8. Online learning

PAC: $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ $\hat{y} \sim y$

训练 + 测试

$x \to \boxed{h} \to \hat{y}$

i.i.d. $\mathcal{D}$

Online: ① mixing traing and testing phases

② no distribution assumption

③ time-series (T)

T rounds: at the tth round:

① receive $x_t \in \mathcal{X}$

② make prediction $\hat{y}_t \in \mathcal{Y}$

③ receive $y_t \in \mathcal{Y}$

④ incur loss $L(\hat{y}_t, y_t)$

⑤ update the parameters

$$L = \sum_{t=1}^{T} L(\hat{y}_t, y_t)$$

Online learning with expert advice → model

at the tth round, receiving $x_t \in \mathcal{X}$

receiving advice $y_{t,i} \in \mathcal{Y}$. $i \in [N]$ ⇒ N expert.

⇒ prediction $\hat{y}_t = f_t(y_{t,1}, \cdots, y_{t,N}) \sim y_t$

def (Regret) 遗憾

$$R_T = \underbrace{\sum_{t=1}^{T} L(\hat{y}_t, y_t)}_{\text{当前算法的累计损失}} - \underbrace{\min_{i=1}^{N} \sum_{t=1}^{T} L(\hat{y}_{t,i}, y_t)}_{\text{"最好"算法的累计损失}}$$

遗憾界 (regret bound)

e.g. 天气预报

advice →

晴	雨	多云	晴
bbc.com	cctv.com	bilibili.com	sohutv.com

Loss: Zero-one loss $\sum_{t=1}^{T} \mathbb{1}_{\hat{y}_t \neq y_t} = \text{mistakes}(A)$

mistake (regret under 0/1 loss)

concept: $M_A(c) = \max_{x_1,\dots,x_T} |\text{mistakes}(A, c)|$

concept class: $M_A(C) = \max_{c \in C} M_A(c)$ bound

① realizable case (at least one of the experts makes no errors)

Halving ($\mathcal{H}$) — the set of all experts!

1. $\mathcal{H} \leftarrow H$
2. for $t = 1$ to T do
   - Receive (x_t)
   - $\hat{y}_t = \text{MajorityVote}(\mathcal{H}_t, x_t)$ — 多数投票的方法得到预测
   - Receive (y_t)
   - if $\hat{y}_t \neq y_t$ then: — $\mathbb{1}_{y_t \neq \hat{y}_t}$
     - ☆ $\mathcal{H}_{t+1} \leftarrow \{h \in \mathcal{H}_t : h(x_t) = y_t\}$
   - else $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t$.
3. return H_{T+1}

↳ 预测错误 则把投大多数票的experts从假设空间中删去

⇒ 预测正确 无事发生;

Thm Let $\mathcal{H}$ be finite hypothesis set: $M_T = M_{\text{Halving}}(\mathcal{H}) \leq \log_2 |\mathcal{H}| = \log_2 N$

T轮总错误数

(at each mistake, the active set is reduced by at least half)

Let $\mathrm{opt}(\mathcal{H})$: optimal mistake bound for $\mathcal{H}$ (最多只需犯多少个错误就能找到正确答案)

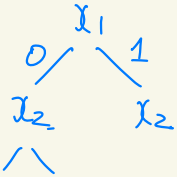
$$\mathrm{VCdim}(\mathcal{H}) \le \mathrm{opt}(\mathcal{H}) \le M_{\mathrm{Halving}}(\mathcal{H}) \le \log_2 N$$

$\mathrm{VCdim}(\mathcal{H}) = d$ (这 d 个样本每一种 label 都有对应的 prediction 存在)

w.c

"强制算法在每一层都犯错"

adversarial argument



② non-realizable setting

Weight-Majority (WM) algorithm $\Rightarrow$ weight vector $\in \mathbb{R}^N$

$\mathbb{1}_{\sum_{y_i=1} w_i \ge \sum_{y_i=0} w_i}$ (加权投票)

Weighted-Majority (N)

1. for $i = 1$ to N do

   $w_{1,i} \leftarrow 1$ uniform

2. for $t = 1$ to T do

   Receive (x_t)

   $\hat{y}_t = \mathbb{1}_{\sum_{i: y_{t,i}=1} w_{t,i} \ge \sum_{i: y_{t,i}=0} w_{t,i}}$ 加权多数投票

   Receive (y_t)

   if $\hat{y}_t \ne y_t$ then

   for $i = 1$ to N do

   if $(y_{i,t} \ne y_t)$ then

   $w_{t+1,i} \leftarrow \beta \cdot w_{t,i}$ 预测错误 weight decay ($\beta = 0$: Halving)

   else $w_{t+1,i} \leftarrow w_{t,i}$ 预测正确 weight maintain.

