L16: Random Projections

for Dimensionality Reduction

Mar 17, 2025

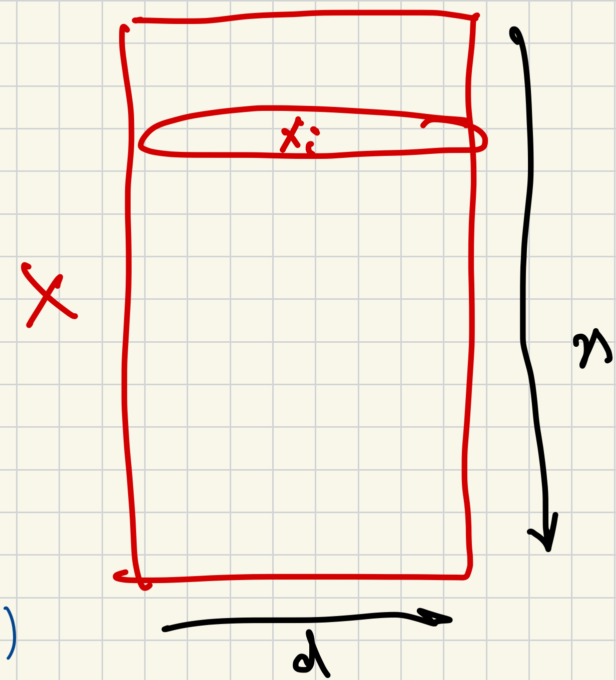
Data Mining

Jeff M. Phillips

Dimensionality Reduction

Data $X = \{x_1, x_2, \dots x_n\}$ $\quad x_i \in \mathbb{R}^d$

$$D(x_i, x_j) = \|x_i - x_j\|$$



When dimension d
very large

want map
$u(x_i) \to \mathbb{R}^k \qquad k \ll d.$

$$\|u(x_i) - u(x_j)\| \approx \|x_i - x_j\|$$

$v_1, v_2, v_3, \dots, v_k, v_{k+1}, v_{k+2}$

Random Projection

Make Sense

$d = 100{,}000$ or more ← native

1-hot encoding

bag-of-words

pixels

$k = 500$ or more

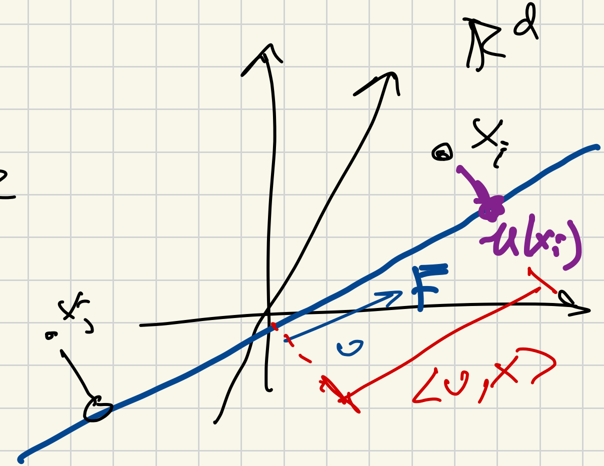
When can we apply DR.

$D(x_i, x_j) = \|x_i - x_j\|$ make sense.

- all coordinates mean same thing/unit
- learned / distributed embedding

$X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$

Projection $u(x_j) \to \mathbb{R}^k$



---

start F 1-dimensional

$$F = \{u \in \mathbb{R}^d \mid \|u\| = 1\}$$

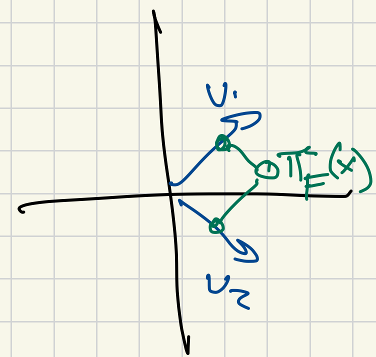
$$u = \pi_u(x) = \boxed{\underbrace{\langle x, u\rangle}_{\text{1-dimensional}}\, u} \in \mathbb{R}^d$$

F k-dimensional $F = \{u_1, u_2, \ldots u_k\}$

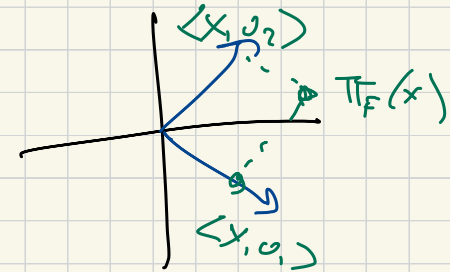
$u_j \in \mathbb{R}^d$, unit vector

$u_j \perp u_{j'}$

$$\pi_F(x) = \sum_{j=1}^{k} \langle x, u_j\rangle u_j$$



$F = \{v_1, v_2, \dots v_k\}$



$\pi_F(x) \Rightarrow \mathbb{R}^k$

$\pi_F(x) = \left(\langle x, v_1\rangle, \langle x, v_2\rangle, \dots, \langle x, v_k\rangle\right)$

How do we choose $F = \{v_1, \dots v_k\}$

?

Random choice

$v_i \sim S^{d-1}$

Random Projection

For $j = 1$ to k

choose $\underline{u_j \sim S^{d-1}}$

$g_1, g_2 \sim \mathrm{Unif}(0, 1)$

$g_1, g_2 - \mathrm{BoxM}(g_1, g_2)$

$(g_{j1}, g_{j2}, \dots g_{jd})$

$q_{ij} =$ jth coordinate $\pi(x)$

$= \langle u_j, x \rangle$

$q_i = \sqrt{\frac{d}{k}}(q_{i1}, q_{i2}, \dots q_{ik})$

- $E[\|\langle u_j, x \rangle\|^2] = \frac{1}{d}\|x\|^2$
- $k = O\left(\frac{1}{\varepsilon^2}\log\left(\frac{n}{\delta}\right)\right)$ n data points

$$\|x_i - x_j\|(1-\varepsilon) \leq \|\mu(x_i) - \mu(x_j)\| \leq (1+\varepsilon)\|x_i - x_j\|$$

true w.p. $1-\delta$ for all n points

$$\mu : \mathbb{R}^d \to \mathbb{R}^k$$

$$k = O\left(\frac{1}{\varepsilon^2} \log \frac{n}{\delta}\right)$$