Lecture 10: Locality Sensitive Hashing

- Near neighbor search
- Giving structure
- Conventional approaches - k-d tree
  - Exponential dependence on dimension $k$
- Hashing to the rescue

- Build a data structure
- Query garniver
  - $\text{return } a \in S$
    - $d(q, a) = \min_{a' \in S} d(q, a')$

Metric space $(X, d)$
$S \subseteq X$ 
Corpus distance
\[ S = \text{set of ships} \]

- Document = bag of words

\[ \begin{align*}
&\text{freq}(a) = 2 \\
&\text{freq}(b) = 2 \\
&\text{freq}(c) = 1 \\
&\text{freq}(d) = 0 \\
&\text{freq}(e) = 1 \\
\end{align*} \]

TF/IDF \( w = \frac{\text{freq}(w) \times \log \left( \frac{\text{#words in corpus}}{\text{#w appears in corpus}} \right)}{\text{avg doc}} \)

- \( X = \mathbb{R}^k \)

\[ d = \| x - y \|_2 = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \]

\[ a = \| x - y \|_1 \]

\[ = \sum_{i=1}^{k} |x_i - y_i| \]
\( k \text{-d tree} \)

Solution:
1. **generic dim \( \mathbb{R}^n \)** (lectures later)
2. **LSH**

\( k \) runtime curve of dimensionality
LSH vs Hashing

\[ S \subseteq U, \quad h : U \rightarrow \{ 0, 1 \} \]

\[ h(x) = h(y) \]

\[ \Pr [ h(x) \neq h(y) ] \text{ is high} \]

if \( x = y \)

\( x \neq y \)

---

distance between points

\( x, y \) close

\( x, y \) far

\[ \Pr [ h(x) = h(y) ] \text{ is high} \]

\[ \Pr [ h(x) \neq h(y) ] \text{ is high} \]
Solving Near Neighbor

$(X,d)$ metric space  $S \subseteq X  |S|=n$

Build a data structure

- on seeing queries $q \in X$
  - return $a \in S$

  $s.t. d(q,a) \leq (1+\varepsilon) \min_{a' \in S} d(q,a')$

- small space / preprocessing
- small query time

Given $\varepsilon, \delta \in \mathbb{R}^+$

Build ad struct

- for $q \in X$
  - if $\exists a \in S$ s.t. $d(q,a) \leq \varepsilon$
    - return a pt $a'$
    - $d(q,d) \leq (1+\varepsilon) R$
    - if no point at distance $\leq R$ from $q$
      - then either return $a'$ or $d(q,a) \leq (1+\delta) R$
  - or return NO point.
\text{LSH:}
\[d(x, y) \leq r \Rightarrow \Pr_{h \in H} [h(x) = h(y)] \geq P_{\text{close}}\]
\[d(x, y) \geq (1 + \varepsilon)r \Rightarrow \Pr_{h \in H} [h(x) = h(y)] \leq P_{\text{far}}\]

\[H \text{ hash family}\]
LSH for Hamming metric

\[ X = 30, 13^k \]

\[ d(x, y) = \text{# positions where they differ} \]

Weak LSH:

1. Pick a random bit position of \( x \),
   \[ h(x) = x_i \]

   if \( x, y \) close \( \implies d(x, y) \leq r \)
   far \( \implies d(x, y) \geq 2r \)

   \[ \Pr[h(x) = h(y)] = 1 - \frac{d(xy)}{k} \]

   \[ \Pr(\text{collision}) \geq 1 - \frac{r}{k} \]

   \[ \leq 1 - \frac{2r}{k} \]

Metic dependent

\[ e = 1 \]
Step 2: Parallel Repetition

\[ H' \]

\[ h'(x) = (h_1(x), h_2(x), \ldots, h_t(x)) \]

\[ \Pr[h'(x) = h'(y)] \leq \beta^t \]

\[ \Pr(\text{collision under } h') \geq (1 - \frac{\beta}{K})^t \leq e^{-\frac{\beta t}{K}} \leq \frac{1}{\sqrt{n}} \]

\[ \Pr(\text{collision under } h') \leq (1 - \frac{2\beta}{K})^t \leq e^{-\frac{2\beta t}{K}} \leq \frac{1}{n} \]

\( \iff t = \frac{K}{2\beta} \cdot \log n \)