DataPrivacy—hw1

Terence Wang

2023/11/05

Contents

1 Q1

1.1 a

The quasi-identifier attributes: **Zip Code Age Salary Nationality**

1.2 b

After cell-level generalization:

Sequence	Zip Code	Age	Salary	Nationality	Condition
1	130**	[21,40]	[13k,22k]	*	Heart Disease
2	130**	[21,40]	[23k,25k]	*	Heart Disease
3	130**	[21,40]	[13k,22k]	Japanese	Viral Infection
4	130**	[21,40]	[13k,22k]	*	Viral Infection
5	1485*	[41,55]	[13k,22k]	*	Cancer
6	1485*	[41,55]	[13k,22k]	*	Heart Disease
7	1485*	[41,55]	[13k,22k]	*	Viral Infection
8	1485*	[41,55]	[13k,22k]	*	Viral Infection
9	130**	[21,40]	[13k,22k]	*	Cancer
10	130**	[21,40]	[23k,25k]	*	Cancer
11	130**	[21,40]	[13k,22k]	Japanese	Cancer
12	130**	[21,40]	[13k,22k]	*	Cancer

That is:

Sequence	Zip Code	Age	Salary	Nationality	Condition
1	130**	[21,40]	[13k,22k]	*	Heart Disease
4	130**	[21,40]	[13k,22k]	*	Viral Infection
9	130**	[21,40]	[13k,22k]	*	Cancer
12	130**	[21,40]	[13k,22k]	*	Cancer
2	130**	[21,40]	[23k,25k]	*	Heart Disease
10	130**	[21,40]	[23k,25k]	*	Cancer
3	130**	[21,40]	[13k,22k]	Japanese	Viral Infection
11	130**	[21,40]	[13k,22k]	Japanese	Cancer
5	1485*	[41,55]	[13k,22k]	*	Cancer
6	1485*	[41,55]	[13k,22k]	*	Heart Disease
7	1485*	[41,55]	[13k,22k]	*	Viral Infection
8	1485*	[41,55]	[13k,22k]	*	Viral Infection

Generalization hierarchies are as follows: Figure 1 Figure 2

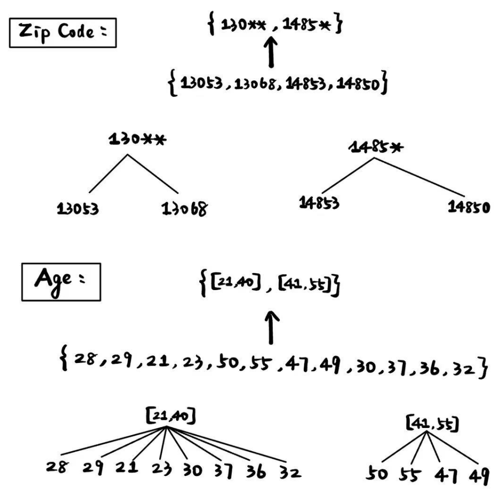

Figure 1: generalization hierarchies

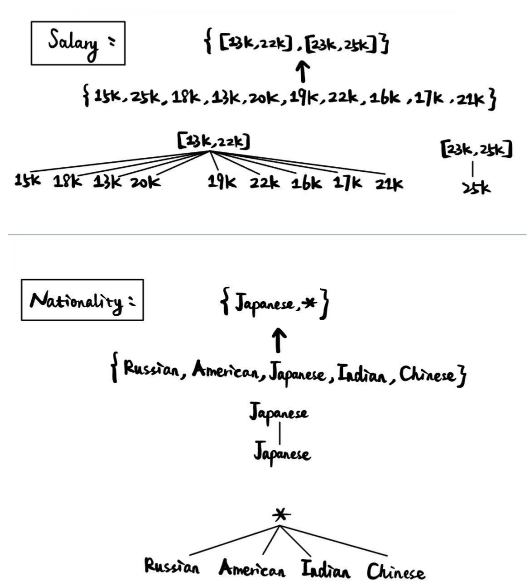

Figure 2: generalization hierarchies

Calculation of the LM:

Zip Code:

$T[13053] = \frac{2-1}{4-1} = \frac{1}{3}$
$T[13068] = \frac{2-1}{4-1} = \frac{1}{3}$
$T[14853] = \frac{2-1}{4-1} = \frac{1}{3}$
$T[14850] = \frac{2-1}{4-1} = \frac{1}{3}$
Therefore, $LM_{ZipCode} = \frac{1}{3}$

Age:

$T[21-40] = \frac{40-21}{55-21} = \frac{19}{34}$
$T[41-55] = \frac{55-41}{55-21} = \frac{14}{34}$
Therefore, $LM_{Age} = (8 \times \frac{19}{34} + 4 \times \frac{14}{34}) \times \frac{1}{12} = \frac{26}{51}$

Salary:

$T[13-22] = \frac{22-13}{25-13} = \frac{3}{4}$
$T[23-25] = \frac{25-23}{25-13} = \frac{1}{6}$
Therefore, $LM_{Salary} = (10 \times \frac{3}{4} + 2 \times \frac{1}{6}) \times \frac{1}{12} = \frac{47}{72}$

Nationality:

$LM_{Nationality} = \frac{3}{4} \times \frac{4}{5} = \frac{3}{5}$
In conclusion, $LM = \frac{1}{3} + \frac{26}{51} + \frac{47}{72} + \frac{3}{5} = \frac{12827}{6120} \approx 2.1$

2 Q2

2.1 a

meet recursive (2,2)-diversity

Figure 3

> We say that a $q^\star$-block is $(c, 2)$-diverse if $r_1 < c(r_2 + \cdots + r_m)$ for some user-specified constant c. For $\ell > 2$, we say that a $q^\star$-block satisfies *recursive* (c, ℓ)-*diversity* if we can eliminate one possible sensitive value in the $q^\star$-block and still have a $(c, \ell - 1)$-diverse block. This recursive definition can be succinctly stated as follows.

