Last time

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

Today

Locality sensitive hashing (LSH)
$\Rightarrow$ 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)
  $d(x,y)$ = dist between x & y

2 problems:

1. Find all pairs i j s.t. $d(x^{(i)}, x^{(j)}) \le r$

   Naive: $O(n^2 k)$

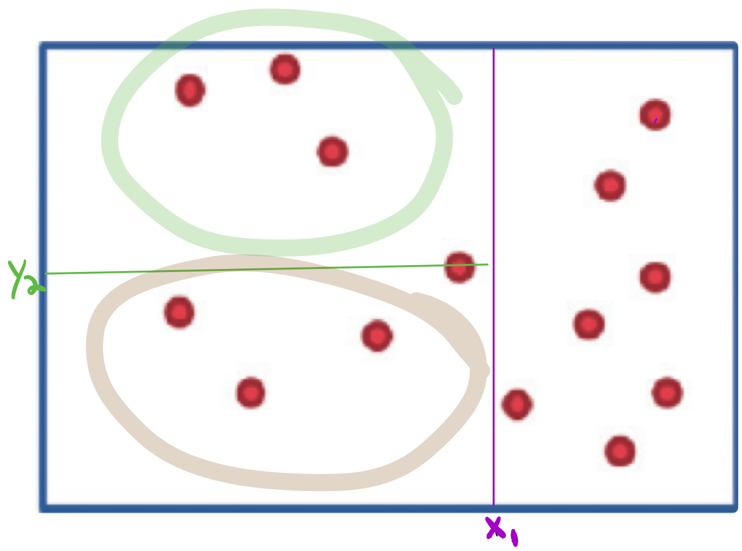
2. Preprocess database & then efficiently respond to queries

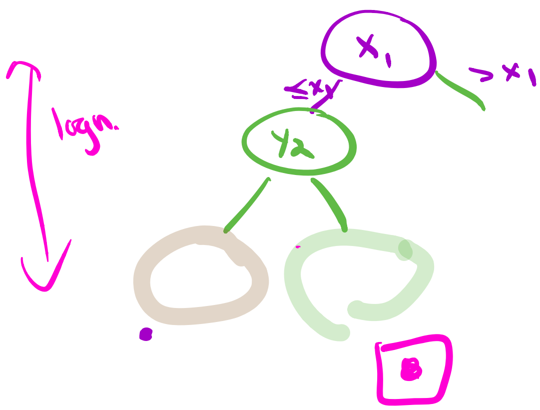
★ query $q \in \mathbb{R}^k$
return all pts $x^{(i)}$ in DB s.t. $d(q, x^{(i)}) \le r$

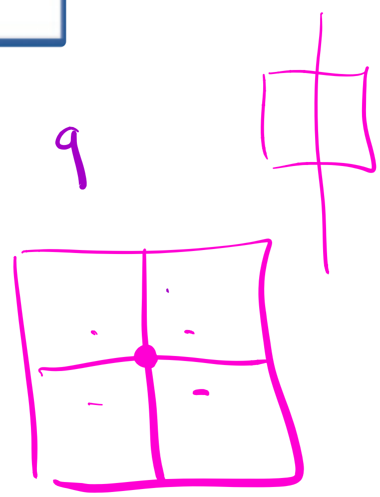
$O(nk)$/query

If k small, say < 15, there are space partitioning data structures that are reasonably efficient

k-d trees





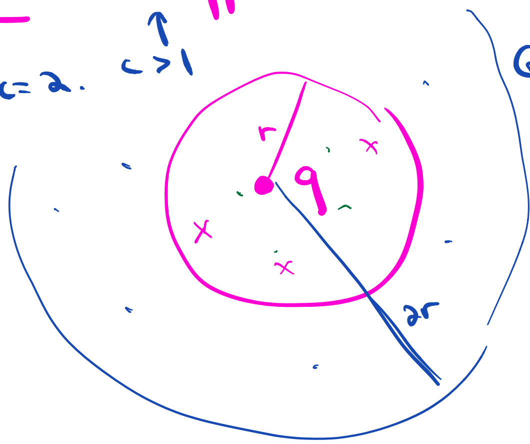


curse of dimensionality.

LSH

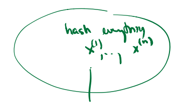
c-approx r-nearest neighbors problem

think $c=2$. $c>1$



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

$d(x,y) = \begin{cases} 0 & \text{identical } x=y \\ \infty & \text{o.w.} \end{cases}$



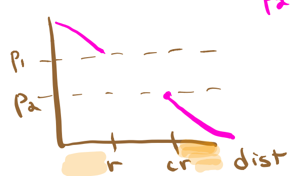
Primitive

$\mathcal{H}$: family of hash fns that map pts $\in \mathbb{R}^k \longrightarrow$

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

$\left.\begin{array}{l} \text{If } d(x,y) \le r \Rightarrow \Pr_{h\in\mathcal{H}}(h(x)=h(y)) \ge p_1 \\ \text{If } d(x,y) \ge cr \Rightarrow \Pr_{h\in\mathcal{H}}(h(x)=h(y)) \le p_2 \end{array}\right\}$ defn.

