

# Data Mining LZ

## Statistical Principles + Hashing

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

$X$  iid  $\mathcal{U}$  unknown distribution

iid: independent and identically distributed.

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

IP addresses  
all possible words  
# days in the year  
 $m = 365$

---

known distribution  $\mu = \text{uniform}$

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

Hashing (Hash Table)

# (Random) Hash Function

---

$h : \text{Domain} \rightarrow \text{Range}$   
 $\in^k \quad [m]$

deterministic

$\uparrow$  Sigma

Randomly select  $h_a \in \mathcal{H}$  ← family of hash functions

$$P_{h_a \in \mathcal{H}} [h_a(x) = h_a(y)] = \frac{1}{m} \quad x \neq y$$

- Built-in Hash function

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

$a = \text{salt}$

$$\text{Input } x \rightarrow \text{SHA-1}(\text{concat}(x, a))$$


---

- Multiplicative Hashing

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

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

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

$a$  large number

- Modular Hashing  $h(x) = x \bmod m$   
Do NOT USE

$$U \in \text{Unif}(0, 1)$$

$$\lfloor U \cdot m \rfloor \rightarrow j \in [m]$$

- How many samples  $x_1, x_2, \dots, x_n \in [m]$   
so avg two  $x_i = x_j$   
a collision

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

$$\hookrightarrow n \Rightarrow \Theta(\sqrt{m^2})$$

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

9

Pr[Collision]

$> .5$

after  $n = 23$

Pr [collision, domain  $[m]$ ,  $n$  steps]

$$n=1 \rightarrow P_r = 0$$

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

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

When

$$n = m+1$$

↳ most

collide

pigeon hole  
principal

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

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

$$P_r [\text{collision}] \approx 1 - \left(1 - \frac{1}{m}\right)^{\binom{m}{2}}$$

set

$$n = \sqrt{2m}$$

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

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

• Birthdays not coll

mode Oct 5

• 366 (Feb 29)

• Twins



How many  $n$  until we see

all  $j \in [m]$

?  $n \geq m$  "Coupon collector's"

$$m^2$$

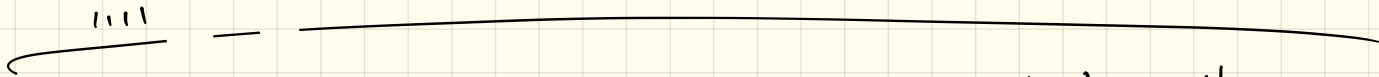
$$m\sqrt{m}$$

$$m \lg m$$

$$E[R_m] = \sum_{i=1}^m E[t_i] = \sum_{i=1}^m \frac{m}{m-i+1} = m \sum_{j=1}^m \frac{1}{j} = m \left( 0.6 + \log m \right)$$

$H_m$   
 Harmonic #

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



$$\begin{aligned}
 & E[R_m] \\
 &= E\left[ \sum_{i=1}^m t_i \right] \\
 &= \sum_{i=1}^m E[t_i]
 \end{aligned}$$

$r_i$  = # trials until  $i$ th distinct item

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

epoch  $t_i = r_i - r_{i-1}$

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

truth ( $\mu$ )

sample ( $X$ )

$$d(\mu, \text{Alg}(X)) \leq \epsilon$$

error  
↓

$$P_r[d(\mu, \text{Alg}(X)) > \epsilon] \leq \delta$$

Probably Approximately  
(PAC)

Correct  
↑  
prob. failure