2021.5.10

1.

当 $l=2$ 时，若攻击者已知"Bob does not have heart disease"，则攻击者知道Bob可能患的病为概率为 $r_2,\cdots,r_m$ 对应的病.

当 $l>2$ 时，共有 $r_1, r_2, r_3, \cdots, r_{l-1}$，和 $\sum_{i=l}^{m} r_i$ l 个频次比较大的集合，若攻击者已知 $l-2$ 条形如"Bob does not have heart disease"的信息，则攻击者仍不知道Bob患病的病属于~~哪个~~ l 个中的哪个集合。但 $l-1$ 条时，则攻击者可排除 $l-1$ 个集合，则可以确定Bob所患的病属于已知哪个集合.

2.

Race: R_1	ZIP: Z_0	Race: R_1	ZIP: Z_1	Race: R_1	ZIP: Z_2
person	94138	person	9413*	person	941**
person	94138	person	9413*	person	941**
person	94138	person	9413*	person	941**
person	94138	person	9413*	person	941**
person	94142	person	9414*	person	941**
person	94142	person	9414*	person	941**
person	94142	person	9414*	person	941**
		person	9414*	person	941**
		person	9414*	person	941**

No.

Date.

9×8

Race: R_0 ZIP: Z_2

asian 941**

asian 941**

asian 941**

asian 941**

black 941**

black 941**

black 941**

3.

~~(a) $D[P_1,Q]=\frac{1}{8}($~~

(a) $r_1=\frac{1}{3}-\frac{1}{9}=\frac{2}{9}$ $r_2=r_4=r_5=r_6=r_8=r_9=-\frac{1}{9}$ $r_3=\frac{1}{3}-\frac{1}{9}=\frac{2}{9}$ $r_7=\frac{1}{3}-\frac{1}{9}-\frac{2}{9}$

$D[P_1,Q]=\frac{1}{8}(\frac{2}{9}+\frac{1}{9}+\frac{3}{9}+\frac{2}{9}+\frac{1}{9}+0+\frac{2}{9}+\frac{1}{9}+0)=\frac{1}{6}=\frac{12}{72}$

$r_1=r_2=r_3=r_5=r_7=r_8=-\frac{1}{9}$ $r_4=r_6=r_9=\frac{2}{9}$

$D[P_2,Q]=\frac{1}{8}(\frac{1}{9}+\frac{2}{9}+\frac{3}{9}+\frac{1}{9}+\frac{2}{9}+0+\frac{1}{9}+\frac{1}{9}+0)=\frac{11}{72}$

$r_2=r_5=r_8=\frac{2}{9}$ $r_1=r_3=r_4=r_6=r_7=r_9=-\frac{1}{9}$

$D[P_3,Q]=\frac{1}{8}(\frac{1}{9}+\frac{1}{9}+0+\frac{1}{9}+\frac{1}{9}+0+\frac{1}{9}+\frac{1}{9}+0)=\frac{6}{72}$

~~$t=\max$~~

$D[P,Q]=0$

$t=\max_i[D[P_i,Q]]=\frac{12}{72}=\frac{1}{6}$

No.

Date.