$p_2 < p_1$



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

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

$\mathcal{H} = \{ h(x) = x_i \mid 1 \le i \le k \}$

$d(x,y) = \text{Hamming dist} = \sum_{i=1}^{k} \mathbb{1}_{x_i \ne y_i}$

(# bits where different)

$d(x,y) \le r$

$p_1 = \Pr(h(x)=h(y)) \ge 1 - \frac{r}{k} \approx e^{-\frac{r}{k}}$

$d(x,y) \ge cr$

$p_2 = \Pr(h(x)=h(y)) \le 1 - \frac{cr}{k} \approx e^{-cr/k}$

$x = x_1 \boxed{x_2} \cdots \boxed{x_k}$

$y = y_1 \boxed{y_2} \cdots \boxed{y_k}$

$\left(r, cr, 1-\frac{r}{k}, 1-\frac{cr}{k}\right)$ family of hash fns.

combine these hash fns to amplify difference between p_1 & p_2

$h(x) \in \{0,1\}$
$g(x) \in \{0,1\}^d$

Where we are:
- preprocessing time: $n \cdot \ell \cdot d$ hash fn computations
- space: $n \cdot \ell$ + actual pts
- exp time to process query: ℓd (hash fn computations) $+ \underbrace{E(\#\text{ far pts} > cr \text{ that are in } S_q)}_{\le \ell \cdot n \cdot p_2^d}$
- Prob miss a close pt $\Rightarrow (1-p_1^d)^\ell \to 1$

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

If $d(x,y) \le r \Rightarrow \Pr_{h \in \mathcal{H}}(h(x)=h(y)) \ge p_1$

If $d(x,y) \ge cr \Rightarrow \Pr_{h \in \mathcal{H}}(h(x)=h(y)) \le p_2$

$p_2 < p_1$

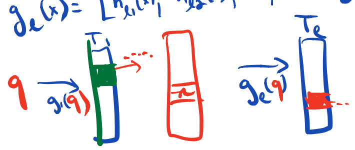
① $g(x) = [h_1(x), h_2(x), \dots, h_d(x)]$

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

AND step q, y $d(q,y) \ge cr$

$\Pr(g(q)=g(y)) \le p_2^d$

②
$g_1(x) = [h_{11}(x), h_{12}(x), \dots, h_{1d}(x)]$
$g_2(x) = [h_{21}(x), h_{22}(x), \dots, h_{2d}(x)]$
$\vdots$
$g_\ell(x) = [h_{\ell 1}(x), h_{\ell 2}(x), \dots, h_{\ell d}(x)]$



dist fn,
start w/ r, c
get $\mathcal{H}$ (r, cr, p_1, p_2)

③ On query q, compute $g_1(q), \dots, g_\ell(q)$

OR $S_q = \{x^{(i)} \mid g_j(x^{(i)}) = g_j(q) \text{ for some } 1 \le j \le \ell\}$

need to select
- d
- ℓ

④ compute $d(q,x)$ $\forall x \in S_q$
output all of those that have $d(q,x) \le r$

$d(x,q) < r$

$\Pr(x \notin S_q) =$

$\Pr(g_1(x) \ne g_1(q), g_2(x) \ne g_2(q), \dots, g_\ell(x) \ne g_\ell(q))$

$= (\Pr(g_1(x) \ne g_1(q)))^\ell$

$1 - \Pr(g_1(x) = g_1(q))$

$\Pr(h_1(x) = h_1(q)), \dots h_{1d}(x) = h_{1d}(q))$

$\le (1 - p_1^d)^\ell \approx e^{-p_1^d \cdot \ell}$

set $n p_2^d = 1$

$\equiv \le 1$ bad pt per table

$$\boxed{d = \frac{\log n}{\log\left(\frac{1}{p_2}\right)}}$$

$\Pr(\text{missing close pt}) = \frac{1}{e}$

$p_1^d \cdot \ell = 1 \Longrightarrow$

$$\boxed{\ell = n^{\boxed{\frac{\log\left(\frac{1}{p_1}\right)}{\log\left(\frac{1}{p_2}\right)}}}}$$

$$\boxed{d = \frac{\log n}{\log\left(\frac{1}{p_2}\right)}}$$

$\ell = n^\rho$

$$\boxed{\frac{\log\left(\frac{1}{p_1}\right)}{\log\left(\frac{1}{p_2}\right)} \triangleq \rho}$$

$$\boxed{p_1 = \tfrac{1}{2} \quad p_2 = \tfrac{1}{4}}$$

$$\ell = n^{1/2}$$

Where we are:

