L18: Random Projections

10.29.2025

Data Mining

Jeff M. Phillips

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

d extremely large

≈ 100 million

goal $Q = \{q_1, q_2 \dots q_n\} \subset \mathbb{R}^k$

$k \ll d$

$k \approx 500$

$\forall$ all pairs $x_i, x_j \in X$

$$(1-\varepsilon)\,\|x_i - x_j\| \le \|q_i - q_j\| \le (1+\varepsilon)\,\|x_i - x_j\|$$

error tolerance

$\varepsilon \in (0,1)$

$\varepsilon \approx 0.1$

or 0.01

Johnson Lindenstrauss

Random Projections

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

w/ distortion $(1 \pm \varepsilon)$

Define set $v_1, v_2 \dots v_k$ $v_j \in \mathbb{R}^d$

$\|v_j\| = 1 \cdot \frac{\sqrt{d}}{\sqrt{k}}$

$q = \mu(x) = \bar{\pi}_{v_1 \dots v_k}(x)$

$q_j = \langle x, v_j \rangle$

jth coord of q

How do we choose v_j

generate $g \in \mathbb{R}^d$

$g_j \sim N(0,1)$

$g \sim N_d(0,1)$

$v_j = v = \frac{g}{\|g\|} \cdot \frac{\sqrt{d}}{\sqrt{k}}$

Why does random projection work?

$v \sim N_d(0,1)$

$\bar{v} = \frac{v}{\|v\|}$

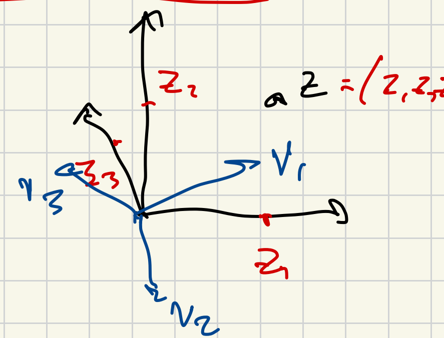
$x_i - x_j = z \in \mathbb{R}^d$

$$E_v\left[\langle z, v\rangle^2\right] = \boxed{\frac{1}{d}\cdot\|z\|^2}$$

$$\|z\|^2 = \sum_{j=1}^{d} z_j^2$$

$$\|z\|^2 = \sum_{j=1}^{d} \langle v_j, z\rangle^2$$

pick one $v_j = v$ at random



$\|v_j\| = 1 \quad \langle v_1, v_2\rangle = 0$

unbiased

$$E\left[d\cdot\langle z, v\rangle^2\right] = \|z\|^2$$

$$E\left[\frac{1}{k}\sum_{j=1}^{k} d\cdot\langle z, v_j\rangle^2\right] = \|z\|^2$$

$$E\left[\frac{d}{k}\|z\cdot U\|^2\right] = \|z\|^2$$

$$U = \left[v_1 \ v_2 \ \dots \ v_k\right]$$

Chernoff-Hoeffding bound n R.V. X_i:

subGaussian

$$M = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\Pr\left[\left| M - E[M] \right| > \varepsilon \right] \le \delta$$

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

$\approx$ Holds for any one $z = X_i - X_j$ pair.

For all distances $\delta' = \frac{\delta}{n^2}$ $\quad n^2 \approx$ # pairs

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