DSA5105 lec8

Recap:

up to now, there are supervised learning algos.

$\updownarrow$

labelled dataset and learn the oracle function

Hypothesis class

Unsupervised Learning → understand the data (Aim)

why?

- ① labelled data expensive
- ② labelled data impossible to attain
- ③ different application scenarios

① find some structure of data

② dimension reduction → reduce cost

will lose part of information (depends on the situation)

③ clustering → cluster users

④ density estimation (non-parametric stats)

$$p(x)\,dx = \mathbb{P}(X \in dx)$$

Application: generate new samples

⑤ Generative Model (GANs)

↓

Stable diffusion model (SOTA)

Gaussian Noise → Model GAN → samples similar

PCA

① Recall

a) $Au = \lambda u \longrightarrow \begin{cases} \text{eigenvalue: } \lambda \\ \text{eigenvector: } u \end{cases}$

b) Diagonalization of $A \iff A = P \Lambda P^{-1}$ P is invertible

$\iff AP = P\Lambda \iff \boxed{A p_i = \lambda_i p_i}$

$\implies$ column of P is eigenvector

c) symmetric $A \Rightarrow \boxed{A = U \Lambda U^T}$

where $\begin{cases} UU^T = I \ (U \text{ is orthonormal}) \\ \Lambda \text{ diagonal} \end{cases}$

d) PD $A \iff x^T A x > 0$ for all $x \in \mathbb{R}^n$ and $x \neq 0$

PSD $A \iff x^T A x \geq 0$ for all $x \in \mathbb{R}^n$

e) symmetric A PD $\iff A = U \Lambda U^T$

and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$

$\downarrow$

$\boxed{\text{eigenvalue}}$

② PCA $\begin{cases} \text{min error} \\ \\ \text{max variance} \end{cases}$ $\boxed{2 \text{ interpretation}}$ (formulation)

a) max var derivation

$\mathcal{D} = \{ (x_i) \}_{i=1}^{N}$

某一

沿着$\boxed{\text{某个方向}}$的 projection

$\boxed{\text{Goal:}}$ Find a direction $u \in \mathbb{R}^d$ s.t "Variance" of the projection is maximized

$A \cdot x$

Assume $\|u\| = 1$.

1. How to compute projection

x

u $\longrightarrow$ $\boxed{\langle x, u\rangle \cdot u}$

$\Downarrow$

Coordinate given by $\langle x, u\rangle$

2. Variance $= \frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2$

Then we can compute :

(Assume $\bar{x} = \frac{1}{N}\sum x_i = 0$) $\rightarrow$ we can normalize $\tilde{x}_i = x_i - \bar{x}$

then let $\boxed{z_i = x_i^T u}$ $\Rightarrow$ $z = Xu$ $\quad X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix}$

$$\Rightarrow \bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{1}{N}\sum_{i=1}^{N} x_i^T u = 0$$

$$\Rightarrow \text{Variance} = \frac{1}{N}\sum_{i=1}^{N} z_i^2$$

$$= \frac{1}{N} z^T z$$

$$= \frac{1}{N} u^T \underline{X^T X} u$$

$\rightarrow$ $\boxed{\text{symmetric, PSD}}$

$$= u^T \underline{\left(\frac{1}{N} X^T X\right)} u$$

$\rightarrow$ $\boxed{\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T}$

Variance $:=$ $\underline{u^T S u}$ $\quad \Sigma$ is symmetric & PSD

$\longrightarrow$ max variance formulation

$$\Rightarrow \boxed{\begin{array}{ll} \max\limits_{u \in \mathbb{R}^d} & u^T S u \\ \text{s.t} & \|u\| = 1 \end{array}} \longrightarrow L(u;\lambda) = u^T S u + \lambda(1 - \|u\|_2^2)$$

$$\boxed{\nabla L(u;\lambda) = 2Su - 2\lambda u}$$

$\downarrow$

Convex prog $(S \succeq 0)$ $\nearrow$ $\nabla_u L(\hat{u};\hat{\lambda}) = 0$

$\Leftrightarrow$ $\boxed{S\hat{u} = \hat{\lambda}\hat{u}}$

Strong duality

Stationarity

$\Downarrow$

$\hat{u}$ is the eigenvector of some eigenvalue $\hat{\lambda}$

$$\begin{aligned} \max_{u \in \mathbb{R}^d} \quad & \lambda u^T u \\ \text{s.t} \quad & \|u\| = 1 \\ & u \text{ is eigenvector of } S \text{ for some } \lambda \end{aligned}$$

