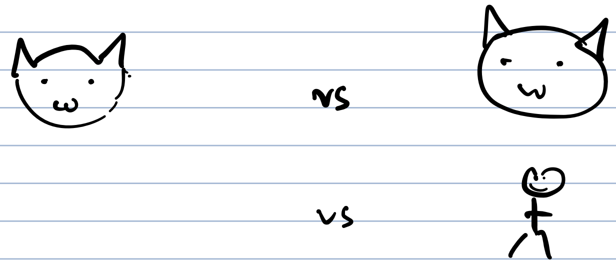
Lecture 3: metric data

- So far, we haven't really thought about <u>structure</u> of data.
- But usually this is what actually matters!

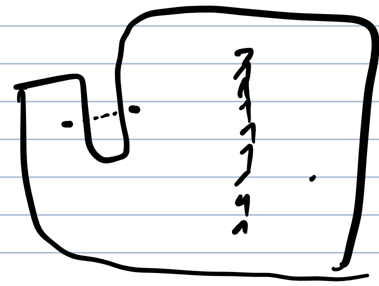


Q: What is the most similar data point to mine?

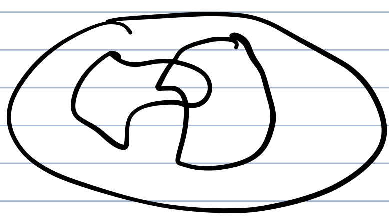
<u>Q</u>: How similar are two data points?

High level question: <u>distance between data</u>

- points on a map
  - Euclidean ?



- Sets of things



distance = ?

<u>Metric spaces</u>: A metric space is specified by

$$(X, d)$$

X: a collection of points

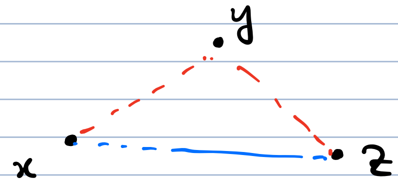
d: a distance function $d: X \times X \to \mathbb{R}_{\geq 0}$.

d must satisfy 3 properties:

(i) $d(x,y) = 0 \Rightarrow x = y. \quad \forall x, y$

(ii) (Symmetry) $d(x,y) = d(y,x) \quad \forall x, y.$

(iii) (Triangle inequality)



$d(x,z) \le d(x,y) + d(y,z).$
$\forall\ x, y, z.$

examples of metric spaces:

(0) ℓ_p - distances. $X = \mathbb{R}^d$

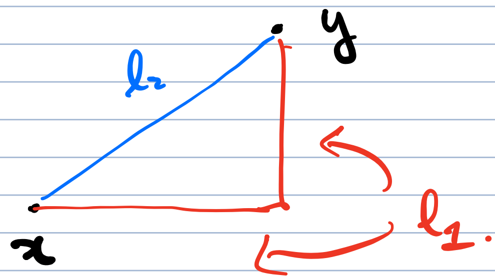
$$d(x,y) = \|x-y\|_p = \left(\sum_{i=1}^{d} |x_i - y|^p\right)^{1/p}$$

$(p \in [1, \infty])$.

e.g. $p = 2 \rightarrow$ Euclidean distance.

$p = 1 \rightarrow$ "Manhattan distance"

$$\|x-y\|_1 = \sum_{i=1}^{d} |x_i - y_i|$$



$p = \infty$

$$\|x-y\|_\infty = \max(|x_1 - y_1|, \dots, |x_d - y_d|).$$
$$= \lim_{p\to\infty} \|x-y\|_p.$$

Note: ℓ_2 is very special!
- basis independent.
- self-dual

1). String distances. given s_1, s_2 strings,

edit distance:

$d_{edit}(s_1, s_2)$ = min # insertions + deletions to go from s_1 to s_2

s_1 = A G T A C A

s_2 = G T A A T

$d_{edit}(s_1, s_2) \le 3$ ←

~~A~~ G T A C A

↓

G T A ~~C~~ A

↓

G T A A

↓

G T A A T

"basic" DP : $O(mn)$

$O(n^{1+\delta})$ time to get $(1+\delta)$-approx

↳ best paper FOCS 2018

[Chakraborty, Das, Goldenberg, Koucky, Saks].

