Course: Comtemporary Algorithms T.II/2019-20

Lecture 4: Nearest Neighbours II

15 January 2020

Lecturer: Dr. Kanat Tangwongsan *Scribe: Kanokpon & Kanokpon*

Input: a collection $D = \{P_1, P_2, P_3, ...P_n\} \subset \chi$ of n points
a dist function $d : \chi * \chi \to \mathbb{R}_+ \cup \{0\}$

Query: given a query $q \in x$ find (all)points near q

Two ingredients in the above formulation:

1. a (dis)aimilarity measure.
2. an efficient d/s & algo baseline: $O(n)$ probes

Similarity = -dist

1 Euclidian distance (l_2 distance)

$$\begin{aligned}
\vec{x}, \vec{y} &\in \mathbb{R}^d \\
d(\vec{x}, \vec{y}) &= ||x - y||_{l_2} \\
&= \sqrt{(\vec{x} - \vec{y})^T(\vec{x} - \vec{y})} \\
&= \sqrt[2]{\sum_{i=1}^{d}(x_i - y_i)^2} \\
\text{For } p > 0, ||x - y||_{l_p} &= \sqrt[p]{\sum(x_i - y_i)^p} \\
p = 0 &: \text{How many non-zero coordinate } ||\vec{x} - \vec{y}||_{l_0} \\
p = \infty &: \max x_i - y_i \\
p = 1 &: \text{Manhattan Dist}
\end{aligned}$$

2 Jaccard

$$S, T \subseteq u$$
$$J(S, T) = \frac{|S \cap T|}{S \cup T}$$

3 Cosine Similarity

$$\frac{\langle x, y \rangle}{||x||\ ||y||}$$

Angular distant = $\cos^{-1}(\frac{\langle x,y \rangle}{||x||\ ||y||})$

4 Low-dimensional Space

4.1 d = 1

- balance search tree
- sorted array
- $O(\log n)$ time/query
- $O(n)$ space
- $O(n \log n)$ preprocessing

4.2 d = 2

- Voronoi
- KD tree
- OCT tree

Voronoi
For each point in D, construct (and store) the region that contains all of its closest neighbors

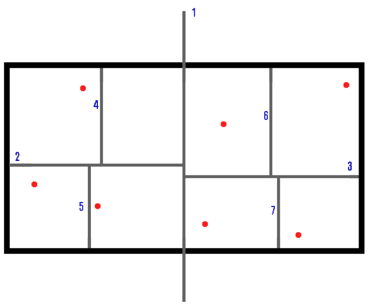
$$VR_i = \{x \in \chi | d(x_i, p_i) \leq d(x_j, p_j) \forall j \neq i\}$$

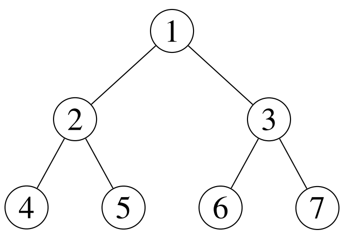
- $n \log n$ time to build
- $O(n)$ space
- $O(\log n)$ query(via point location)

KD trees

- space partitioning technique
- alternate b/w vertical and horizontal cuts
  - find a median split point

– recurse until 1 point
– yield a BST





Rectangle range query
find all points of D in $D \cap R$

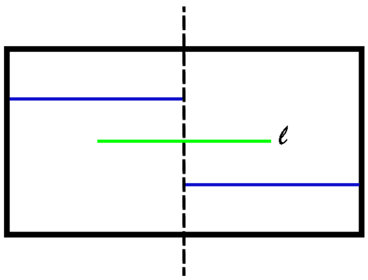
Start at root & recurse

case i: This node region $\subseteq R$
return everything in this subtree

case ii: this node's region $\cap R = \emptyset$
return $\emptyset$

case iii: this node's region $\cap R \neq \emptyset$
recurse & return the union of the answers from2 subtrees