3. return w_{T+1}

Thm $\beta \in (0,1)$ m_T: # mistakes made by algorithm WM.

m_T^*: # mistakes made by the best of N experts (T rounds)

$$\Rightarrow \boxed{m_T \le \frac{\log N + m_T^* \log \frac{1}{\beta}}{\log \frac{2}{1+\beta}}}$$

$\Rightarrow m_T = O(\log N) + \text{constant} \times |\text{mistakes of best expert}|$

Regret: $m_T - |\text{mistakes of best expert}|$

Proof: ① potential function ② upper bound. lower bound ③ combine

potential function method 势函数法

$W_t = \sum_{i=1}^{N} W_{t,i}$

$\beta \in (0,1)$

upper: if makes mistake at round t, $W_{t+1} \le \left(\frac{1}{2} + \frac{1}{2}\beta\right) W_t = \frac{1+\beta}{2} \cdot W_t$

$W_1 = N$ m_T mistakes $\Rightarrow W_T \le \left(\frac{1+\beta}{2}\right)^{m_T} N$

lower: $\forall i$. $W_T \ge W_{T,i} = \beta^{m_{T,i}} \ge \beta^{m_T^*}$

$\Rightarrow \beta^{m_T^*} \le \left(\frac{1+\beta}{2}\right)^{m_T} \cdot N \Rightarrow m_T^* \log \beta \le m_T \cdot \log \frac{1+\beta}{2} + \log N$

$\Rightarrow m_T \le \frac{m_T^* \log \beta + \log N}{\log \frac{1+\beta}{2}}$ $\square$

realizable: $m_T^* = 0$.

$m_T = O(\log N)$

Halving WM $\not\Rightarrow$ deterministic in 0/1-loss:

cannot achieve $R_T = o(T)$ over all sequence

T轮: adverserial! $\left\langle \begin{array}{l} 0 \to 1 \\ 1 \to 0 \\ \cdots \end{array} \right.$ $\Rightarrow m_T = T$, $m_T^* \le \frac{T}{2}$ $\Rightarrow R_T = m_T - m_T^* \ge \frac{T}{2}$

③ Stochastic. (randomized algorithm)

a set $A = \{1, \cdots, N\}$ of N actions (experts)

at each round $t \in [T]$, $\mathcal{A}$ selects a $\boxed{\text{distribution } p_t}$ over a set of actions

receives a loss vector l_t where $l_{t,i} \in [0,1]$ is the loss with action i

$$L_t = \sum_{i=1}^{N} p_{t,i}\, l_{t,i} = \mathbb{E}_{p_t}[l_t]$$

cumulative loss: $L_T = \sum_{t=1}^{T} L_t$

$$L_{T,i} = \sum_{t=1}^{T} l_{t,i}$$

$$\Downarrow$$

$$L_T^{min} = \min_{i \in A} L_{T,i}$$

def (Regret) $R_T = L_T - L_T^{min}$ △

assume $l_{t,i} \in \{0,1\}$ is zero-one loss

Randomized-Weighted-Majority(N)

```
for i = 1 to N do
    W1,i = 1
    P1,i = 1/N
for t = 1 to T do
    Receive(lt)
    for i = 1 to N do
        if lt,i = 1 then
            Wt+1,i = β · Wt,i
        else Wt+1,i = Wt,i
    Wt+1 = Σ_{i=1}^{N} Wt+1,i
    for i = 1 to N do:
        pt+1,i = Wt+1,i / Wt+1
```

⇒ 根据weights更新 probability

3 return W_{T+1}.

Thm $\beta \in [\frac{1}{2}, 1)$ RWM: $L_T \le \frac{\log N}{1-\beta} + (2-\beta) \cdot L_T^{min}$

