L2 Statistical Principles & Hashing

Wednesday, January 8, 2020 2:58 PM

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

where $X \overset{iid}{\sim} U$ ← unknown distribution

↑ independent and identically distributed

To start, each $x_i \in [m] = \{1, 2, \ldots, m\}$

IP Addresses
all possible words
# of days in the year.

Here, U is a known distribution

U = uniform distribution

↳ $P[x_i = j] = 1/m \quad \forall j \in [m]$

↳ Appears in algos' data structures when Hashing (Hash table)

(Random) Hash Function:

h: Domain → Range
$\sum^k$ $[m]$

* deterministic (same string maps to same hash)

- Randomly select $h_a \in \mathcal{H}$ (family of hash functions)

$P[h_a(x) = h_a(y)] = \frac{1}{m}$ where $x \neq y$

3 versions of hash functions.

- Built-in hash functions
  - ↳ SHA-1 : $(\Sigma = \{0,1\})^k \rightarrow [m = 2^{160}]$

    a = salt

    ↳ input $x \rightarrow$ SHA-1(concat(x, a))

- Multiplicative Hashing
  - ↳ $h_a(x) = \lfloor m \cdot \text{frac}(x \cdot a) \rfloor$ where frac(17.32) = 0.32

    $= (xa / 2^q) \bmod m$ q large number

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

Ex.

1. $u \sim$ Unif $[0,1]$
2. $\lfloor u \cdot m \rfloor \rightarrow j \in [m]$

Q: How many samples $x_1, x_2, \ldots, x_n \in [m]$ do avg two $x_i = x_j$ (a collision)

A: $n \approx \Theta(\sqrt{m})$

Analyzing:

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

$n=1 \rightarrow P=0$
$n=2 \rightarrow P=\frac{1}{m}$
$n=3 \rightarrow P=\left(\frac{1}{m}\right)\left(\frac{3}{2}\right) \approx \left(\frac{1}{m}\right)^{n^2/2}$

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

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

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

set $n=\sqrt{2m}$

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

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

Problems with the analysis

1) Birthdays are not iid.
2) Some years have 366 days
3) Twins exist

$\Rightarrow$ but when $n=m+1$, there must exist a collision by the pigeon hole principle.

Q: How many n until we observe all $j \in [m]$? (coupon collectors problem)

A: $n = m \cdot \ln(m)$

Simulation Analysis

$E[r_m]$ where r_i = # of trials until ith distinct item.

$= E\left[\sum_{i=1}^{m} t_i\right]$ $\quad r_1 = 1,\ r_2 \approx 2$ and epoch $t_i = r_i - r_{i-1}$

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

where

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

Now, $E[r_m] = \sum_{i=1}^{m} E[t_i] = \sum_{i=1}^{m} \frac{m}{m-i+1} = m\sum_{j=1}^{m}\frac{1}{j}$ where $j = m-i+1$

$\approx m(0.6 + \ln(m))$

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

truth(M)

Sample(x)

$d(M, Alg(x)) \leq \varepsilon$ ← error

$P[d(M, Alg(x)) > \varepsilon] < \delta$

$$P[a(M, Alg(x)) > \epsilon] < \delta$$

↖ prob. failure

Probably Approximately Correct

(PAC)