Total query cost = $\overbrace{\text{\# nodes visited}}^{O(\sqrt{n})} + \overbrace{\text{\# pts reported}}^{s}$
<u>Observation:</u> #nodes visited $\leq$ # nodes where regions are intersected by an edge of R



$$f(n) \leq 3 + 2f(\frac{n}{4})$$
$$f(n) \leq O(\sqrt{n})$$

- space: $1 + 2 + 2^2 + ... + 2^{\log n} = O(n)$
- preprocessing: $O(n \log n) \Leftarrow n \log n$(sorting) + $[T(n) = 2T(n/2) + O(n)]$
- query: $O(\sqrt{n} + s)$

4.3 d > 2

- Voronoi
  - has $\Omega(n^{\lceil \frac{d}{2} \rceil})$ sites
  - $O(n^d)$ space, $(d + \log n)^{O(1)}$ query
- KD-tree
  - query: $O(dn^{1-\frac{1}{d}} + s)$ [still prefer if $d \leq 50$]

5 Aim

#1 Dimensionality reduction JL

#2 Hash & Approximate LSH

#3 Data-dependent schemes PCA

5.1 Johnson-Lindenstrauss

JL-Lemma states that any n points in high dimensional euclidian space can be mapped onto k dimensions where $k \geq O(\frac{1}{\varepsilon^2} \log n)$ without distorting the euclidian distance between any two more than a factor

Find a mapping $f : u^D \to V^d$ s.t. (embedded space) $d_V(f(x), f(y)) \in (1 \pm \varepsilon) d_u(x, y)$

Lemma 1 **(Johnson-Lindenstrauss)**

Let $0 < \varepsilon < \frac{1}{2}$ Given any set of points
$\{p_1, p_2, \ldots, p_n\} \subseteq \mathbb{R}^D$ there exists a mapping $f : \mathbb{R}^D \mapsto \mathbb{R}^k$
with $k = O(\frac{1}{\varepsilon^2} \log n)$ such that $||f(p_i) - f(p_j)||_2^2 \in (1 \pm \varepsilon)||p_i - p_j||_2^2$

$$M_X = \begin{bmatrix} M_1 \\ \vdots \\ M_k \end{bmatrix}_{K \times D} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad ; M_{n,i} \sim N(0,1) \tag{1}$$

$$= \begin{bmatrix} \langle M_1, x \rangle \\ \vdots \\ \langle M_k, x \rangle \end{bmatrix} \tag{2}$$

$$f(x) = \frac{1}{\sqrt{k}} M_x \tag{3}$$

Lemma 2 **(Distributional JL)**

Let $0 < \varepsilon < \frac{1}{2}$ If f is constructed
per above with $k = 8\varepsilon^{-2} \ln \frac{2}{\delta}$ with $x \in \mathbb{R}^D$, with $||x||_2 = 1$ then
$\mathbf{Pr}[||f(x)||_2^2 \in 1 \pm \varepsilon] \geq 1 - \delta$

$$
\begin{aligned}
\Delta &= f\left(\frac{\overbrace{P_i - P_j}^{x}}{||p_i - p_j||}\right) && (4)\\
&= \frac{1}{\sqrt{k}} M(\cdots) && (5)\\
&= \frac{1}{||p_i - p_j||}(f(p_i) - f(p_j)) && (6)\\
||\Delta||_2^2 &= \frac{1}{||p_i - p_j||_{l2}^2}||f(p_i) - f(p_j)||_2^2 && (7)\\
& && (8)\\
\text{Let } x &= \frac{P_i - P_j}{||p_i - p_j||} && (9)\\
||x||_2 &= 1 && (10)\\
\frac{||f(p_i) - f(p_j)||_2^2}{||p_i - p_j||_2^2} &= ||\Delta||_2^2 \in 1 \pm \varepsilon \quad \text{Proved Lemma 1} && (11)\\
& && (12)\\
\text{Let } \delta &= \frac{2}{n^2} && (13)\\
k &= \frac{8}{\varepsilon^2} \ln \frac{2}{\frac{2}{n^2}} && (14)\\
&= \frac{8}{\varepsilon}^2 \ln n^2 && (15)
\end{aligned}
$$

for pair (i, j), $\mathbf{Pr}\left[\Delta \notin 1 \pm \varepsilon\right] < \frac{1}{n^2}$
$\binom{n}{2}$ pairs total $\Rightarrow \mathbf{Pr}$ [at least one has $\Delta \notin 1 \pm \varepsilon$]
$\overset{\text{union}}{\text{bound}} < \binom{n}{2}\frac{1}{n^2} \leq \frac{1}{2}$
w.p. $\geq \frac{1}{2}$, all pairs are good

