Last time

- short reviews from probability
  - variance & tail bounds
  - Gaussians & CLT
- Distinct sets
- Similarity search & dimension reduction

Today

Locality sensitive hashing (LSH)
⇒ efficient approx. similarity search

Similarity search, NNS

large collection of data items
- \( x^{(1)}, x^{(2)}, \ldots, x^{(n)} \in \mathbb{R}^k \)
- notion of distance (or equivalently similarity)
  \( d^*(x, y) = \text{dist between } x \text{ & } y \)

2 problems:

1) Find all pairs \( i, j \) s.t.
   \[ d(x^{(i)}, x^{(j)}) \leq r \]

   Naive:
   \( O(n^2k) \)

2) Preprocess database \( \mathcal{D} \)
   from efficiently respond to queries
   \( \text{query } q \in \mathbb{R}^k \)
   return all pts in \( \mathcal{D} \) s.t.
   \[ d(q, x^{(i)}) \leq r \]

   \( O(nk) / \text{query} \)

If \( k \) small, say \( \leq 15 \), there are
space partitioning data structures that are reasonably efficient
- \( k \)-d trees
curse of dimensionality.
**LSH (c-approx r-nearest neighbors problem)**

Think $c=2$, $c > 1$

Given query pt $q$, return
- all pts $x^{(i)}$ s.t.
  
  $$d(x^{(i)}; q) \leq r$$
  
  (w.h.p.)
- may return some pts $x^{(i)}$
  
  s.t.
  
  $$d(x^{(i)}; q) \leq c \cdot r$$

**Primitives**

$X$: family of hash funs that map pts $\in \mathbb{R}^k \rightarrow$

$X$ is $(r, cr, p_1, p_2)$-sensitive if $\forall x, y \in \mathbb{R}^k$

If $d(x, y) \leq r$ \implies $\Pr_{h \in X} (h(x) = h(y)) \geq p_1$

If $d(x, y) \geq cr$ \implies $\Pr_{h \in X} (h(x) = h(y)) \leq p_2$

$p_2 < p_1$

**Example:** Suppose pts $e \{0, 1\}^k$

$\{0, 1\}^k \rightarrow \{0, 1\}^k$

$X = \{h(x) = x_i \mid 1 \leq i \leq k\}$

$d(x, y) \leq r$

$p_1 = \Pr(h(x) = h(y)) \geq 1 - \frac{c \cdot r}{k} \approx e$

$x = x_1, x_2 \ldots x_k$

$y = y_1, y_2 \ldots y_k$

$d(x, y) > cr$

$p_2 = \Pr(h(x) = h(y)) \leq 1 - \frac{c \cdot r}{k}$

$e^{-\frac{c \cdot r}{k}}$

$(r, cr, \frac{1}{k}, 1 - \frac{c \cdot r}{k})$

family of hash funs.
Combine these hash fins to amplify difference between p, & q:

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

**Purpose:** reduce chance that pts that are far away map to some value.

Amplify \( g(y) \) \( \Delta(0,y) \geq \epsilon \) \( p \leq d \)

\[ g_1(x) = [h_1(x), h_2(x), \ldots, h_n(x)] \]
\[ g_2(x) = [h_1(x), h_2(x), \ldots, h_n(x)] \]
\[ g_3(x) = [h_1(x), h_2(x), \ldots, h_n(x)] \]

1. \( q(x) \) \( \Delta(q,y) \geq \epsilon \)
2. \( \Pr(g(qp)=g(y)) \leq p \)
3. \( S_q = \{ x \mid g_3(x)=q(x), 0 \leq x < \epsilon \} \)
4. \( d(q,y) \) \( \forall x \in S_q \) compute all of those that have \( d(q,y) \leq r \)

\[ d(x,y) \leq r \]
\[ \Pr(x \notin S_q) = \frac{(Pr(Y \notin S_q))^t}{1 - Pr(Y \notin S_q)} \]
\[ \leq (1 - p)^t = e^{-pt^t} \]

\[ d = \frac{\log n}{\log(1/p^t)} \]

**Where we are:**

- **Preprocessing time:** \( n \cdot d \) hash fn computation
- **Space:** \( n \cdot d + \text{actual pts} \)
- **Exp time to process query:**
  \[ ld (\text{hash fn computation}) + E(\# \text{far pts in question}) \]
  \[ \leq l \cdot n \cdot p \]
- **Prob miss a close pt** \( \Rightarrow \)
  \( (1 - p^t)^t \)

