Data Mining L2

Statistical Principles + Hashing

Data $X = \{x_1, x_2, \ldots x_n\}$

$X \underset{\boxed{iid}}{\sim} \mu$ ← unknown distribution

iid: independent and identically distributed.

Each $x_i \in [m] = \{1, 2, , \dots, m\}$

IP addressess

all possible words

days in the year

$m = 365$

---

known distribution $\mathcal{U}$ = uniform

$\Pr[x_i = j] = \frac{1}{m}$ for any $j \in [m]$

Hashing (Hash Table)

(Random) Hash Function

$h : \text{Domain} \rightarrow \text{Range}$ deterministic

Σ^{k} ↑ \Sigma

$[m]$

family of hash functions

Randomly select $h_a \in \mathcal{H}$

$$\Pr_{h_a \sim \mathcal{H}} [h_a(x) = h_a(y)] = \frac{1}{m} \qquad x \neq y$$

- Built-In Hash function

$SHA\text{-}1 : (\Sigma = \{0,1\})^k \to [m = 2^{160}]$

$a = \text{salt}$

input $x \to SHA\text{-}1(\text{concat}(x, a))$

- <u>Multiplicative Hashing</u>

$\text{frac}(17.32) = 0.32$

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

$= (xa \,/\, 2^8) \bmod m$

a large number

- <u>Modular Hashing</u> $h(x) = x \bmod m$

DO NOT USE

$u \in \mathrm{Unif}(0,1)$

$\lfloor u \cdot m \rfloor \to j \in [m]$

- How many samples $x_1, x_2, \dots x_n \in [m]$

so any two $x_i = x_j$

a <u>collision</u>

$? \left(\frac{1}{m}\right)^2 \cdot m = \frac{1}{m} \to \Theta(m)$

$\hookrightarrow \quad n \approx \Theta(\sqrt{m_2})$

Jan	26, 12
Feb	
Mar	
Apr	
May	23, 4
June	
July	
Aug	
Sept	
Oct	$\boxed{13}$, 24
Nov	24, 23
Dec	21

9

Pr[collision] > .5

after n = 23

$$\Pr[\text{collision, domain } [m], n \text{ steps}]$$

$$n = 1 \rightarrow \Pr = 0$$

$$n = 2 \rightarrow \Pr = \frac{1}{m}$$

$$n = 3 \rightarrow \Pr = \cancel{\left(\frac{1}{m}\right)^{\binom{n}{2}}} \approx \cancel{\left(\frac{1}{m}\right)^{n^2/2}}$$

When $n = m+1$

↳ most collide

pigeon hole principal

$$\binom{n}{2} \text{ pairs} \approx \frac{n^2}{2}$$

$$\Pr[\text{no collision}] = \left(1 - \frac{1}{m}\right)^{(n^2/2)}$$

$$\Pr[\text{collision}] \approx 1 - \left(1 - \frac{1}{m}\right)^{(n^2/2)}$$

set $n = \sqrt{2m}$

$$1 - \left(1 - \frac{1}{m}\right)^m \approx 1 - \frac{1}{e}$$

$$\Pr[\text{coll}] = 1 - \left(\frac{m-1}{m}\right)\left(\frac{m-2}{m}\right)\left(\frac{m-3}{m}\right)\cdots\left(\frac{m-(n-1)}{m}\right)$$

- Birthdays not iid

  mode Oct 5

- 366 (Feb 29)

- Twins

How many n until we see

<u>all</u> $j \in [m]$

? $n \geq m$ "Coupon Collectors"

m^2 $\quad m\sqrt{m}$ $\quad m \log m$

$$E[r_m] = \sum_{i=1}^{m} E[t_i] = \sum_{i=1}^{m} \frac{m}{m-i+1} = m \boxed{\sum_{j=1}^{m} \frac{1}{j}} = m\begin{pmatrix} 0.6 \\ + \\ \log m \end{pmatrix}$$

i $\quad j \quad j = m-i+1$

H_m

Harmonic #

$$H_m = 0.577 + \ln(m)$$

---

r_i = # trials until ith distinct item

$r_1 = 1$, $r_2 \approx 2$

<u>epoch</u> $t_i = r_i - r_{i-1}$

$$E[r_m] = E\left[\sum_{i=1}^{m} t_i\right] = \sum_{i=1}^{m} E[t_i]$$

$$E[t_i] = \frac{1}{p_i} = \frac{1}{\left(\frac{m-i+1}{m}\right)} = \frac{m}{m-i+1}$$

truth (μ)

sample (X)

error

$$d(\mu, Alg(X)) \leq \varepsilon$$

$$Pr[d(\mu, Alg(X)) > \varepsilon] \leq \delta$$

prob. failure

Probably Approximately Correct

(PAC)