Methodology, Ethics and Practice of Data Privacy HW1

王嵘晟 PB1711614

1.

(a.)

Age Gender Nationality Salary

(b.)

Name	Age	Gender	Nationality	Salary	Condition
Irene	17-26	F	American	16K-30K	Viral Infection
Fox	17-26	F	American	16K-30K	Flu
Martin	17-26	F	American	16K-30K	Viral Infection
Jean	17-26	F	American	16K-30K	Cancer
Helen	17-26	F	American	16K-30K	Cancer
Bluce	27-36	M	*	30k-40k	Flu
Dick	27-36	M	*	30k-40k	Flu
Ann	27-36	M	*	30k-40k	Viral Infection
Gary	27-36	M	*	30k-40k	Flu
ken	>37	M	American	>40K	Viral Infection
Cary	>37	M	American	>40K	Heart Disease
Lewis	>37	M	American	>40K	Heart Disease
Eshwar	>37	M	American	>40K	Heart Disease

对于 Age: $t[17-26]=\frac{5-1}{10-1}=\frac{4}{9}$ $t[27-36]=\frac{1}{3}$ $t[37-120]=\frac{3}{82}$

对于 Salary: $t[16K-30K]=\frac{4}{15K}t[30k-40k]=\frac{3}{10K}$ 设 Salry 最大值为 70K,$t[40k-70k]=\frac{3}{30K}$

(c.)

设数据共有 n 个元组，首先将待泛化的属性值进行排序，然后从 K=2 开始选择 K 匿名的 K 值，并进行泛化操作，若泛化结果失败则 K+1 直到 $\frac{n}{2}$，当泛化成功时记录 LM 的值，并利用深度优先搜索的思想在 K 不变的情况下继续搜索有没有不同的泛化方法同样记录 LM 的值，如果得到更优的则做替换，否则 K+1 重复循环，直到 $\frac{n}{2}$ 为止。

2.

(a.)

令 $Y = R_i(x)$

当 $X = 0$ 时，对于 $R_1(X)$，先验概率：$P = 0.01$，后验概率：$P(X=0|Y) = \frac{P(X=0)P(Y|X=0)}{P(X=0)P(Y|X=0)+P(X=0)P(Y|X\neq 0)} = 0.2+0.8*\frac{1}{1001}$ 对于 $R_2(x)$，先验概率 $P = 0.01$，后验概率：$P(X=0|Y) = \frac{P(X=0)P(Y|X=0)}{P(X=0)P(Y|X=0)+P(X=0)P(Y|X\neq 0)} = \frac{1}{201}$。对于 $R_3(x)$，先验概率 $P = 0.01$，后验概率：$P(X=0|Y) = \frac{P(X=0)P(Y|X=0)}{P(X=0)P(Y|X=0)+P(X=0)P(Y|X\neq 0)} = 0.5*(\frac{1}{201}+\frac{1}{1001})$。

当 $X \in [200,800]$，先验概率：$P = 0.59499$ 对于 3 种方法都成立，后验概率：$P(X \in [200,800]|Y) = \frac{P(X\in[200,800])P(Y|X\in[200,800])}{P(X\in[200,800])P(Y|X\in[200,800])+P(X\notin[200,800])P(Y|X\in[200,800])}$。带入得：$R_1(x)$ 的后验概率为：$P = 0.83$,$R_2(x)$ 的后验概率为：$P = 1$，$R_3(x)$ 的后验概率为 $P = 0.708$

(b.)

R_3 最好，因为根据先验概率和后验概率的计算发现第三种方法最小概率揭露了实际信息。

3.

证明：反证法，如果允许 (α,β) 隐私侵犯，则必须要满足 (2)(3) 两项不等式。但由于 (4) 式中对于 $\gamma - amplifying$ 的定义，$\frac{p(R(u_1)=v)}{p(R(u_2)=v)} \leq \gamma \leq \frac{\beta(1-\alpha)}{\alpha(1-\beta)}$，再由贝叶斯公式，可计算发现与 (2)(3) 两式矛盾，所以原命题成立。