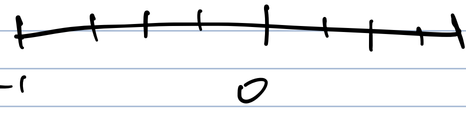
Lecture 4: dimensionality reduction.

The curse of dimensionality.

$\mathbb{R}^k = \{(x_1, \cdots, x_k) : x_i \in \mathbb{R}^k\}$

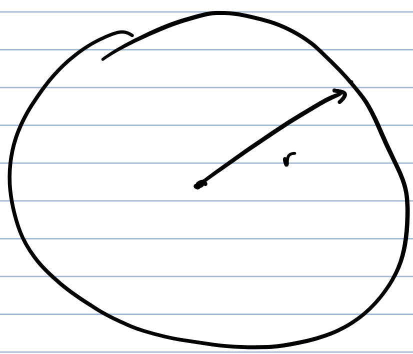
Let $B(\mu, r) = \{x \in \mathbb{R}^k : \|x - \mu\|_2 \le R\}$.

Q: How many balls of radius $1/8$ can you fit in a ball of radius 1? $k = 1$:

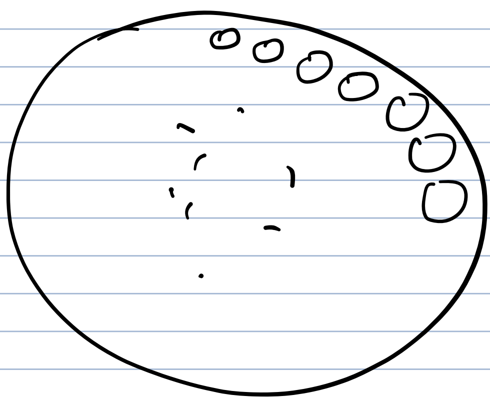


8.

general k



Volume $= C(r)^k \rightarrow \frac{\pi^{k/2}}{\Gamma(\frac{k}{2}+1)} \cdot r^k$



Fact: $\exists X \subseteq B(0,1)$ s.t. $\forall x \neq y \in X$, $\|x - y\|_2 \ge 1/4$, and $|X| \ge c^k$.

$\Rightarrow$ if you take balls of radius $1/8$ around every $x \in X$, they don't intersect!

volume of ball of radius $1/4 \rightarrow \left(\frac{1}{4}\right)^k$ "

tiny.

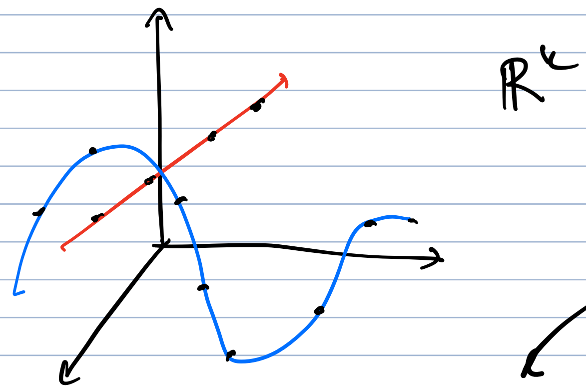
"Pf: proceed greedily. keep removing balls of radius $1/4$. Each one removes $\sim \left(\frac{1}{4}\right)^k$ mass, so you can do this $\exp(k)$ times.

Dimensionality reduction:

Can we replace our dataset w/ another in lower dimension that still preserves the relevant info?

"Intrinsic Dimensionality"

Oftentimes, high dimensional data secretly has low dimensional structure.



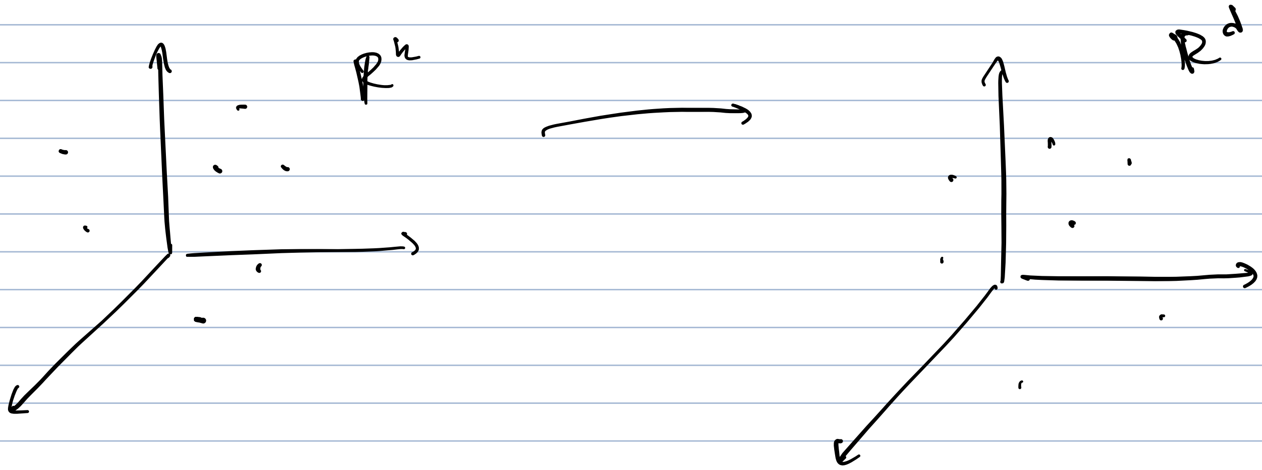
"metric embedding."