Hamming distance: # of replacements

$$\begin{matrix} A & G & T & A & C & A \\ G & T & A & A & T & \emptyset \end{matrix}$$

$d_{Hamming}(s_1, s_2) = 5.$

3). Jaccard (dis)-similarity given S, T sets

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

not a metric!

larger $J(S,T)$ ⟶ more similar S & T

$J(S,T) = 1. \Rightarrow S = T.$

$d_J(S,T) = 1 - J(S,T)$ ⟵ is a metric.

triangle inequality rather nonobvious but true!

Q: How similar are different metrics?

metric embeddings:

given two metric spaces (X_1, d_1), (X_2, d_2), is there a bijection $f: X_1 \to X_2$ s.t.

$$d_1(a,b) \approx d_2(f(a), f(b)) \quad \forall a,b \in X_1 \;?$$

or can I design X_2 so that f exists?

Active area of study!

Nearest-neighbor problem (X, d)

data set $D \subseteq X$

preprocess D s.t. given a query $q \in X$,

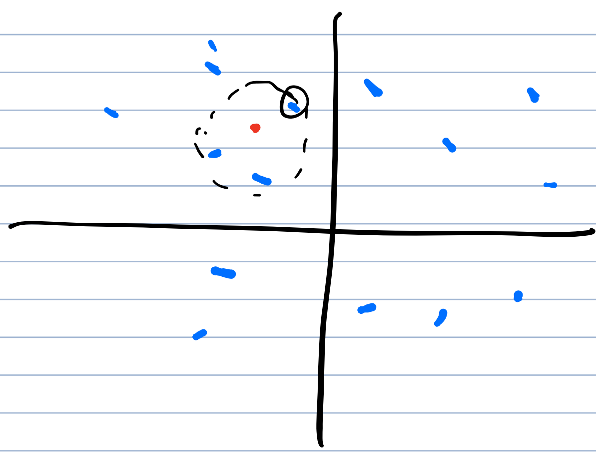
quickly find $x \in D$ s.t. $d(x,q)$ is minimized.

mostly focus on $X = \mathbb{R}^k$,

$d = \|\cdot\|_2$.

Approx-NN: given parameter $\varepsilon > 0$, return $x \in D$ s.t.

$$d(x,q) \leq (1+\varepsilon) \min_{x^* \in D} d(x,q).$$



Goal: If $|D| = n$, use space $O(n)$ and answer queries in time $O(\log n)$.

"Space partitioning"	Random projection
- kd-trees	- locality sensitive hashing (LSH)

this lecture

next lecture.

k-d tree [Bentley '75].

build a <u>search tree</u> to partition space.

many variants of this!

<u>leaf nodes:</u> Subset that contains 1 data element.

<u>non-leaf node:</u> A subset of points w/ size > 1. Some dimension i, and value m.

$\{x \in D : x_i < m\}$ $\qquad$ $\{x \in D : x_i \geq m\}$.

How to build a k-d tree: Initially, root node is all points in D.

at node v: Let D_v be associated set of points to v.

- if $|D_v| = 1$, v is leaf.
- if $|D_v| > 1$:

