L18: Random Projections

Data Mining
Nov 3, 2026
Jeff M. Phillips

Input $X \subset \mathbb{R}^d$ $x_1 \dots x_n \in \mathbb{R}^d$

- d too big
- do not trust Euclidean

↳ Learn a Mahalanobis distance

$$d_M(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$

scaling + coordinates

$M \in \mathbb{R}^{d \times d}$ s.t. $M \succeq 0$

if $M = I \Rightarrow$ Euclidean

↳ Learn M.

Dim-Reduction

vs low rank

M psd $\Rightarrow$ $M = A^T A$

$M_k \Leftarrow \text{rank}(k)$

implicitly $\Rightarrow \pi(x_i)$

$\hookrightarrow q_i$

$\text{eig}(M)$ $\quad M = V \Lambda V^T$

$A^T = V \Lambda^{1/2}$

$A = \Lambda^{1/2} V^T$

$\Rightarrow M_k = V_k \Lambda_k V_k^T$

truncate

$A_k = \Lambda_k^{1/2} V_k^T$

$D_M(x_1, x_2) = \sqrt{(x_1 - x_2)^T M_k (x_1 - x_2)}$

$= \sqrt{(x_1 - x_2)^T A_k^T A_k (x_1 - x_2)}$

$= \sqrt{(A_k x_1 - A_k x_2)^T (A_k x_1 - A_k x_2)}$

$= \sqrt{(q_1 - q_2)^T (q_1 - q_2)} = \|q_1 - q_2\|$

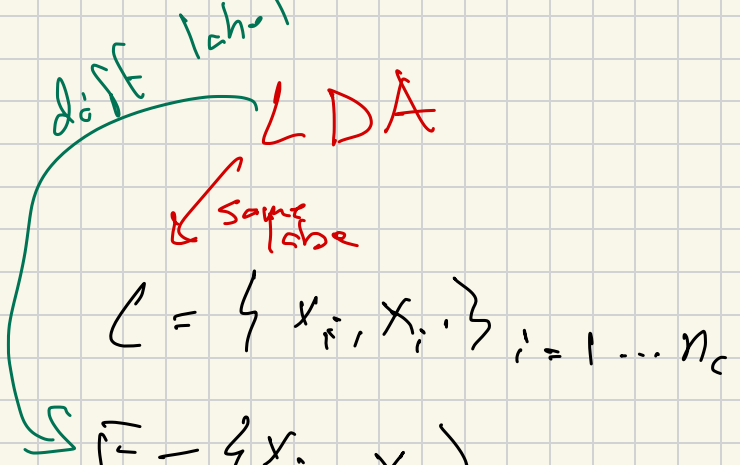
$\pi(x_i) \to q_i \in \mathbb{R}^k$

$A_k x_i = q_i$

How to learn M?

More input $x \subset \mathbb{R}^d$

- close pairs
- far pairs



① LDA $\Sigma_B^{-1}\Sigma_W = \Sigma_F^{-1}\Sigma_C = M$

② Constraints $\det(\Sigma_C) = 1$ easier

force for pairs as long as possible

③ in Bonus Question

Random Projections

$d = 10$ mill

$k = 100s - 1000$

Input $X = \{x_1, \ldots x_n\} \in \mathbb{R}^d$

Goal $Q = \{q_1 \ldots q_n\} \in \mathbb{R}^k$

$u = \frac{1}{\sqrt{k}}$

Johnson-Lindenstrauss Lemma

set error tolerance $\varepsilon \in (0,1)$ $\varepsilon = 0.1$

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

holds for all pair $x_i, x_j \in X$.