- preprocessing time: $n \cdot \ell \cdot d$ hash fn computations $\Longleftarrow$ $n \cdot n^{\rho} \cdot \log n / \log\left(\frac{1}{p_2}\right)$
- space: $n \cdot \ell$ + actual pts $\quad n^{1+\rho}$ + pts
- exp time to process query:
  - ℓd (hash fn computations)
  - $+ \underbrace{E(\#\text{ far}_{>cr}\text{ pts that are in } S_q)}_{\le \ell \cdot n \cdot p_2^d}$ $\to 1$

$$\boxed{\underline{n^{\rho} \log n / \log(1/p_2)} + n^{\rho}}$$

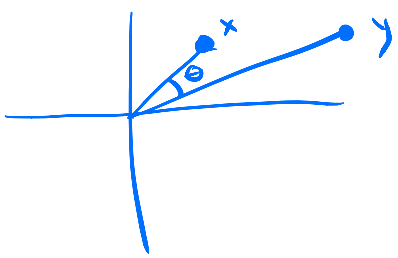
- Prob miss a close pt
  $\Rightarrow (1 - p_1^d)^{\ell}$ $\quad \boxed{\frac{1}{e}.}$

Cosine similarity

$x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^k$

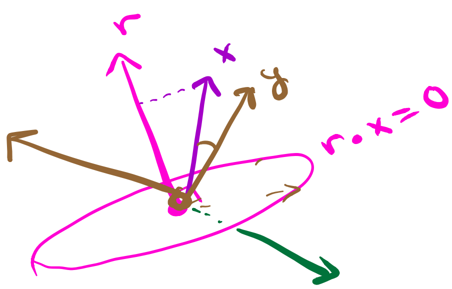
similar (close) if $\boxed{\theta \le 0.1}$ $\ge p_1$

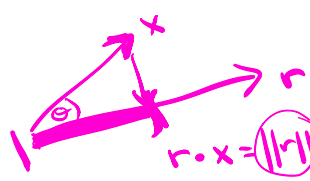
far $\theta \ge 0.2$ $\le p_2$



$$\left[\ \mathcal{H} = \left\{ h(x) = \underline{\text{sign}(\underbrace{r \cdot x}_{\sum_{i=1}^{k} r_i x_i})} \;\middle|\; \begin{array}{l} r = (r_1, \dots, r_k) \\ r_i \sim N(0,1) \\ \text{indep} \end{array} \right\}\right.$$

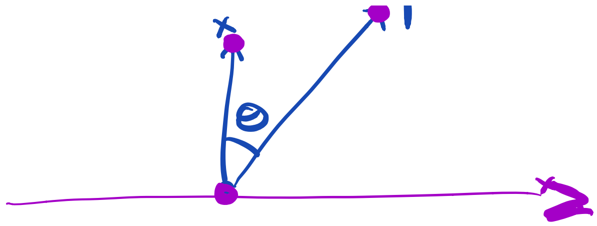
$h: \mathbb{R}^k \to \{0,1\}$





$r \cdot x = \|r\| \|x\| \cos\theta$

$$\text{sign}(r \cdot x) = \begin{cases} 1 & \text{if } r \cdot x \ge 0 \\ 0 & r \cdot x < 0 \end{cases}$$

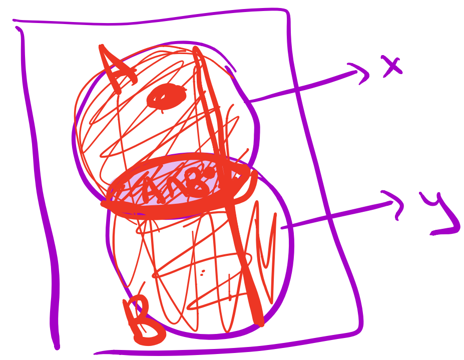


Jaccard Similarity

$x = (x_1, \dots, x_k)$

$$x_i = \begin{cases} \#\text{occurrences word } i \text{ is in doc} \\ 0 \quad \text{o.w.} \end{cases}$$

$$J(x,y) = \boxed{\frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}} = \frac{|\text{intersection}|}{|\text{union}|}$$



$$\begin{array}{lccccccc} & 1 & 2 & \boxed{3} & 4 & 5 & 6 & 7 \\ x = (& 0 & 0 & 1 & 1 & 0 & 1 & 0) \\ \hline \pi = & 5 & \boxed{3} & 2 & 4 & 7 & 1 & 6 \end{array}$$

$h_\pi(x) = 0 \ \boxed{1}$

$$Pr(h_\pi(x) = h_\pi(y)) = \frac{|A \cap B|}{|A \cup B|}$$

$\pi' = 7 \ 2 \ 4 \dots$

$h_{\pi'}(x) = 4$