$\Downarrow$

$(\hat{\lambda}, \hat{u})$ should be the largest eigenvalue $(\hat{\lambda})$ and corresponding eigenvector $(\hat{u})$

$$\Rightarrow \begin{cases} \hat{u} : \text{ the first principal compoment} \\ \boxed{z_i = x_i^T \hat{u} : \text{ the scores}} \end{cases}$$

x_i's score for one direction

Generalization to several directions: (u_1, u_2)

Modification

$$\begin{cases} \max_{u \in \mathbb{R}^d} & u^T S u \\ \text{s.t} & \|u\| = 1 \\ & u^T u_1 = 0 \end{cases} \longrightarrow \mathcal{L}(u; \lambda, \beta) = u^T S u + \lambda (1 - u^T u) + \beta u^T u_1$$

$$\nabla_u \mathcal{L}(u; \lambda, \beta) = 2Su - 2\lambda u + \beta u_1$$

$$\xrightarrow{\text{strong duality}} (\hat{u}; \hat{\lambda}, \hat{\beta}) \text{ s.t } \begin{cases} \nabla_u \mathcal{L}(\hat{u}; \hat{\lambda}, \hat{\beta}) = 0 \\ \hat{u}^T u_1 = 0 \end{cases}$$

$$\Rightarrow 2 u_1^T S \hat{u} + \hat{\beta} = 0$$

$$\Rightarrow 2 \lambda_1 \underline{u_1^T \hat{u}} + \hat{\beta} = 0$$

$$\Rightarrow \hat{\beta} = 0 \xrightarrow{\text{代入}} \nabla_u \mathcal{L}$$

$$\Rightarrow S \hat{u} = \hat{\lambda} \cdot \hat{u} \Rightarrow \underline{\text{eigenvalue \& eigenvector}}$$

$\max \quad \lambda u^T u$ | $\max \quad \lambda$

$s.t \quad \|u\|=1 \quad \Leftrightarrow$ | $s.t$ ① λ is eigenvalue

$\boxed{u^T u_1 = 0} \rightarrow$ ② $\lambda \neq \lambda_1$

(u, λ) is the (eigenvector, eigenvalue)

$\Downarrow$

$(\hat{\lambda}, \hat{u})$ should be the $\boxed{\text{second largest}}$ eigenvalue

and corresponding eigenvector

✲ $\boxed{\text{explained variance}} \rightarrow$ ① $U_m = (u_1, \ldots, u_m)$

$Z_m = X U_m$ is the score matrix

② $\sum_{i=1}^{m} \lambda_i$ is the explained variance

---

b) $\boxed{\text{min error}}$ (second interpretation)

(projection error)

defn

$$\|x - (x^T u) \cdot u\|_2^2$$

1) if $x = \alpha u$, then projection error $= 0$

the overall projection error: $= \frac{1}{N} \sum_{i=1}^{N} \|x_i - (x_i^T u) \cdot u\|_2^2$

$\|X - X \cdot u \cdot u^T\|_F^2$

our aim: $\min_u \quad \boxed{\frac{1}{N} \sum_{i=1}^{N} \|x_i - (x_i^T u) \cdot u\|_2^2}$

$s.t \quad \|u\| = 1$

$XU = (x_i^T u_1, \ldots, x_i^T u_m)$

calculation: $\sum_{i=1}^{N} \|x_i - (x_i^T u) u\|_2^2$

$$= \sum_{i=1}^{N} \|x_i\|_2^2 - 2 (x_i^T u)^2 + \|(x_i^T u) \cdot u\|_2^2$$

$$= \sum_{i=1}^{N} \|x_i\|_2^2 - (x_i^T u)^2$$

$$\min \sum_{i=1}^{N} \left(\|x_i\|_2^2 - (x_i^T u)^2\right)$$

$\Leftrightarrow \max \sum_{i=1}^{N} (x_i^T u)^2$

$\Leftrightarrow \boxed{\max \ u^T X^T X u} \longrightarrow$ $\boxed{\text{max variance interpretation}}$!!!

