L5 Locality Sensitive Hashing

Wednesday, January 29, 2020 2:03 PM

- Family of hash functions $\mathcal{H}$

↳ We want similar items to have a larger chance of collision.

$$P_{h \in \mathcal{H}} [h(p) = h(q)] \approx sim(p, q)$$

= Jaccard Triangle

≈ Euclidian (dot product)

1. 1 hash Function

$$\hat{JS}(p,q) = \begin{cases} 1 & h(p) = h(q) \\ 0 & ow \end{cases}$$

hash table ??

2. K hash functions

$$\hat{JS}_K(p,q) = \frac{1}{K} \sum_{j=1}^{K} I\{h_j(p) = h_j(q)\}$$

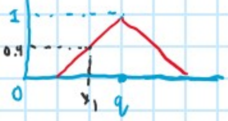
Algo??

Large Number of Objects

$X = \{x_1, x_2, \ldots, x_n\}$ (maybe documents from k-grams)

Q1: Which pairs are similar? (n^2 time)

Q2: Given a query 'q', which $x_i \in X$ are similar to q? (n time)



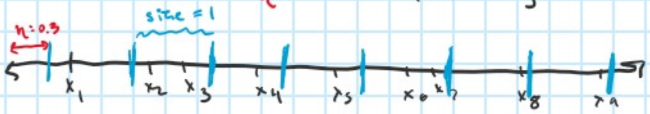
- Pretend $x_1, \ldots, x_n \in \mathbb{R}$

↳ Then similarity $S_\Delta(q, x_i) = \max\{0, 1 - |q - x_i|\}$

1. Sort $x_1, x_2, \ldots, x_n$
2. Build binary tree
3. Find q in tree

$\eta \in Unif(0,1)$

$h_\eta(x) \rightarrow$ a bin Runtime: $\log(n) + K$ # similar items



* two items collide if in the same bin

- It can be shown that $P[h(x) = h(x')] = S_\Delta(x, x')$

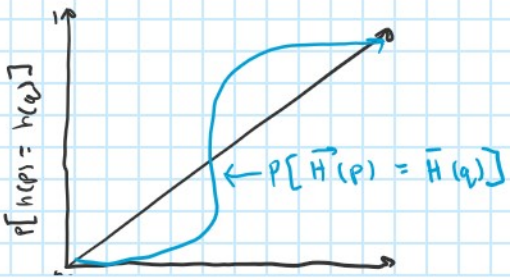
* Though we know there exists data structure solutions in 1D, we will explore the algorithmic solution for problems in higher dimensional spaces!

Banding: How to combine hash functions...

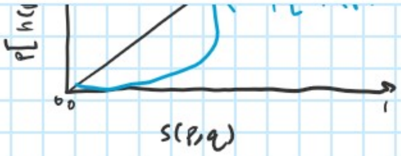
- $H = \{h_1, h_2, \ldots, h_t\} \in \mathcal{H}$
- $\vec{H}$ ← single super hash function.

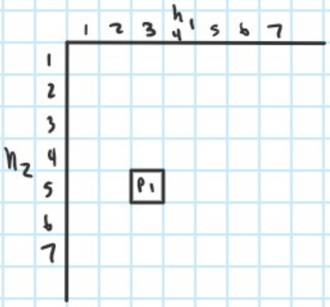
band b=2

	h_1	h_2	h_3	h_4	h_5	h_6
p_1	3	5	0	2	4	3
p_2	3	4	1	0	4	2
p_3	0	2	4	3	5	1
p_4						



p_4						
p_5						
p_6						
q	3	5	1	2	3	2



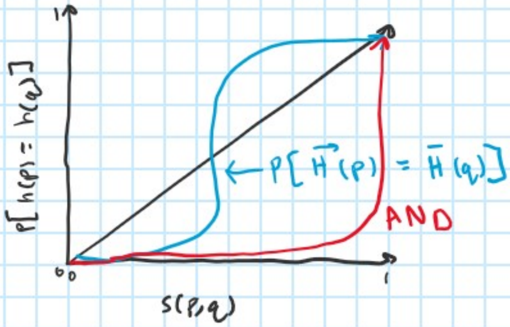


$p_1 = [3,5]$

$P_2[h_1,h_2(p) = h_1h_2(q)]$
$= s^2$

ie, much more selective

$\Rightarrow$



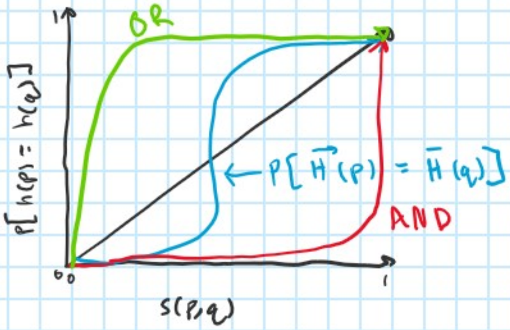
	h_1	h_2	h_3	h_4	h_5	h_6
p_1	3	5	0	2	4	3
p_2	3	4	1	0	4	2
p_3	0	2	4	3	5	1
p_4						
p_5						
p_6						
q	3	5	1	2	3	2

H_1 H_2 H_3

# of bands = r

$\vec{H}(p,q) = OR(H_1, H_2, H_3)$

$\Rightarrow$



Analysis:

- r bands, each with b hash functions.
  - $\hookrightarrow$ t = # of hash functions.
    - $\hookrightarrow$ $t \geq r \cdot b$
- $S(p,q) = s$
  - $\hookrightarrow$ s^b = Probability p, q collide in one band
- $(1-s^b)$ = Probability p, q do not collide.
- $(1-s^b)^r$ = Probability p, q don't collide in r bands.
- $f(s) = 1-(1-s^b)^r$ = Probability p, q collide in at least 1 band.
  - $\hookrightarrow$ r ↑ increases the OR count (→ ⌞)
  - $\hookrightarrow$ b↑ increases the AND count (← ⌞)

(✱) (time is used creating hash functions but lookup is then constant)

Q: How to choose values for b and r given t, T

S1. Plot f(s) (we want the most sleep line)

S2. $\begin{cases} b = -\log_T (t) \\ r = t/b \end{cases}$

LHS For Euclidean Distance!

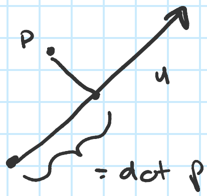
- $d_E(p,q) \iff s_E(p,q)$

$$= \langle p,q \rangle$$
$$= \sum_{j=1}^{d} p_i \cdot q_i$$

- $h_{n,u} : \mathbb{R}^d \rightarrow [m]$ where:

$\begin{cases} n \in \text{Unif}(0,T) & \text{(random shift)} \\ u \in \mathbb{R}^d \text{ s.t. } \|u\| = 1 & \text{(random unit vector)} \end{cases}$

- $h_{n,u}(p) = \left(\lfloor \langle p,u \rangle - n \rfloor \bmod m \right)$



(projecting $\vec{p}$ onto $\vec{u}$)

= dot product of p, u