Announcement:

Recitation — Friday 3–4 pm — Remote

Feb 15 2022

Dimension Reduction

Data in high dimensions.

E.g. Documents, Movie or product ratings by users, gene expression data

Curse of ↓ dimensionality (high)

E.g. Nearest-neighbor search

Dimen. Redun.: Transform the vectors into lower dimension while retaining useful properties

① pair-wise distances

② variance in the data points

1. Johnson-Lindenstrauss Transform (JL)

- Linear transform

$$x \longrightarrow Ax$$

↖ linear transformation mx.

- maintain pairwise distances ($L2$ norm) btwn data points

Set of points $X = \{x_1, x_2, \ldots, x_n\}$, $\in \mathbb{R}^D$

Each x_i

n ← # of data points

D ← initial dim

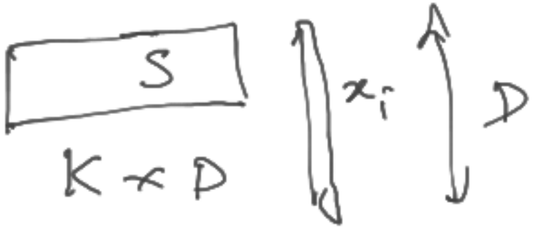
Want Linear Transform $S : \mathbb{R}^D \longrightarrow \mathbb{R}^K$

K ← final dim

$K < D$

Any vector $x_i \longrightarrow Sx_i$

$K \times D$



JL Lemma

Let $\varepsilon \in (0, 1/2)$. $X = \{x_1, x_2 \dots x_n\}$ each x_i in $\mathbb{R}^D$

There exist a map (transformation) $S : \mathbb{R}^D \to \mathbb{R}^k$

with $K = O\left(\frac{\log n}{\varepsilon^2}\right)$ s.t

$$(1-\varepsilon)\|x_i - x_j\|^2 \leq \|Sx_i - Sx_j\|^2 \leq (1+\varepsilon)\|x_i - x_j\|^2$$

$v = [v_1 \dots v_d]$

$$\|v\|^2 = \sum_{i=1}^{d} v_i^2$$

Obs:

① - K is indep. of D

② Log factor!

E.g. 10 billion points $\sim 10^{10}$ pts $\sim n \sim 2^{30} \Rightarrow K = O\left(\frac{30}{\varepsilon^2}\right)$

Construction:

Matrix A

each entry of the matrix $\sim N(0,1)$ (standard Normal / Gaussian distribution)

Transformation matrix:

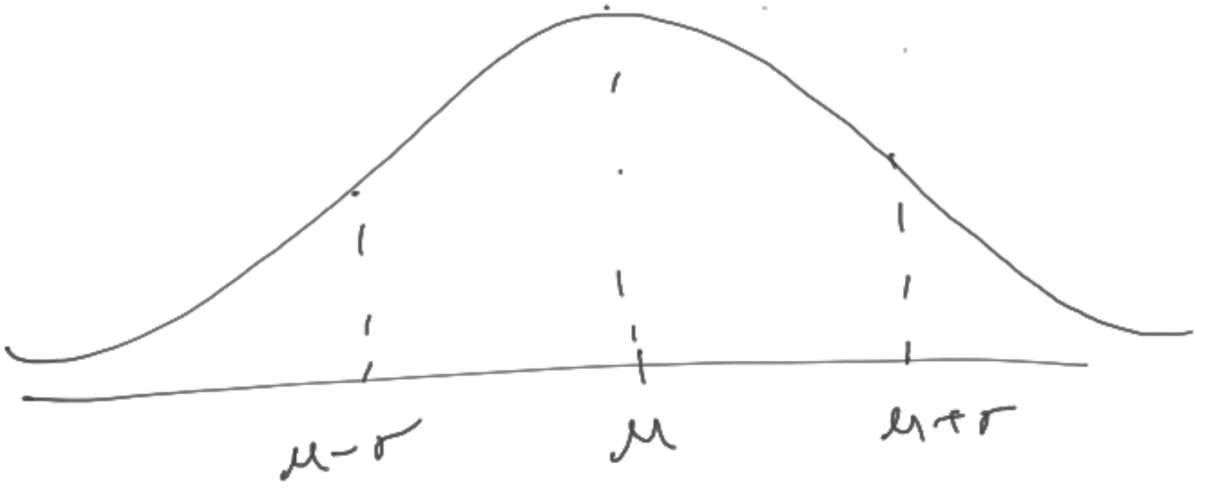
$$S = \frac{1}{\sqrt{K}} A$$

Gaussian R.V-s :

Cont. R.V.

Notation $\mathcal{N}(\mu, \sigma^2)$ — μ: mean, σ^2: variance

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



- exponential tapering off the mean

$e^{-\frac{y^2}{2\sigma^2}}$

properties:

1. $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ & indep.