(Approximate) dimensionality reduction

Given a subset $X \subseteq \mathbb{R}^k$, target dimension $d \ll k$, approx error $\varepsilon > 0$,

find $F : \mathbb{R}^k \to \mathbb{R}^d$ s.t. $\forall x, y \in X$,

$$(1-\varepsilon) \cdot \|x-y\|_2 \leq \|F(x) - F(y)\|_2 \leq (1+\varepsilon)\|x-y\|_2.$$



Johnson-Lindenstrauss Lemma [84]:

Form a random $k \times d$ matrix A

$$A = \begin{array}{cc} & d \\ k & \left[\begin{array}{c} \frac{x_{ij}}{\sqrt{d}} \end{array} \right] \end{array}$$

$A_{ij} = \frac{\pm 1}{\sqrt{d}}$ also works

where $A_{ij} = \frac{x_{ij}}{\sqrt{d}}$, and each $x_{ij} \sim N(0,1)$ are independent.

$A : \mathbb{R}^d \to \mathbb{R}^k \quad x \longmapsto Ax$

$\ln(n)$

Then, $\forall \varepsilon \in (0,1)$ take $d := \left\lceil \frac{24 \log n}{\varepsilon^2} \right\rceil$ $\left(d = O\left(\frac{\log n}{\varepsilon^2}\right)\right)$

Then, for any set $X \subset \mathbb{R}^k$ w/ $|X| = n$,

w.p. $\geq 1 - \frac{1}{n^2}$: $\forall x, y \in X$:

$$(1-\varepsilon) \cdot \|x-y\|_2 \leq \|Ax - Ay\|_2 \leq (1+\varepsilon)\|x-y\|_2$$

"with high probability"

can choose any, will change constants.

<u>Proof:</u> Fix $x, y \in X$. We will show that

$$\Pr\left[\left| \|Ax - Ay\|_2 - \|x-y\|_2 \right| \leq \varepsilon \cdot \|x-y\|_2\right] \leq \frac{1}{n^4}.$$

Then we can union bound over all choices of x, y.

$$\Pr[\exists x, y \text{ s.t. } (*) \text{ fails}] \leq \binom{n}{2} \cdot \frac{1}{n^4}$$

n^2

$$\leq n^2 \cdot \frac{1}{n^4} \leq \frac{1}{n^2}$$

Let $z = x - y$. Want to show:

$$\Pr\left[\left| \|Az\|_2 - \|z\|_2 \right| \leq \varepsilon \cdot \|z\|_2\right] \leq 1/n^4$$

Let $u = \frac{z}{\|z\|_2}$

$$\iff \left| \|Au\|_2 - 1 \right| \leq \varepsilon.$$

$$\text{“}\iff\text{”} \quad \left| \|Au\|_2^2 - 1 \right| \leq \varepsilon/3$$

$$(1 \pm \varepsilon)^2 = 1 \pm 2\varepsilon + \varepsilon^2$$

$$\frac{1}{\sqrt{d}} \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ x_{21} & \cdots & x_{2k} \\ & \vdots & \\ x_{d1} & \cdots & x_{dk} \end{bmatrix} \begin{bmatrix} u \end{bmatrix} = \frac{1}{\sqrt{d}} \begin{bmatrix} \langle \vec{x}_1, u \rangle \\ \langle \vec{x}_2, u \rangle \\ \vdots \\ \langle \vec{x}_d, u \rangle \end{bmatrix}$$

Fact: If $\vec{x}$ is a Gaussian vector, then

$$\langle \vec{x}, u \rangle \sim \mathcal{N}(0,1)$$

So let $Y_i = \langle \vec{x}_i, u \rangle \sim \mathcal{N}(0,1)$.