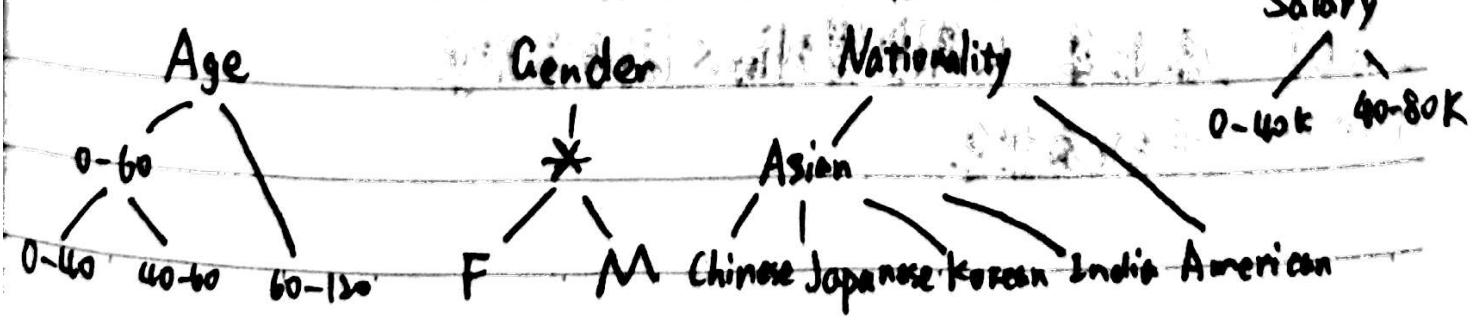
4.

(a)

Age, Gender, Nationality, Salary

(b)

Age	Gender	Nationality	Salary	Condition
0-40	F	Asian	0-40k	VI
0-40	~~M~~ *	American	0-40k	Flu
~~40-60~~ 0-60	*	Asian	40-80k	HD
0-40	M	Asian	0-40k	Flu
~~40-60~~ 0-60	*	Asian	40-80k	HD
0-40	~~M~~ *	American	0-40k	Flu
0-40	M	~~American~~ Asian	0-40k	Flu
0-40	F	Asian	0-40k	Cancer
0-40	~~M~~ *	American	0-40k	VI
0-40	F	Asian	0-40k	Cancer
0-60	*	American	40-80k	VI
0~~40~~-60	*	American	40-80k	HD
0-40	~~M~~ *	American	0-40k	VI



No.

Date.

the loss for attribute

Age $4 \times \frac{1}{2} \times \frac{1}{13} = \frac{2}{13}$

Gender $8 \times 1 \times \frac{1}{13} = \frac{8}{13}$

Nationality $7 \times \frac{3}{4} \times \frac{1}{13} = \frac{21}{52}$

Salary 0

$LM = \frac{2}{13} + \frac{8}{13} + \frac{21}{52} = \frac{61}{52}$

(c)

K-Optimize(k, head set H, tail set T, best cost c)

T ← PRUNE-USELESS-VALUES(H,T)

c ← MIN(c.COMPUTE-COST(H))

T ← PRUNE(H,T,c)

T ← REORDER-TAIL(H,T)

while T is non-empty do

v ← the first value in the ordered set T

H_{new} ← H∪{v}

T ← T-{v}

參考KACA算法

KACA(D,k)

INPUT: D數據集 k期望得到的k-匿名的數據集

OUTPUT: k-匿名的數據集

Date.

1. 由数据集D生成初始等价类，等价类中各个元组的QI准标识符的值相等
2. 若若不存在元组个数小于k的等价类，转8
   否则转3
3. 随机选择一个大小小于k的等价类C
4. 计算C与其它所有等价类的距离
5. 选找到距C最近的等价类C'
6. 将C和C'并为一类，并泛化C和C'
7. 转2
8. 返回k-匿名后的数据集

注：距离计算参考EMD和Ordered Distance
泛化过程中选取LM小的泛化方式，因数据量因k有限，故复杂度不会太高.

No.

Date.

5.

(a)

~~Priar~~

$P(R_1(X)=0 \mid X=0) = 20\% + 80\% \times \frac{1}{1001} \approx 0.2008$

$P(R_1(X)=0 \mid X=i) = 80\% \times \frac{1}{1001} \approx 0.0008$

$i \in \{200, \cdots, 800\}$

$P(R_2(X)=0 \mid X=0) = \frac{1}{201} \approx 5 \times 10^{-3}$

$P(R_2(X)=0 \mid X=i) = 0$

$i \in \{200, \cdots, 800\}$

$P(R_3(X)=0 \mid X=0) = 50\% \times \frac{1}{201} + 50\% \times \frac{1}{1001} \approx 3 \times 10^{-3}$