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

2. $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow aX \sim \mathcal{N}(a\mu, a^2\sigma^2)$

3. $X_i \sim \mathcal{N}(0, 1)$ → Std. normal

$a = (a_1, \dots a_d)$

$$\sum_{i=1}^{d} a_i X_i \sim \mathcal{N}\left(0, \sum_{i=1}^{d} a_i^2\right)$$

$$= \mathcal{N}(0, \|a\|^2)$$

... JL Transform

Fact 1

$x \in \mathbb{R}^D$

$$A = \underset{K \times D}{\begin{bmatrix} - A_1 - \\ - A_2 - \\ \vdots \\ - A_K - \end{bmatrix}}$$

(D columns)

Recall: $x \to Sx = \frac{1}{\sqrt{K}} A x$

A: i.i.d $N(0,1)$ entries

Transformation

$$Sx = \frac{1}{\sqrt{K}} A \cdot x = \begin{bmatrix} \frac{1}{\sqrt{K}} \langle A_1, x \rangle \\ \vdots \\ \frac{1}{\sqrt{K}} \langle A_K, x \rangle \end{bmatrix} \quad (K \text{ rows})$$

given vector

$\langle A_1, x \rangle \to \sum A_{1i} x_i$

$A_{1i} \sim N(0,1)$, x_i: scalar

$$\sim \begin{bmatrix} N\left(0, \frac{\|x\|^2}{K}\right) \\ \vdots \\ N\left(0, \frac{\|x\|^2}{K}\right) \end{bmatrix} \quad (K \text{ rows})$$

Fact 2 For any $Y \in \mathbb{R}^D$

$$\mathbb{E}\left[\|SY\|^2\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{K} \frac{1}{K} \langle A_i, Y\rangle^2\right]$$

$$= \sum_{i=1}^{K} \frac{1}{K} \mathbb{E}\left[\langle AX, Y\rangle^2\right] \longrightarrow$$

$$= \|Y\|^2$$

$Sx_i - Sx_j$

↓

$S(x_i - x_j)$

SY

Variance

if mean = 0,

$\mathbb{E}[X^2]$ = variance

$$\Rightarrow \mathbb{E}\left[\|SY\|^2\right] = \|Y\|^2$$

$$\Rightarrow \mathbb{E}\left[\|Sx_1 - Sx_2\|\right] = \|x_1 - x_2\|^2 \Rightarrow$$

result holds in expectation.

Lemma (Conc. bound for squared-Gaussians)

Let $v_1, \ldots v_k$ be i.i.d $\mathcal{N}(0, \sigma^2)$

Let $Z = \sum_{i=1}^{k} v_i^2$

$$E[Z] = k\sigma^2$$

Then for $\varepsilon \in (0, 1/2)$, some constant c

$$P(|Z - E[Z]| \geq \varepsilon E[Z]) \leq e^{-\frac{k\varepsilon^2}{c}}$$

Conc. bound)

$$P(|Z - E[Z]| \geq \alpha) \leq$$

... JL Transform

As before $Y = x_1 - x_2$

$$P\left[\left|\|SY\|^2 - E[\|SY\|^2]\right| \geq \varepsilon \|Y\|^2\right] \leq ?$$

LHS

$$\|SY\|^2 = \sum_{i=1}^{k} v_i^2 \quad \text{where} \quad v_i \sim N\left(0, \frac{\|Y\|^2}{k}\right)$$

From Conc. bound Lemma

$$\text{L.H.S.} \leq e^{-\frac{k\varepsilon^2}{c}}$$

Choose $k = \frac{2c \log n}{\varepsilon^2}$

want $\leq \frac{1}{n^2}$

to use union bound over all pair-wise distances.

all pairwise dist = $\binom{n}{2} < n^2$

By union bound, over all $\binom{n}{2}$ pair-wise distances.

then squared lengths maintained w.p. at least

$$1 - \binom{n}{2} \cdot \frac{1}{n^2} \geq \frac{1}{2}$$

$\Rightarrow$ want $\frac{1}{n}$

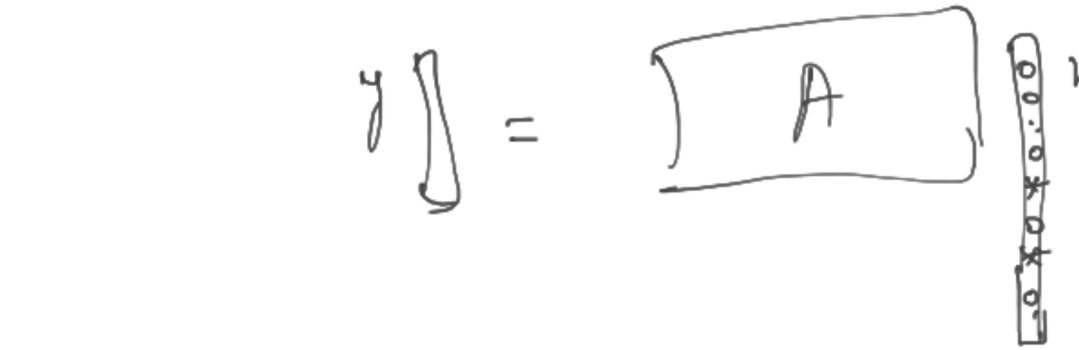
make this $\frac{1}{n^3}$ gives

↳ Choose $K = 3c \cdots$

Discussions:

1. Runtime.
2. Applications: ① fast (but approximate) mat. mul.

② Compressive Sensing



E.g. MRI imaging, cameras