L5: Locality Sensitive Hashing

Jeff M. Phillips

January 22, 2020
Family hash functions $H$

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

1. 1 hash function

$$\delta_S(p, q) = \sum \{ 1 \mid h(p) = h(q) \}$$

hash table

2. k hash functions

$$\delta_S(p, q) = \frac{1}{k} \sum_{j=1}^{k} \delta_S(p_j, q_j)$$

Also:

$$1(b) = \begin{cases} 1 & \text{if } b = \text{True} \\ 0 & \text{if } b = \text{False} \end{cases}$$

Also:

Euclidean (dot product)

Triangle

Jaccard

Indicator Function
Large number of objects \( X \)

\[ X = \{ X_1, X_2, \ldots, X_n \} \]

(documents, 17-grams, IP addresses, customers)

Q1: Which pairs are similar?

\( n^2 \) time

Q2: Given \( q \) going \( b \), which \( X_i \in X \) are similar to \( q \)?

\( n \) time
x_1, x_2, ..., x_n ∈ R

Similarity \quad \Delta_\theta (\theta, x_i) = \max \{0, 1 - \log(x_i)\}

1. Sort \ x_1, x_2, ...

2. Build binary tree \ T

3. Find \ g \ in \ T \quad \Pr[h(x) = h(x)] \geq \log n + \log \frac{n}{\# similar items}
Banding: How to combine hash functions

\[ H = \{ h_1, h_2, \ldots, h_3 \} \in X \]

\[ H \subseteq \text{single super hash function} \]

\[ P_1 \]
\[ 3 \]
\[ 3 \]
\[ 0 \]
\[ 1 \]
\[ 2 \]
\[ 3 \]

\[ P_2 \]
\[ 3 \]
\[ 4 \]
\[ 1 \]
\[ 0 \]
\[ 2 \]

\[ P_3 \]
\[ 3 \]
\[ 2 \]
\[ 4 \]
\[ 3 \]
\[ 5 \]

\[ P_4 \]
\[ 3 \]
\[ 5 \]
\[ 0 \]
\[ 2 \]
\[ 4 \]
\[ 1 \]

\[ P_5 \]
\[ 3 \]
\[ 3 \]
\[ 1 \]
\[ 2 \]
\[ 3 \]
\[ 2 \]

\[ \text{bands} \]
\[ = 5 \]

\[ H(\{3,3\}) = \text{OR}(H_1, H_2, H_3) \]
\[ P \rightarrow (3, 5) \]

\[ P_0 \left[ h_1 h_2 (p) = h_1 h_2 (q) \right] = s^2 \]

Much more selective
r bands, each with h hash functions

\( t = \# \text{hash functions} \quad t \geq rh \)

\[ S(p, b) = s \]

\[ s^b = \Pr \text{pig collide in one band.} \]

\[ (1 - s^b) = \Pr \text{pig don't collide.} \]

\[ (1 - s^b)^r = \Pr \text{pig don't collide in } r \text{ bands} \]

\[ f(s) = 1 - (1 - s^b)^r = \Pr \text{pig collide in at least one band.} \]
LSH $b = 3$ and $r = 5$

Probability of found collision $= 1 - (1 - s^b)^r$
LSH \( b = 3 \) and \( r = 15 \)

Probability of found collision \( = 1 - (1 - s^b)^r \)
LSH $b = 3$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$
LSH $b = 6$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$
LSH \( b = 6 \) and \( r = 15 \)

\[ t = 90 \]

Probability of found collision = \( 1 - (1 - s^b)^r \)
LSH $b = 10$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$
LSH $b = 10$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$
LSH $b = 8$ and $r = 100$

Probability of found collision $= 1 - (1 - s^b)^r$
LSH $b = 8$ and $r = 100$

Probability of found collision = $1 - (1 - s^b)^r$
LSH \((b = 3, r = 5) \& (b = 6, r = 15) \& (b = 8, r = 100)\)

Probability of found collision = \(1 - (1 - s^b)^r\)
LSH \((b = 3, r = 5) \& (b = 6, r = 15) \& (b = 8, r = 100)\)

Probability of found collision = \(1 - (1 - s^b)^r\)
LSH for Euclidean Dist.

\[ d_E(p, q) \iff s_E(p, q) \]

\[ \langle p, q \rangle \]

\[ = \sum_{i=1}^{m} p_i \cdot q_i \]

\[ h : \mathbb{R}^d \rightarrow [m] \]

\[ h_{\theta, u}(p) = \left \lfloor \frac{\langle p, u \rangle - \theta}{m} \right \rfloor \mod m \]

\[ u \in \text{unif}(0, 1) \]

\[ v \in \mathbb{R}^d, \quad \|v\| = 1 \]