  pick dimension $i_v \in \{1, \cdots, k\}$

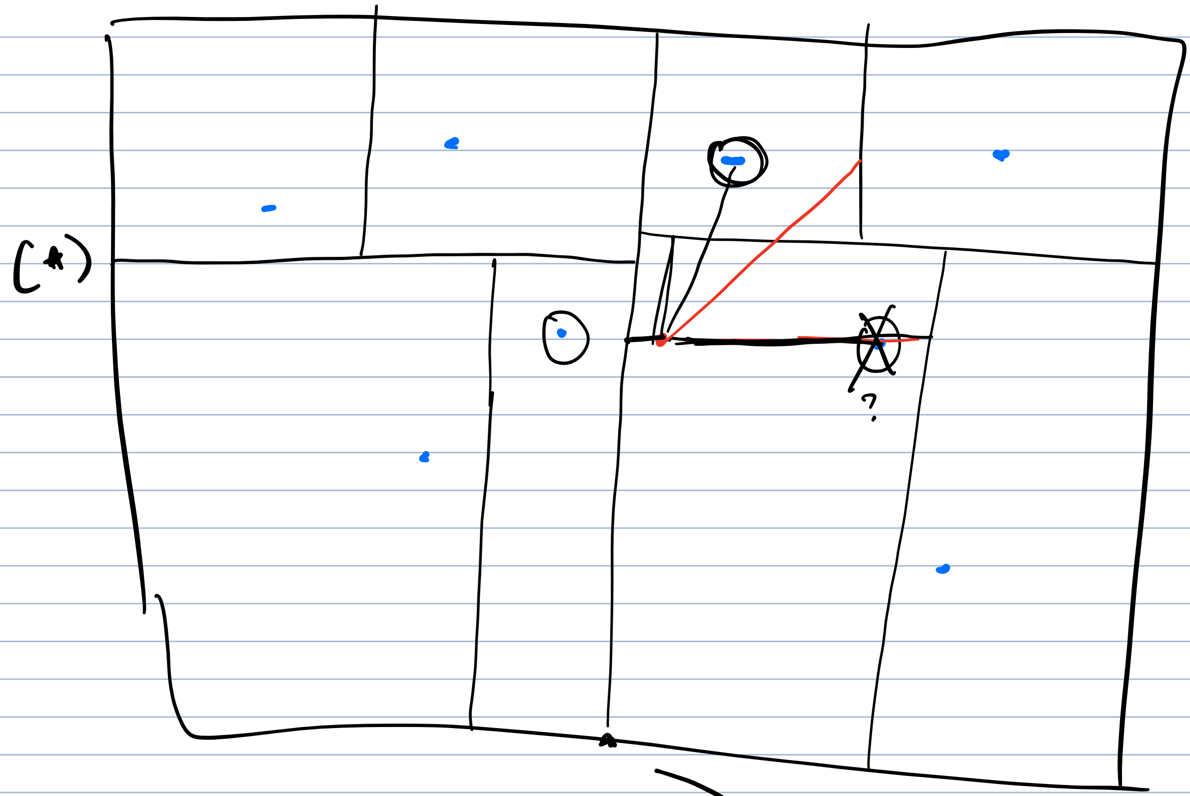
  let $m_v = \underset{x \in D_v}{\text{median}}(x_{i_v})$

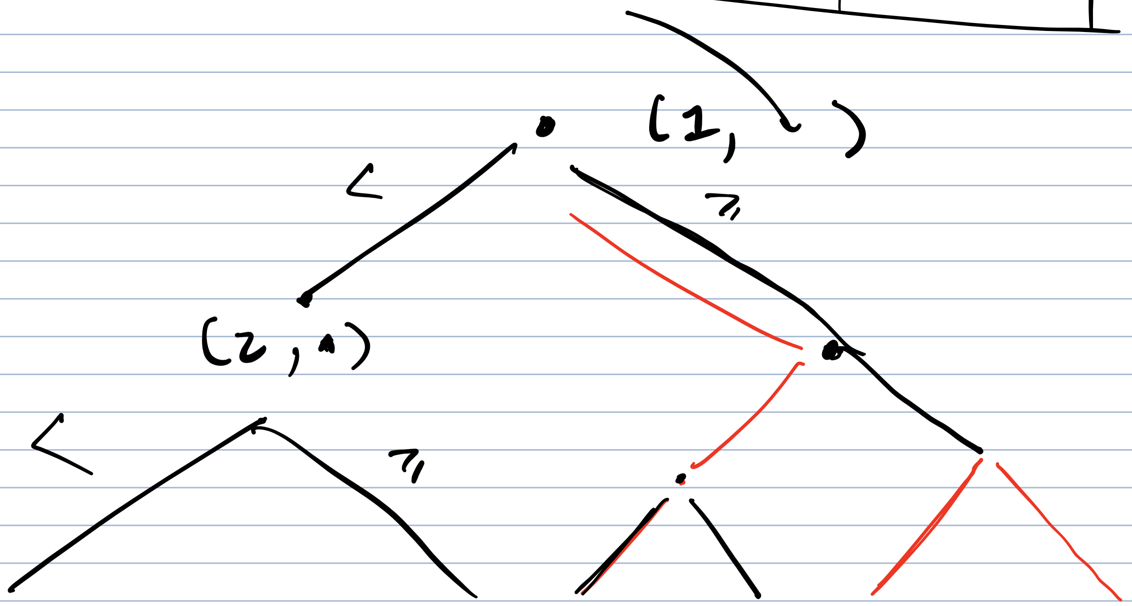
  store i_v, m_v at v.

