11/11/8

Plan:

11/13	important talk
11/18	go over for test.
11/20	Test 2
11/25	Go over test 2
12/2 12/4	Presentations. & projects

Privacy issues

Privacy for the data that is used

Problem. We have data that we actually use, but we need to persene privacy.

We have.

statistical Data base : You are interested in statistics and no in individual records.

∘ usual databases. objective get individual records

e.g.,

+ Census data base ;– good source for correlations
+ Medical database :–

(Need to get data without disclosing sensitive info.)

* Many traditional ways to protect privacy.

+limit size of the sample below.

$x_1, \quad x_2, \quad \dots, x_n.$

– Problem: data 1, ... data 1003

data 2 1001

$$\frac{s_1 + \dots + s_K}{K} = \bar{S}_1$$

$$\frac{s_1 + \dots + s_{K-1}}{K-1} = \bar{S}_2$$

$$\rightarrow \bar{S}_1 \cdot K = S_1 + \dots + S_{K-1} + S_K$$

$$\rightarrow \bar{S}_2(K-1) = S_1 + \dots + S_{K-1}$$

$$\bar{S}_1 \cdot K - \bar{S}_2(K-1) = \dots S_K$$

11|11|08

$$\begin{array}{cccc|l} 10 & 100 & 80 & 50 & \longrightarrow 60 = \bar{S}_1 \\ S_1 & S_2 & S_3 & S_4 & \longleftarrow \frac{S_1+S_2+S_3}{3} \end{array}$$

60
4)240

$$\frac{10+100+80+50}{4} = 60$$

$$\frac{10+100+80}{3} = \frac{190}{3} = 63.333$$

$$S_1+S_2+S_3+S_4 = \bar{S}_1 \cdot 4 = 240$$
$$S_1+S_2+S_3 = \bar{S}_2 \cdot 3 = 190$$

<u>Another trick.</u>

you always get the same random #.

$$\begin{cases} S. + \Delta S_1 = S_1 \\ S. + \Delta S_2 = S_2 \\ \dots \\ S + \Delta S_n = S_n \end{cases}$$

Radically different approach.

- Don't Keep exact data
- Keep only bounds. [20, 30], [30, 40], ...,

Salary : [10, 30], [20, 100] [100, 300]

11 | 11 | 08

+ Privacy is perserved.

– how to compute statistical characteristics.

If we know s_i

$$S = \frac{s_1 + \dots + s_n}{n}$$

$$V = \frac{1}{n} \sum_{i}^{n} (s_i - s_{av})^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} s_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} s_i\right)^2$$

we only know

$s_i \in [\underline{s}_i, \overline{s}_i]$

$$\underline{S}_{av} = \frac{\underline{s}_1 + \dots + \underline{s}_n}{n}$$

$$\overline{S}_{av} = \frac{\overline{s}_1 + \dots + \overline{s}_n}{n}$$

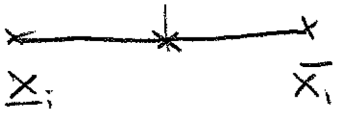
in this case the fcn is monotonic.

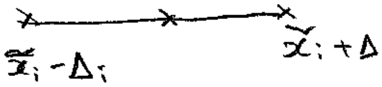
$y = f(x_1, \dots, x_n)$

Previously: $[x_1 - \Delta_i, x_i + \Delta_i]$

We know: $X_i \in [\underline{x}_i, \overline{x}_i]$

Want range $\{ y = f(x_1, \dots, x_n) : x_i \in [\underline{x}_i, \overline{x}_i] \}$





$$\boxed{\Delta_i = \frac{\overline{X}_i - \underline{X}_i}{2}}$$

half width.

Use Monte Carlo when f is easy.

11/11/08

Monte Carlo Method

☑ fewer calls to f

— more computations $\sum_{i=1}^{n} |f(\tilde{x}_1, \ldots, \tilde{x}_{i-1}, \tilde{x}_i + \Delta_i, \tilde{x}_{i+1}, \ldots, \tilde{x}_n) - \tilde{y}|$

⚠ Question

- Here is the characterstic
- Here is the Input
- Find the values.

interval	mid point	half width
[10, 20]	15	5
[20, 30]	25	5
[60, 70]	65	5