then $Au = \frac{1}{\sqrt{d}} \begin{bmatrix} Y_1 \\ \vdots \\ Y_d \end{bmatrix}$ Y_i's are independent

so $\|Au\|_2^2 = \frac{1}{d} \cdot \underbrace{\sum_{i=1}^{d} Y_i^2}_{\chi^2 \text{ with } d \text{ degrees of freedom.}}$. $\mathbb{E}\, Y_i^2 = 1$

$$\mathbb{E}\left[\|Au\|_2^2 \right] = \mathbb{E}\left[\frac{1}{d} \sum_{i=1}^{d} Y_i^2 \right] = \frac{1}{d} \sum \mathbb{E}[Y_i^2] = d.$$

Fact: (Chernoff-ish)

$$\Pr\left[\left| \|Au\|_2^2 - 1 \right| \geq \varepsilon \right] \leq e^{-ck \cdot \varepsilon^2}$$

sum of d independent "nice" r.v's.

set $k = \frac{1}{c} \frac{4 \log n}{\varepsilon^2}$

$$= \exp\left(- \cancel{c} \cdot \cancel{\tfrac{1}{c}} \cdot 4 \tfrac{\log n}{\cancel{\varepsilon^2}} \cdot \cancel{\varepsilon^2}\right)$$

$$= n^{-4}$$

Locality Sensitive Hashing (LSH)

Typical hash functions hash to random places.

Can we hash in a way that respects data geometry?

x_1, x_2 close $\longrightarrow$ $h(x_1), h(x_2)$ close

far $\longrightarrow$ " " far.

e.g. JL!

Another example: Jaccard similarity

Recall: for two sets $S, T \subseteq U$

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|} \qquad J = 1 \to S = T \qquad J = 0 \quad S \cap T = \emptyset$$

An LSH for Jaccard: MinHash

Suppose universe $|U| = n$.

wlog $U = \{1, 2, \dots, n\}$.

Our LSH: Choose a random permutation $\pi : U \to U$.

and define, for all $S \subseteq U$

$$h_\pi(S) = \operatorname*{argmin}_{x \in S} \pi(x). \quad h : 2^U \to \mathbb{R} \quad 1\text{-D!}$$

$$\{1, \; 2, \; 3, \; 4, \; 5, \; 6\}$$

$$\pi \quad 5 \quad 1 \quad 6 \quad 4 \quad 1 \quad 2$$

$$S = \{1, 4, 6\} \qquad h_\pi(S) = 2.$$

$$\downarrow \quad \downarrow \quad \downarrow$$

$$5 \quad 4 \quad 2$$

Claim: $\forall\, S, T \subseteq U$,

$$\Pr_\pi [h_\pi(S) = h_\pi(T)] = J(S,T)$$

pf: Let $x \in S \cup T$ have the smallest label of all elts in $S \cup T$.

Then $h_\pi(S) = h_\pi(T) \iff x \in S \cap T$

x is a random element of $S \cup T$.

$$\Rightarrow \Pr[h_\pi(S) = h_\pi(T)] = \Pr_{x \sim \mathrm{Unif}(S \cup T)} [x \in S \cap T]$$

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

expectation is right, but variance is large.

Variance reduction:

$\pi_1, \ldots, \pi_\ell$ are iid random permutations.

$$J^H(S,T) = \frac{\# \{ i : h_{\pi_i}(S) = h_{\pi_i}(T) \}}{\ell}.$$

$$\mathbb{E}[J^H(S,T)] = \frac{1}{\ell} \mathbb{E}\left[\# \{ i : h_{\pi_i}(S) = h_{\pi_i}(T) \} \right]$$

$$= \frac{1}{\ell} \mathbb{E}\Big[\sum_{i=1}^{\ell} \underbrace{\mathbb{1}[h_{\pi_i}(S) = h_{\pi_i}(T)]}_{z_i} \Big]$$

$$= \frac{1}{\ell} \sum_{i=1}^{\ell} \Pr[h_{\pi_i}(S) = h_{\pi_i}(T)]$$

$$= \frac{1}{\ell} \cdot \sum_{i=1}^{\ell} J(S,T) = J(S,T)$$

$$J^H(S,T) = \frac{1}{\ell} \sum_{i=1}^{\ell} z_i$$

$$z_i \in [0,1], \quad \mathbb{E}[z_i] = J(S,T)$$

By Chernoff, $\ell = O\left(\frac{\log n}{\varepsilon^2}\right)$

$$|J^H(S,T) - J(S,T)| \leq \varepsilon \quad \text{w.p. } 1 - \frac{1}{n^c}.$$

LSH for nearest neighbor search

Hashing $\rightarrow$ exact duplicate.

near duplicate?

$$S \longrightarrow (h_1(S), \ldots, h_\ell(S)).$$

for query $q \rightarrow (h_1(q), \ldots, h_\ell(q))$, and find S in dataset that matches the most.

See problem set for more!