Figure 3: recursive (c,l)-diversity definition

For every QI-cluster, we have $r_1 = 2,\ r_2 = 1,\ r_3 = 1$. So $r_1 < 2 \times (r_2 + r_3)$ holds (i.e. $2 < 2 \times 2$)

2.2 b

Entropy is a concave(the definition of **concave** may be vague) function. Thus if QI-cluster $q_1^\star,\ \cdots\cdots,\ q_d^\star$ from table T are merged to form the QI-cluster $q^{\star\star}$ of table $T^\star$, then we have $entropy(q^{\star\star}) \geq \min_i(entropy(q_i^\star))$
Since table T satisfies entropy l-diversity, we have $entropy(q^{\star\star}) \geq \min_i(entropy(q_i^\star)) \geq \log(l)$. Therefore, we get $T^\star$ satisfies entropy l-diversity.

3 Q3

3.1 a

To calculate EMD under ordered distance, we just need to consider flows that **transport distribution mass between adjacent elements**. This is because other circumstances can be decomposed into several transportations between adjacent elements.
So let us consider element 1 first. Let us assume that $p_1 - q_1 < 0$, so $q_1 - p_1$ should be transported from other elements to element 1.
We can transport this from element 2. So after transportation, element 2 has an extra amount of $(p_1 - q_1) + (p_2 - q_2)$. The operation is similar element 3 $\cdots\cdots$
Therefore, we can get $D[\mathbf{P},\mathbf{Q}] = \frac{1}{m-1}(|r_1|+|r_1+r_2|+\cdots+|r_1+r_2+\cdots+r_{m-1}|)$

3.2 b

$m = 9,\ ordered\ list = \{3K,\ 4K,\ 5K,\ 6K,\ 7K,\ 8K,\ 9K,\ 10K,\ 11K\}$
Distribution of the whole table:
$\mathbf{Q} = \{\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}\}$
Distribution of each cluster:
$\mathbf{P}_1 = \{0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0, 0, 0\}$
$\mathbf{P}_2 = \{\frac{1}{3}, 0, 0, 0, 0, \frac{1}{3}, 0, 0, \frac{1}{3}\}$
$\mathbf{P}_3 = \{0, 0, 0, 0, \frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3}, 0\}$
$D[\mathbf{P}_1,\mathbf{Q}] = \frac{1}{8} \times \frac{1}{9}(1 + 1 + 3 + 5 + 4 + 3 + 2 + 1) = \frac{5}{18}$
$D[\mathbf{P}_2,\mathbf{Q}] = \frac{1}{8} \times \frac{1}{9}(2 + 1 + 0 + 1 + 2 + 0 + 1 + 2) = \frac{1}{8}$
$D[\mathbf{P}_3,\mathbf{Q}] = \frac{1}{8} \times \frac{1}{9}(1 + 2 + 3 + 4 + 2 + 3 + 1 + 1) = \frac{17}{72}$
$t = \max\{D[\mathbf{P}_1,\mathbf{Q}],\ D[\mathbf{P}_2,\mathbf{Q}],\ D[\mathbf{P}_3,\mathbf{Q}]\} = \frac{5}{18} \approx 0.278$

4 Q4

4.1 a

question1:
$P_{prior}(x=0)=0.01$
$P(R_1(x)=0)=0.3\times 0.01+0.7\times\frac{1}{101}=0.00993$
$P_{posterior}(x=0|R_1(x)=0)=\frac{0.01\times(0.3+0.7\times\frac{1}{101})}{0.00993}=0.3091$
question2:
$P_{prior}=0.01$
$P(R_2(x)=0)=P(x+\xi=0)+P(x+\xi=101)=0.01\times\frac{1}{21}+10\times 0.0099\times\frac{1}{21}+10\times 0.0099\times\frac{1}{21}=0.00990$
$P(R_3(x)=0)=0.5\times 0.00990+0.5\times\frac{1}{101}=0.00990$
$P_{posterior}(x=0|R_3(x)=0)=\frac{0.01\times(0.5\times\frac{1}{21}+0.5\times\frac{1}{101})}{0.00990}=0.0291$
question3:
$P_{prior}(x\in[20,80])=61\times 0.0099=0.6039$
$P_{posterior}(x\in[20,80]|R_2(x)=0)=\frac{0}{0.00990}=0$

4.2 b

R_3 is more suitable. This is because $P_{posterior}$ should be close enough to P_{prior}, in this way we can protect the privacy. Apparently, R_3 is closest to *nothing* in both $X=0$ and $X\notin\{200,\cdots,800\}$. Therefore, R_3 is more suitable.

5 Q5

5.1 a

Figure 4

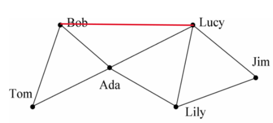

Figure 4: 2-anonymous

5.2 b

Figure 5

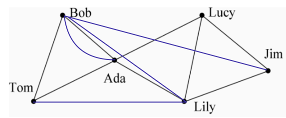

Figure 5: 3-anonymous

5.3 c

Figure 4:
$L(G, G') = 1 - \frac{8}{9} = \frac{1}{9}$
Figure 5:
$L(G, G') = 1 - \frac{8}{12} = \frac{1}{3}$