac{7.6}{15}$

$LM(r_2) = \frac{0.4}{5-1}(3-1) + \frac{0.4}{16-1}(5-1) + \frac{0.2}{2-1}(2-1) = 0.2 + \frac{1.6}{15} + 0.2 = \frac{7.6}{15}$
$LM(r_3) = \frac{0.4}{5-1}(3-1) + \frac{0.4}{16-1}(5-1) + \frac{0.2}{2-1}(2-1) = 0.2 + \frac{1.6}{15} + 0.2 = \frac{7.6}{15}$
$LM(r_4) = \frac{0.4}{5-1}(5-1) + \frac{0.4}{16-1}(3-1) + \frac{0.2}{2-1}(1-1) = 0.4 + \frac{0.8}{15} + 0 = \frac{6.8}{15}$
$LM(r_5) = \frac{0.4}{5-1}(5-1) + \frac{0.4}{16-1}(3-1) + \frac{0.2}{2-1}(1-1) = 0.4 + \frac{0.8}{15} + 0 = \frac{6.8}{15}$
$LM(r_6) = \frac{0.4}{5-1}(2-1) + \frac{0.4}{16-1}(16-1) + \frac{0.2}{2-1}(1-1) = 0.1 + 0.4 + 0 = 0.5 = \frac{7.5}{15}$
$LM(r_7) = \frac{0.4}{5-1}(2-1) + \frac{0.4}{16-1}(16-1) + \frac{0.2}{2-1}(1-1) = 0.1 + 0.4 + 0 = 0.5 = \frac{7.5}{15}$

Therefore, $LM(T) = 3\frac{7.6}{15} + 2\frac{6.8}{15} + 2\frac{7.5}{15} = \frac{51.4}{15}$.

Answer 4: I have used Python for this question, see `q4.py`. The parameters such as SA, QIs, directory of the DGH, and path to the CSV data is set there.

```
SA = 'income'
QIs = ['education', 'gender', 'marital-status', 'native-
country', 'occupation', 'relationship', 'workclass']
dghDirectory = 'DGHs'
dataPath = 'adult.csv'
```

Note that the DGHs are imported according to the path **dghDirectory/qi.txt** where **qi** is the name of the quasi-identifier, such as "education" or "gender", chosen from the QIs array. The code is also designed for a single SA only.

The DGHs are stored as a tree. From this tree, we obtain three dictionaries:

(1) `qi_depths` stores the depth of each value of every QI in it's respective DGH.
(2) `qi_descendant_leave_counts` stores the number of descendant leaves of each value of every QI in it's respective DGH.
(3) `qi_hierarchical_mappings` stores the value a QI can take, depending on the level of DGH.

These dictionaries are used to enhance the performance of calculating distortion metric and loss metric from a dynamic programming aspect. In those calculations, for example instead of calculating depth or number of descendant leaves everytime, we just refer to this dictionary which we compute at the beginning. Together with built-in map and sum functions, this greatly increases the performance.

To showcase my distortion metric and loss metric functions, I create a completely anonymized table (where all values are Any) and calculate the cost. For the given `adults.csv` dataset, I obtain:

- CostMD: 650751
- CostLM: 45222.0

Though CostMD depends on the structure of DGH, CostLM tells much more about the correctness of the function. The loss metric cost is equal to number of records, and since all of values are generalized to the maximum so we should indeed expect the maximum error, which is the number records.

Anonymization: I couldn't do this part sadly. I guess I couldn't really understand top-down anonymization, and I am unable to implement it. I have tried many methods but I am stuck. I hope the prior

parts of this question get graded, as I believe they are correct nevertheless. You can also use Spyder and see how DGHs are stored in the variable explorer there.

Answer 5: Let I denote our item set as $I = \{$Wine, Diapers, Beer, Pregnancy Test$\}$.

(a) No. The transactions T1, T2, T6, T7 and T8 break 2-anonymity.

(b) Yes. 2 distinct items $i, j \in I$ are not enough to identify less than 2 transactions in the dataset.

(c) No. If you know that a transaction contains {Wine, Diapers, Pregnancy Test} then you can identify T2.

(d) Yes. 2 distinct items $i, j, k \in I$ are not enough to identify less than 3 transactions in the dataset.

(e) True.

Proof. Suppose that the set-valued dataset is k-anonymous, however is not k^m-anonymous for some m but for same k. Let T be the set of transactions that can be identified given m items, where $|T| < k$ as per the definition of k^m-anonymity. k-anonymity assures that at best you can identify k records. This contradicts that $|T| < k$, therefore we can say that if the set-valued dataset is k-anonymous then it is k^m-anonymous for the same value of k and for all $m \geq 1$.

Another way to think about this proof is to notice that k-anonymity is more strict than k^m-anonymity, and it is by intuition that we can expect k^m-anonymity to hold given k-anonymity. □

Answer 6:

(a) Charlie has 4 connections and Dave has 3 connections. The number of connections is the degree of a vertex in the graph. Charlie then knows that it is the vertex with degree 4, and Dave knows it is the vertex with degree 3. Fortunately for them there is only 1 such vertex for both degrees.

Charlie and Dave can then identify Ethan, because they both know him and they can see a mutual connection in the graph. After identfying Ethan, Dave knows 2 of the 3 connections: Charlie and Ethan. Now it knows the remaining connection: Frank.

Charlie can then infer that the remaining connections are Alice and Bob, and they are both negative.

As a result, they can learn that Alice, Bob and Ethan are COVID negative, and the rest of the people are COVID positive.

(b) A graph $G(V, E)$ is k-degree anonymous if every node in V has the same degree as $k-1$ other nodes in V. For 2-degree anonymity, this would mean every node would have 1 more node with the same degree.

We can add an edge between Bob and Dave. The result is given in figure below. There are 2 vertices with degree 1, 2 vertices with degree 2 and 2 vertices with degree 4, therefore the graph is 2-degree anonymous.

FIGURE 1. 2-degree anonymous graph.