Second task of unsupervised learning

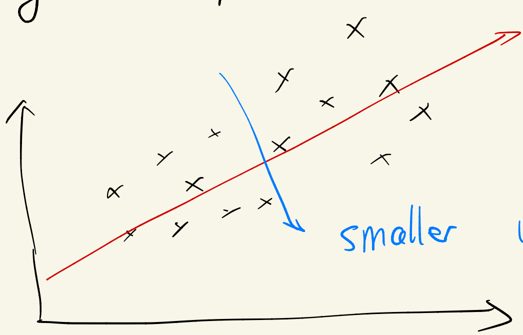
dimensionality reduction

goal: reduce high dimensional data to low-D.

classical method: PCA: principal component analysis

- Given data $x_1, x_2, \dots x_n \in \mathbb{R}^d$.
- Would like to have a reduced-dimension representation $y_1, y_2, \dots y_n \in \mathbb{R}^l$ with $l \ll d$.
  such that important info is kept.

A toy example.



larger variance along this dir.

smaller variance along this direction.

$\longrightarrow$ project onto direction with larger variance.

First interpretation of PCA

$\longrightarrow$ maximize the variance of the reduced data

assume data is centered i.e.

$$\sum_{i=1}^{n} x_i = 0.$$

want to find a <u>direction</u> u.

$u \in \mathbb{R}^d \qquad \|u\|_2 = 1.$

s.t. the projections

$\{u^T x_i\}_{i=1}^{n}$ have max variance.

$$\Longleftrightarrow \max_{u:\|u\|_2=1} \sum_{i=1}^{n} (x_i^T u)^2.$$

$$\Longleftrightarrow \max_{u:\|u\|_2=1} u^T \underbrace{\sum_{i=1}^{n} x_i x_i^T}_{} u$$

$$\triangleq XX^T, \text{ where}$$

covariance matrix

$$X = d\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}^{\,n}$$

This gives us the first PC. (principal component)

$$u_1 = \underset{u: \|u\|_2 = 1}{\arg\max} \sum_{i=1}^{n} (x_i^T u)^2 .$$

How to obtain the second PC ??!

$$u_2 = \underset{\substack{u: \|u\|_2 = 1 \\ u^T u_1 = 0}}{\arg\max} \sum_{i=1}^{n} (x_i^T u)^2 .$$

↳ the 2nd eigenvector of XX^T.

⋮ can extend all the way to k PCA.

Algorithm for PCA.

- center the date set.
- compute eigen-decomposition of $XX^T = U \Sigma U^T$.
- return $U = [\underbrace{u_1 \; u_2 \; \cdots \; u_k}_{\text{top-}k \text{ eigenvectors.}} \; u_{k+1} \cdots u_d]$.

What's the low-D representation of x ???!

$$x \to z = [u_1^T x, u_2^T x, \dots u_k^T x] \in \mathbb{R}^k.$$

---

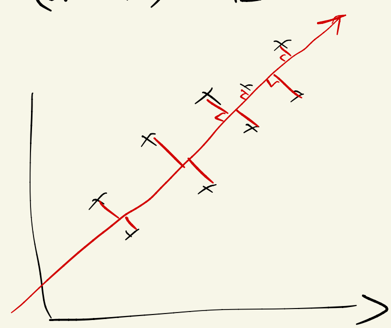
second interpretation: minimizing reconstruction error.

$$x \to (u^T x)\, u \quad \text{for any } \|u\|_2 = 1.$$

would like to have small reconstruction error.

$$\| x - (u^T x) u) \|_2^2 \quad \text{is small.}$$

$$\to \quad \underset{u:\, \|u\|_2 = 1}{\arg\min} \sum_{i=1}^n \| x_i - (u^T x_i) u \|_2^2$$



Claim: this yields exactly the same direction as u_1.

more generally, we have.

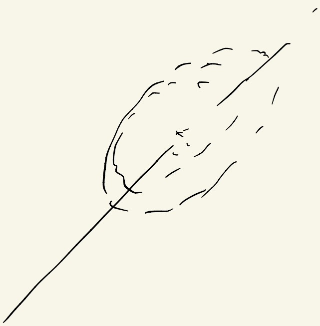
$$\min_{u_1, u_2, \ldots u_k} \sum_{i=1}^{n} \| x_i - \sum_{j=1}^{k} (u_j^T x_i) u_j \|_2^2$$

s.t. $u_i^T u_i = 1 \quad \forall i.$ $\quad \boxed{u_i^T u_j = 0 \quad i \neq j}$

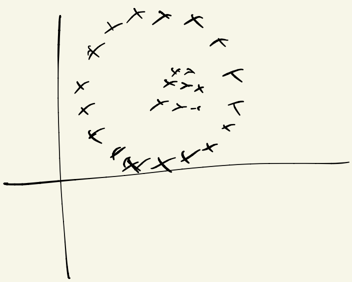
How to choose k ??!

- choose k s.t. the PCs have low reconstruction error.
- cross validation on downstream tasks.

PCA likes Gaussian data.



dislikes other structured data.



$\longrightarrow$ kernel PCA.

Advanced: random projections

and Johnson–Lindenstrauss lemma.

Given $x_1, \dots x_n \in \mathbb{R}^d$ $\rightarrow$ hard to store and manipulate.

construct a mapping $\pi: \mathbb{R}^d \to \mathbb{R}^k$ $k \ll d$.

s.t. all distances are nearly preserved, i.e.

$$\underbrace{\|x_i - x_j\|_2}_{\mathbb{R}^d} \approx \| \underbrace{\pi(x_i)}_{\mathbb{R}^k} - \pi(x_j) \|_2 .$$

example: <u>k-means clustering</u>.

More precisely: we want to achieve.

$$1-\varepsilon \le \frac{\|\pi(x_i) - \pi(x_j)\|_2}{\|x_i - x_j\|_2} \le 1-\varepsilon \quad *$$

for some small ε say $\varepsilon = 0.001$.

Lemma (Johnson–Lindenstrauss 1984).

As long as $k > \dfrac{4 \log n}{\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}}$

for any set of n data points in $\mathbb{R}^d$.

there exists a map s.t. $(*)$ is true

Remarkable property:

$\underline{k \text{ is dimension independent.}}$

only depends log on $\underline{\underline{n}}$.

In fact: you can achieve $*$ simply by

random projection:

let $W \in \mathbb{R}^{k \times d}$ be Gaussian random matrix

define $\pi(x) = \frac{Wx}{\sqrt{m}}$. this "almost always" works.

why ??! fix some x.

① $\mathbb{E} \|\pi(x)\|_2^2 = \mathbb{E} \frac{\|Wx\|_2^2}{m} = \|x\|_2^2$.

②. $\|\pi(x)\|_2^2$ concentrates well around $\|x\|_2^2$.

③ we only need this to hold for n^2 pairs.

Extensions

— : other distances.