$P(R_3(X)=0 \mid X=i) = 50\% \times 0 + 50\% \times \frac{1}{1001} \approx 5 \times 10^{-4}$

$i \in \{200, 300 \cdots, 800\}$

(b) $P(R_1(X)=0) = 20\% \times 0.01 + 80\% \times \frac{1}{1001} \approx 2.80 \times 10^{-3}$

$P(R_1(X)=i) = 20\% \times 0.00099 + 80\% \times \frac{1}{1001} \approx 1 \times 10^{-4}$

$P(R_2(X)=0) = 0.01 \times \frac{1}{201} + 0.00099 \times 200 \times \frac{1}{201} \approx 1.03 \times 10^{-3}$

$P(R_2(X)=i) = 0.00099 \times \frac{1}{201} + 0.00099 \times 200 \times \frac{1}{201} \approx 9.9 \times 10^{-4}$

$P(R_3(X)=0) = 0.5 \times P(R_2(X)=0) + 0.5 \times \frac{1}{1001} \approx 1.01 \times 10^{-3}$

$P(R_3(X)=i) = 0.5 \times P(R_2(X)=i) + 0.5 \times \frac{1}{1001} \approx 9.9 \times 10^{-4}$

(b)

方法(c)较好，因为该方法的先验概率和后验概率差距较小，能更好地保护隐私.

6.

假设 $\gamma \le \frac{\beta}{\alpha} \cdot \frac{1-\alpha}{1-\beta}$ ~~且 R 满足 且 R 是 γ-amplifying 的~~

且 $\exists u \in D_U \ \exists v \in D_V$ s.t. $Pf(\overline{\in}(u_1)) \le \alpha$ and $Pf(\overline{\in}(u_1) \mid R(u_1)=v) \ge \beta$

$\Rightarrow \quad P(R(u_1)=v) \le \frac{\alpha}{\beta}$

又有 $Pf(\neg\overline{\in}(u_2)) \ge 1-\alpha$ and $Pf(\neg\overline{\in}(u_2) \mid R(u_2)=v) \le 1-\beta$

$\Rightarrow \quad P(R(u_2)=v) \ge \frac{1-\alpha}{1-\beta}$

$\Rightarrow \gamma \ge \frac{P(R(u_2)=v)}{P(R(u_1)=v)} \ge \frac{\beta}{\alpha} \cdot \frac{1-\alpha}{1-\beta}$ 矛盾 (因 R 是 γ-amplifying 的)

假设 $\gamma \le \frac{\beta}{\alpha} \cdot \frac{1-\alpha}{1-\beta}$ ~~且 R 是 γ-amplifying~~

且 $\exists u \in D_U \ \exists v \in D_V$ s.t. $Pf(\overline{\in}(u_1)) \ge \beta$ and $Pf(\overline{\in}(u_1) \mid R(u_1)=v) \le \alpha$

$\Rightarrow P(R(u_1)=v) \ge \frac{\beta}{\alpha}$

又有 $Pf(\neg\overline{\in}(u_2)) \le 1-\alpha$ and $Pf(\neg\overline{\in}(u_2) \mid R(u_2)=v) \ge 1-\alpha$

$\Rightarrow P(R(u_2)=v) \le \frac{1-\beta}{1-\alpha}$

$\Rightarrow \gamma \ge \frac{P(R(u_1)=v)}{P(R(u_2)=v)} \ge \frac{\beta}{\alpha} \cdot \frac{1-\alpha}{1-\beta}$ 矛盾

所以 ~~若 $\gamma \le \frac{\beta}{\alpha} \cdot \frac{1-\alpha}{1-\beta}$~~ 当 R 是 γ-amplifying 的

若 $\gamma \le \frac{\beta}{\alpha} \cdot \frac{1-\alpha}{1-\beta}$, 则 R 满足 (α, β)-privacy