  left child gets $\{x \in D_v : x_{i_v} < m_v\}$

  right child gets $\{x \in D_v : x_{i_v} \geq m_v\}$.

  recurse





size of tree? $O(n)$

depth of tree? $O(\log n)$

How to perform lookup?

need to recurse up tree!

worst case: might have to look at full tree!

in low dimensions: often much faster

"on average": $O(2^k \log n)$

rule of thumb: k-d tree useful when $\boxed{k \le 20}$

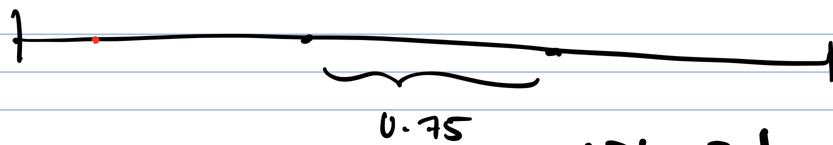
Curse of dimensionality:

Often times, running times for these sorts of space-partitioning methods (and others) scale exponentially w/ dimension.

high-dimensional points have lots of points w/ similar distances!

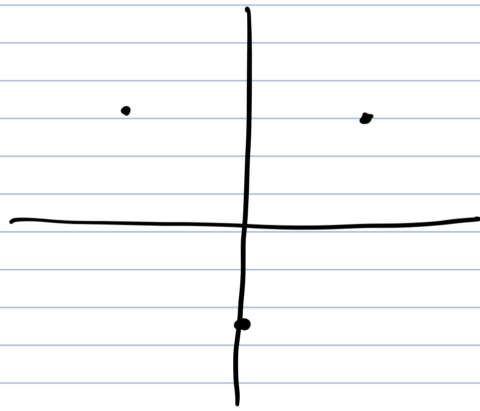
e.g. How large can D be s.t. $\|x-y\|_2 \in [0.75, 1]$ $\forall x, y \in D$?

$k = 1$:



$|D| = 2$!

$k = 2$:



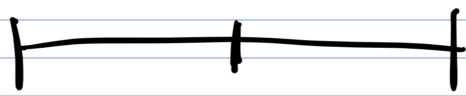
$|D| \le 3$.

$k = 100$? A lot!

in general: $\exists D \subseteq \mathbb{R}^k$ s.t. $|D| \geq 2^{ck}$

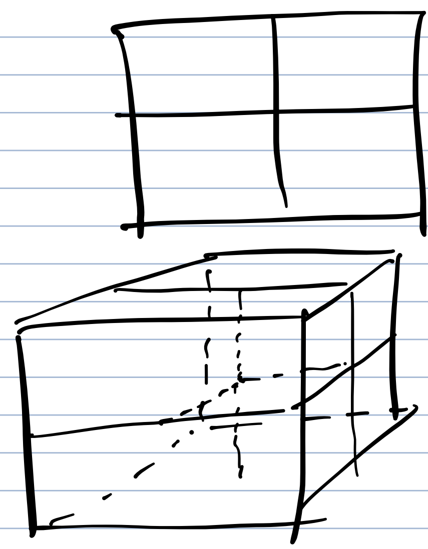
e.g. cell packing: How many boxes of edge length 1 can fit into a box w/ edge length 2?

$k = 1$



$\boxed{2}$

$k = 2$



$\boxed{4}$

$\boxed{2^k}$

need new approach for high dimensions!