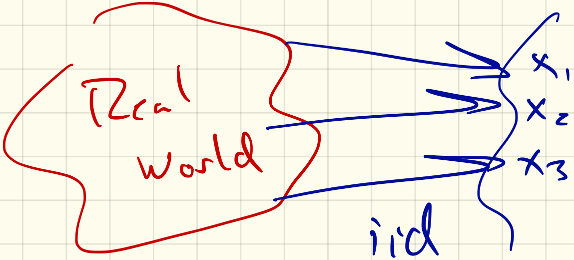
Data Mining L2

Statistical Principals + Hashing

IID Data

Independent and Identically Distributed



input data

$$\underline{X} = \{x_1, x_2, \dots x_n\}$$

set

This lecture

Assume iid set $X = \{x_1, x_2 \ldots x_m\}$

$x_i \in [n] = \{1, 2, 3, \ldots, n\}$

in

Represent $[n]$: All possible IP address

All words in dictionary

All possible birthdays

$n = 365$

Assume each x_i uniform in $[n]$

$$\Pr[x_i = j] = \begin{cases} \frac{1}{n} & \text{if } j \in [n] \\ 0 & \text{o.w.} \end{cases}$$

Hash Table and Hash Function (Random)

Family Hash Functions $\mathcal{H}$

$h_a \in \mathcal{H} \leftarrow$ choice random

$h_a : \Sigma^k \rightarrow [n]$ deterministic

$$\Pr_{h_a \in \mathcal{H}} \left[h_a(\text{string1}) = h_a(\text{string2}) \right] = \frac{1}{n}$$

as long as string1 $\neq$ string2

1. Built in Hash Function

ex. SHA-1

$$h_a(x) = \text{SHA-1}(\text{concat}(a, x))$$

"salt"

some string.

2. Multiplicative Hashing

$$h_a(x) = \lfloor n \cdot \text{frac}(x \cdot a) \rfloor$$

salt

large int

with binary representation

mix 0s 1s

$$\text{frac}(11.278) = 0.278$$

$$h_a(x) = \frac{xa}{2^q} \bmod m$$

3. Modular Hashing

$$h(x) = x \bmod m$$

Do Not Use

Input : sequence distinct strings

each hash with $h_a \in \mathcal{H}$

$h_a(str) \to [n]$

Q1: How many until collision.

"Birthday Paradox"

Jan	1, 9
Feb	8, 15, 27
Mar	
Apr	21, 19
May	17
Jun	25, 8
Jul	
Aug	25, 20
Sep	17
Oct	
Nov	3, 30
Dec	3, 27

18 trials

$$\Pr[s_1 = s_2] = \frac{1}{n}$$

k people $\rightarrow \binom{k}{2} = \frac{k(k-1)}{2}$ pairs

$$\Pr.[\text{no coll } k \text{ pairs}] = \left(1 - \frac{1}{n}\right)^{\binom{k}{2}}$$

$n = 365 \quad k = 23 \Rightarrow 0.467$

pigeon hole principal ? $k = 366 \quad \left(1 - \frac{1}{n}\right)^{\binom{366}{2}} > 0$

- Assume uniform prob.

- True prob after k steps

$$1 - \left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\left(\frac{n-3}{n}\right) \cdots \left(\frac{n-k}{n}\right)$$

$$= 1 = \prod_{i=1}^{k} \left(\frac{n-i}{n}\right)$$

Prob[coll > 50] after $\approx k = \sqrt{2n}$ steps

$k = 18$ Prob(coll) $= 0.34$

$k = 28$ Prob(coll) $= 0.64$

$27 \approx \sqrt{2 \cdot 365}$

QZ: When do we see **all** birthdays?

$\geq n$

${}^{10}\ n^2$? ${}^{5}\ n^{1.5}$? ${}^{25}\ \Theta(n)$?

$\boxed{{}^{40}\ n \log n ?}$ $(n!)$? = n factorial

Coupon Collectors

Analyze Sequence

$s_1 \quad s_2 \quad s_3 \quad s_4 \quad \ldots \quad s_k$

new new x new

Let T_i = # steps until ith distinct observation

$t_i = T_i - T_{i-1}$

= # steps between $(i-1)$th distinct and ith distinct.

Expected total # $= E\left[\sum_{i=1}^{n} t_i\right] = \sum_{i=1}^{n} E[t_i]$

$$E[t_i] = \frac{n-(i-1)}{n}$$ $$= \frac{n}{n-i+1}$$

H_n = nth Harmonic number

$$\sum_{i=1}^{n} E[t_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \sum_{i=1}^{n} \frac{1}{n-i+1} = n \boxed{\sum_{i=1}^{n} \frac{1}{i}}$$

$$\frac{1}{n}, \frac{1}{n-1}, \frac{1}{n-2}, \dots, 1$$

Prob[seeing ith distinct, given i-1 distinct]

$$= \cancel{\frac{1}{n-(i-1)}} = \frac{n-(i-1)}{n} = \frac{n-i+1}{n}$$

$$H_n = \underset{\approx 0.577}{\gamma} + \ln(n) + o(1/n)$$

$$T_n = n H_n = n(0.577 + \ln n)$$

Probably Approximately Correct (PAC)

want to estimate μ

estimate $\hat{x}$

$\delta \in [0,1]$

$\hat{x}, \mu \in [0,1]$

$0 < \varepsilon << 1$

$$\Pr[|\hat{x} - \mu| > \varepsilon] < \delta$$

acceptable error

prob. of failure

\# trials $\approx \frac{1}{\varepsilon^2} \log \frac{1}{\delta}$

$$\Pr[|\hat{x} - \mu| < \varepsilon] \geq 1 - \delta$$