$\boxed{\text{Application}} \longrightarrow$ don't use diagonalization to compute $\begin{cases} \text{Eigenvalue} \\ \text{Eigenvectors} \end{cases}$ for X^TX

$\boxed{\text{SVD instead}}$

$\boxed{\text{Kernel PCA}} \longrightarrow$ focus on Non-linear data (circle ...)

$\boxed{\phi: x \longmapsto \phi(x)}$

$X \longrightarrow \Phi \longrightarrow \boxed{\Phi^T \Phi \in \mathbb{R}^{D\times D}} \longrightarrow$ PCA

$X = \begin{pmatrix} x_i^T \\ \vdots \\ \vdots \end{pmatrix}$, $\Phi = \begin{pmatrix} \vdots \\ \phi(x_i)^T \\ \vdots \end{pmatrix} \in \mathbb{R}^{N\times D}$

$\boxed{\sum_{i=1}^{N} \phi(x_i) \cdot \phi(x_i)^T}$

$\boxed{\text{Recall}}$ Principal component scores $Z = X \cdot U \in \mathbb{R}^{N\times m}$ $\boxed{U \in \mathbb{R}^{d\times m}}$

give them the same importance

Defn transformation $X' = XU\Lambda^{-\frac{1}{2}} = \left(\frac{1}{\sqrt{\lambda_j}} x_i^T u_j \right)$ $\quad X^TX = U\Lambda U^{-1}$

$\begin{cases} \mathbb{E}[X'] = 0 \\ \mathrm{Cov}(X') = I \end{cases}$

normalize the score

$\boxed{\text{whitening}} \longrightarrow$ $\boxed{\text{去相关性}}$

对每个sample的feature做线性组合来消除相关性

PCA $\longrightarrow$ $\boxed{\text{Compression Algorithm}}$

$$Z_m = X \cdot U_m \longrightarrow \text{project on } \underline{\underline{\text{first-}m\text{-component}}}$$

Recover $\longrightarrow$ $X^* = Z_m U_m^T$ Note $U_m \cdot U_m^T = \begin{pmatrix} I & \\ & 0 \end{pmatrix}$

$\downarrow$ Generalization

(AE) $\longrightarrow$ <u>Auto-Encoder</u>

$$\begin{cases} \underline{\text{Encoder}} : & T_{enc}(X;\theta) = Z_m \\ \underline{\text{Decoder}} : & T_{dec}(Z_m;\theta) = X^* \end{cases}$$

$$\begin{tikzcd} X \in \mathbb{R}^{N\times d} \arrow[r, "U_m^T", "\text{Encoder}"'] \arrow[dr, leftrightarrow] & Z \in \mathbb{R}^{N\times m} \arrow[d, "U_m", "\text{Decoder}"'] \\ & \tilde{X} \in \mathbb{R}^{N\times d} \end{tikzcd}$$

$\downarrow$ Generalization

$\boxed{\text{RNN Encoder-Decoder Architecture}}$

What have done in this class?

① $\mathcal{D} = \{(x_i)\}_{i=1}^N \Longrightarrow \underline{X = (x_i^T) \in \mathbb{R}^{N\times d}}$ $\boxed{\bar{X} := \begin{pmatrix} \bar{x}^T \\ \vdots \\ \bar{x}^T \end{pmatrix}}$

projection in direction $\underline{u} \in \mathbb{R}^d$: $\begin{cases} Z = Xu \in \mathbb{R}^N \\ \bar{Z} = \bar{X}u \in \mathbb{R}^N \end{cases}$

then $\mathrm{var}(Z) = \frac{1}{N}(Z-\bar{Z})^T(Z-\bar{Z})$

$= \frac{1}{N} u^T (X-\bar{X})^T (X-\bar{X}) u$

$= \boxed{\frac{1}{N} u^T \tilde{X}^T \tilde{X} u} \longrightarrow$ ☆ <u>this means:</u>

we can normalize the data first,

and then achieve the optimal projection direction → for the original data

$$\boxed{\max_{u} \ var(Z) \quad s.t \ \|u\|_2^2 = 1} \iff \hat{u} = \text{largest eigenvector}$$

$$\boxed{\text{score } \hat{z} = \tilde{X}\hat{u}} \longrightarrow \boxed{\text{under centralization}}$$