$k = \frac{1}{\varepsilon^2} \log \frac{n}{\delta}$ words w/ δ prob. failure

$\varepsilon = 0.1 \Rightarrow 100 \cdot \log \frac{n}{\delta}$

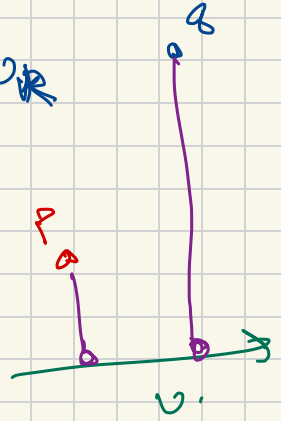
Determine Mapping $u : \mathbb{R}^d \to \mathbb{R}^k$

Define k unit vectors $v_1, v_2, \ldots v_k$

$v_j \in \mathbb{R}^d$

$\|v_j\| = 1$

$$u(x) = \frac{\sqrt{d}}{\sqrt{k}}\left(\langle v_1, x\rangle, \langle v_2, x\rangle \ldots \langle v_k, x\rangle\right)$$



Choose each v_j at random

(ith coord)

$$g_{ji} \sim N(0,1)$$

$$g_j \sim N_d(0,I)$$

$$v_j = \frac{g_j}{\|g_j\|}$$

Why does Rand Proj work?

n^2 pairs $x_i, x_j \in X$

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

$$\mathbb{E}_{v_j}\left[\langle z, v_j\rangle^2\right] = \|z\|^2$$

$$\mathbb{E}_{v_1 \ldots v_k}\left[\frac{d}{k}\sum_{j=1}^{k}\langle z, v_j\rangle^2\right] = \|z\|^2$$

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

$$\|z\| \approx \frac{\sqrt{d}}{\sqrt{k}}\|z \cdot V\|$$

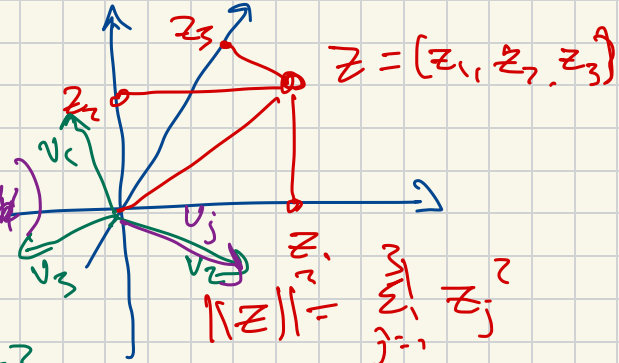
Claim $\mathbb{E}_{v_j}\left[\langle z, v_j\rangle^2\right] = \frac{1}{d}\|z\|^2$

$$\|z\|^2 = \sum_j \langle v_j, z\rangle^2$$

$v_j \sim \text{Unif}(v_1, v_2 \ldots v_d)$

$$\mathbb{E}_{v_j}\left[\langle v_j, z\rangle^2\right] = \frac{1}{d}\|z\|^2$$

$\|v_j\|^2 = 1 \qquad \langle v_j, v_{j'}\rangle = 0$



P3 peer feedback: find your cluster

A Player types
19 NFL players
31 NBA archetypes
36 NFL teams

B Baseball & F1
13 MLB careers
34 Formula 1
42 MLB defense

C Steam games
24 Steam reviews
33 Steam games
40 Steam players

D Embeddings
20 Steam banners
25 Amazon products
41 Image captions

E Music
06 Music audio
21 Spotify charts
30 Playlists

F Social & web
04 Influence ops
10 Web activity
23 Social ties
26 Facebook

G Language
07 Wiktionary
14 TV news
16 LLM activations

H Products
11 Restaurants
18 Amazon reviews
39 Pokémon cards

I Utah data
08 Snow depth
09 Wildfires
22 Air quality
37 UTA buses

J Signals
03 Ocean sound
35 SDSS stars
38 Muscle EMG

K Health
01 Mortality
15 Cancer risk
17 Gut microbiome

L US economy
02 US metros
27 Patents & S&P
32 Housing & votes

M Industry
05 Job skills
12 World energy
28 Factory data

How it works
1. Find your group number.
2. Sit with the groups in your cluster.
3. Give each other feedback on your P3.