\( X \) is \( (r, \epsilon, p, q) \)-sample of \( \forall x \geq \epsilon \) \( k \)

- If \( d(x,y) < r \) \( \Rightarrow \) \( \Pr_{i=1}^k (h_i(x), h_i(y)) = p_i \)
- If \( d(x,y) > r \) \( \Rightarrow \) \( \Pr_{i=1}^k (h_i(x), h_i(y)) = p_{i-1} \)

- **dist fn.
  - Start w/ \( r \)
  - get \( X \) \( (r, \epsilon, p, q) \)
  - need to select
  - \( \epsilon \)
  - \( d \)

\[ \text{set } n_{\text{bad pt}}^d = 1 \]

\[ e \leq 1 \text{ bad pt per table} \]

\[ \frac{d}{\log(1/p^t)} \]

\[ \Pr(\text{missing close pt}) = \frac{1}{2} \]

\[ p^t \cdot e = 1 \]

\[ \Rightarrow \]

\[ l = n^d \]

\[ l = n \]
Where we are:
- Preprocessing time: $n \cdot l \cdot d$ hash fn computing
- Space: $n \cdot d +$ actual pts
- Exp time to process query:
  \[
  l d (\text{hash fn computing}) + E(\# \text{far pts that are in } S_g)
  \leq l \cdot n \cdot d^2.
  \]
- Prob miss a close pt:
  \[
  \Rightarrow (1 - p_i^d)^k.
  \]

\[
\frac{\log \left( \frac{1}{p_i} \right)}{\log \left( \frac{1}{p_2} \right)} = \delta
\]

\[
\left\{ \begin{array}{ll}
p_i = \frac{1}{4} & \text{if } \frac{\sqrt{n}}{4} \leq \alpha \\
p_i = \frac{1}{2} & \text{if } \frac{\sqrt{n}}{2} \leq \alpha \\
\end{array} \right.
\]

\[l = n \]

\[
\frac{n^g}{\log \left( \frac{n^g}{\log \left( \frac{n^g}{n^g + \rho} \right)} \right)}
\]

\[
\frac{n^g}{\log \left( \frac{1}{\rho} \right)} + n^g
\]

\[
\frac{\alpha}{\epsilon}.
\]

\[
\text{Cosine similarity}
\]

\[
\left( x^{(1)}, \ldots, x^{(m)} \right) \in \mathbb{R}^k
\]

\[
\text{similar (close) if } \frac{6 \leq 0.1}{\theta \geq 0.2} \leq \frac{\theta}{\epsilon}
\]

\[
\left( L = \left\{ h(x) = \text{sign}(r \cdot x) \right\} \right.
\]

\[
h : \mathbb{R}^k \rightarrow \{0, 1\}
\]

\[
\text{sign}(r \cdot x) = \left\{ \begin{array}{ll}
1 & \text{if } r \cdot x > 0 \\
0 & \text{if } r \cdot x < 0
\end{array} \right.
\]

\[
\theta = \frac{\epsilon}{\alpha}
\]

\[
r \cdot x = \| r \| \cdot \| x \| \cos \theta
\]

\[
\text{cos }\theta
\]
Jaccard Similarity
\[ J(x, y) = \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} \max(x_i, y_i)} \]
\[ = \frac{\text{intersection}}{\text{union}} \]

\[ x = (x_1, \ldots, x_k) \]
\[ y = (y_1, \ldots, y_k) \]
\[ x_i = \begin{cases} \# \text{occurrence} \text{ of } i \text{ in } x & \text{if } i \text{ is in both } x \text{ and } y \\ 0 & \text{o.w.} \end{cases} \]

\[ x = (1, 2, 3, 4, 5, 6, 7) \]
\[ y = (0, 1, 0, 0, 1, 0) \]
\[ \Pi = \{1, 2, 4, 7\} \]
\[ h_{\Pi}(x) = (0, 1) \]
\[ h_{\Pi}(y) = (1, 0) \]
\[ \text{Pr}(h_{\Pi}(x) = h_{\Pi}(y)) = \frac{|\Pi|}{|A \cup B|} \]

\[ \Pi' = \{1, 2, 4\} \]
\[ h_{\Pi'}(x) = 4 \]