$A(x) = \frac{1}{\sqrt{k}} M_x$

matrix $M_{K \times D}$, $N_{i,j}$ $(0, 1)$

Let $0 < \varepsilon < \frac{1}{2}$ If A is as defined above with $k = 8\varepsilon \ln(\frac{2}{\delta})$ and $x \in \mathbb{R}^D$ with$||x||_2 = 1$ then, $\mathbf{Pr}\left[||A(x)||_2^2 \in 1 \pm \varepsilon\right] \geq 1 - 8$

$N(\mu, \sigma^2)$ mean μ and variance σ^2

pdf: $f(x) = \frac{1}{2\pi\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$

then:

$$
\begin{aligned}
CG_1 &\sim N(c\mu_1, c^2\sigma_1^2) \\
G_1 + G_2 &\sim (\mu_1 + \mu_2, \sigma_1^2, \sigma_2^2) \\
A(x) &= \begin{bmatrix} M_1 \\ \vdots \\ M_k \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix} = \frac{1}{\sqrt{k}} \begin{bmatrix} < M_1 x > \\ < M_2 x > \\ \vdots \\ < M_k x > \end{bmatrix} \begin{matrix} -y_1 \\ -y_2 \\ \vdots \\ -y_K \end{matrix} \\
y_i &= [G_1, G_2, \ldots, G_2] \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix} = \sum_{j=1}^{D} G_j x_j, \text{where } G_j \sim N(0,1) \\
y_i &\sim N(0, \underbrace{||x||_2^2}_{=1}) = N(0,1) \\
Z &= ||A(x)||_2^2 = A(x)^T A(x) = \frac{1}{k} \sum y_i^2 \to \text{chi}^2 \text{ disttribution with } k \text{ doF} \\
\mathbb{E}\left[Z\right] &= \frac{1}{k} \sum \mathbb{E}\left[y_i^2\right] = 1 = ||x||_2^2 \\
\mathbb{E}\left[||A(p) - A(q)||_2^2\right] &= \mathbb{E}\left[||A(p-q)||_2^2\right] = ||p-q||_2^2
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{Pr}\left[Z > 1 + \varepsilon\right] \forall t &> 0 \\
&= \mathbf{Pr}\left[e^{tkz} \geq e^{tk(1+\varepsilon)}\right] \\
&\leq \frac{\mathbb{E}\left[e^{tkz}\right]}{e^{tk(1+\varepsilon)}} \qquad ; e^{tkz} = e^{tk\frac{1}{k}\sum y_j^2} = \Pi e^{ty_j^2} \\
&= \Pi \frac{\mathbb{E}\left[e^{ty_j^2}\right]}{e^{tk(1+\varepsilon)}}
\end{aligned}
$$

claim1: $\mathbb{E}\left[e^{tG^2}\right] = \frac{1}{\sqrt{1-2t}}$ for $t < \frac{1}{2}$

claim2: $\frac{1}{e^t\sqrt{1-2t}} \leq e^{\frac{t^2}{1-2t}}$

$$\begin{aligned}
&= \left(\frac{1}{e^{t(1+\varepsilon)}\sqrt{1-2t}}\right)^k \\
&\leq exp\left\{\frac{kt^2}{1-2t} - k + \varepsilon\right\} \\
\text{use } t &= \frac{\varepsilon}{4} \leq e^{\frac{-k\varepsilon^2}{8}} \\
\text{Recall } \varepsilon &< \frac{1}{2} \\
\text{choose } k &= \frac{8}{\varepsilon^2}\left(\frac{2}{\delta}\right) = \frac{\delta}{2} \\
\mathbf{Pr}\left[Z < 1-\varepsilon\right] &< \frac{\delta}{2}
\end{aligned}$$