② Generally speaking, given ① $X \xrightarrow{\text{centralization}} \tilde{X}$

② $\boxed{S = \tilde{X}^T \tilde{X}} \longrightarrow$ covariance mat

③ achieve first m eigenvalue & vector

$(\lambda_1, u_1), \ldots, (\lambda_m, u_m)$

④ $\boxed{U_m = (u_1, \ldots, u_m)} \longrightarrow$ orthogonal basis

$Z = X U_m \in \mathbb{R}^{N \times m}$ $\boxed{\text{score matrix}}$

explained variance $= \sum_{i=1}^{m} \lambda_i$

$Z = \begin{pmatrix} z_1^T \\ \vdots \\ z_N^T \end{pmatrix}$ $z_1^T \longrightarrow$ the coordinates for sample 1

$\boxed{\text{Encoder}}$

$$\begin{cases} x \xrightarrow{U_m^T} z = U_m^T x \\ X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix} \longrightarrow Z = X \cdot U_m \end{cases}$$

$\boxed{\text{Decoder}}$

$$\begin{cases} z \xrightarrow{U_m} x^* = U_m z \\ Z = \begin{pmatrix} z_1^T \\ \vdots \\ z_N^T \end{pmatrix} \longrightarrow \boxed{X^* = Z \cdot U_m^T} \end{cases}$$

$\boxed{\text{Consider the 2nd formulation}}$

formulation of problem

$\{u_1, \ldots, u_d\} \longrightarrow$ orthogonal basis

$\boxed{x = \sum_{i=1}^{d} (u_i^T x) \cdot u_i}$ $\quad \tilde{x} = \sum_{i=1}^{m} \beta_i u_i \approx x$

optimization

$$\boxed{\begin{aligned} &\min_{u, \beta_i} \ \frac{1}{N} \sum_{i=1}^{N} \| x_i - \tilde{x}_i \|_2^2 \\ &s.t \ \{u_1, \ldots, u_d\} \text{ orthogonal basis} \end{aligned}}$$

$\boxed{\text{linear projection}}$

① $x \longrightarrow Px$ 坐标变换

① $X \longrightarrow X \cdot P^T$ 坐标阵

$\boxed{X = (x_i^T)_{i \in I}}$

$$x_i - \tilde{x}_i = \sum_{j=1}^{m} \left[(u_j^T x_i) - \beta_{ij} \right] \cdot u_j + \sum_{j=m+1}^{d} u_j^T x_i u_j$$

equivalent

$$\boxed{\begin{aligned} &\min_{u, \beta_i} \ \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{m} (\beta_{ij} - u_j^T x_i)^2 + \sum_{j=m+1}^{d} (u_j^T x_i)^2 \right) \\ &s.t \ \ U \text{ orthogonal} \end{aligned}}$$

$\Downarrow$

for $i=1,\dots,N$, $j=1,\dots,m$. $\boxed{\beta_{ij} = u_j^T x_i}$

$$\boxed{\begin{aligned} &\min_{u} \ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=m+1}^{d}(u_j^T x_i)^2 \\ &\text{s.t } \{u_1,\dots,u_d\} \text{ orthogonal} \\ &\beta_{ij} = u_j^T x_i \ (\text{score}) \end{aligned}}$$

$\Downarrow$

One choice of $\underline{\hat{u}_1,\dots,\hat{u}_d}$ is the eigenvector of

$$\boxed{S = \frac{1}{N}X^T X \in \mathbb{R}^{d\times d}}$$

$\underline{Z_m = X\cdot U_m} \rightarrow Z_m = \begin{pmatrix} z_1^T \\ \vdots \\ z_N^T \end{pmatrix}$

$U_m = (u_1,\dots,u_m)$

score (coordinates) given basis $\hat{u}_1,\dots,\hat{u}_m$

$$\boxed{x \xrightarrow{U_m^T} z \xrightarrow{U_m} \tilde{x}}$$

$$\tilde{X} = \begin{pmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_N^T \end{pmatrix} = X U_m U_m^T = \begin{pmatrix} z_1^T \\ \vdots \\ z_N^T \end{pmatrix}\cdot U_m^T$$

then, $\frac{1}{N}\sum_{i=1}^{N}\|x_i - \tilde{x}_i\|_2^2$

$$= \boxed{\sum_{j=m+1}^{d}\lambda_j}$$

$\downarrow$

un explained variance

= average compression loss