Proof: $W_t = \sum_{i=1}^{N} w_{t,i}$

$$W_{t+1} = \sum_{i: l_{t,i}=0} w_{t,i} + \beta \cdot \sum_{i: l_{t,i}=1} w_{t,i}$$

$$= W_t + (\beta - 1) \cdot \sum_{i: l_{t,i}=1} w_{t,i}$$

$$= W_t + (\beta-1) \cdot W_t \cdot \sum_{i: l_{t,i}=1} p_{t,i} = W_t + (\beta-1) \cdot W_t \cdot \sum_i p_{t,i} l_{t,i}$$

$$= W_t + (\beta-1) \cdot W_t \cdot L_t = W_t \cdot (1-(1-\beta) L_t)$$

$$W_1 = N \Rightarrow W_{T+1} = N \cdot \prod_{t=1}^{T} (1-(1-\beta) L_t)$$

lower bound: $W_{T+1} \ge \max_{i \in [N]} w_{T+1,i} = \beta^{L_T^{min}}$ (权重越大, 权重越高)

$$\beta^{L_T^{min}} \le N \cdot \prod_{t=1}^{T} (1-(1-\beta) L_t) \Rightarrow L_T^{min} \cdot \log \beta \le \log N + \sum_{t=1}^{T} \log(1-(1-\beta) L_t)$$

$$\le \log N - (1-\beta) \cdot \sum_{t=1}^{T} L_t$$

$\log(1-x) \le -x$

$$= \log N - (1-\beta) L_T$$

$$\Rightarrow L_T \le \frac{\log N}{1-\beta} - \frac{\log \beta}{1-\beta} L_T^{min}$$

$\le 2-\beta$

只需证 $-\frac{\log \beta}{1-\beta} \le 2-\beta \Rightarrow (2-\beta)(1-\beta) + \log \beta \ge 0$

$$\Rightarrow L_T \le \frac{\log N}{1-\beta} + (2-\beta) \cdot L_T^{min} \quad \square.$$

$f(\beta) = \beta^2 - 3\beta + 2 + \log \beta$

$f'(\beta) = 2\beta - 3 + \frac{1}{\beta} \le 0$

$\Rightarrow f(\beta) \ge f(\frac{1}{2}) = 2 - \log 2 + \frac{1}{4} - \frac{3}{2}$

$= 0.75 - \log 2 > 0$

$$L_T \le \frac{\log N}{1-\beta} + (1-\beta) L_T + L_T^{min}$$

$$\le \frac{\log N}{1-\beta} + (1-\beta) \cdot T + L_T^{min}$$

$\beta = \max\left\{\frac{1}{2}, 1-\sqrt{\frac{\log N}{T}}\right\} \Rightarrow \boxed{R_T = \mathcal{L}_T - \mathcal{L}_T^{min} \le O(\sqrt{T\log N})}$ by RWM 算法

Average Regret per round. $\frac{R_T}{T} = O\left(\frac{1}{\sqrt{T}}\right)$ $\boxed{\text{Optimal}}$ ☆☆

Thm $N=2$. $\exists$ stochastic sequence of losses, for which, the regret of any on-line learning algorithms verifies $\mathbb{E}[R_T] \ge \sqrt{T/8}$

Proof: $\forall t \in T$. $l_t = \begin{cases} l_{01} = (0,1)^T & \text{w.p. } \frac{1}{2}. \\ l_{10} = (1,0)^T & \text{w.p. } \frac{1}{2}. \end{cases}$ $\vec{p}_t = \begin{pmatrix} p_{t,1} \\ p_{t,2} \end{pmatrix}$

$$\mathbb{E}[\mathcal{L}_T] = \mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot l_t\right] = \sum_{t=1}^{T} p_t \cdot \mathbb{E}[l_t] = \sum_{t=1}^{T} \frac{1}{2}p_{t,1} + \frac{1}{2}(1-p_{t,1}) = T/2$$

$$\mathcal{L}_T^{min} = \min\{\mathcal{L}_{T,1}, \mathcal{L}_{T,2}\} = \frac{1}{2}\left(\mathcal{L}_{T,1} + \mathcal{L}_{T,2} - |\mathcal{L}_{T,1} - \mathcal{L}_{T,2}|\right)$$

$$= \frac{T}{2} - \left|\mathcal{L}_{T,1} - \frac{T}{2}\right|$$

$\mathcal{L}_{T,1} + \mathcal{L}_{T,2} = T$

$$\Rightarrow \mathbb{E}[R_T] = \mathbb{E}[\mathcal{L}_T] - \mathbb{E}[\mathcal{L}_T^{min}] = \mathbb{E}\left[\left|\mathcal{L}_{T,1} - \frac{T}{2}\right|\right]$$

R.V. $\sigma_i \in \{-1, 1\}$ ($\frac{1}{2}$, $\frac{1}{2}$)

$$\mathcal{L}_{T,1} = \sum_{t=1}^{T} \frac{\sigma_t + 1}{2} = \frac{T}{2} + \frac{1}{2}\sum_{t=1}^{T}\sigma_t$$

$$\Rightarrow \mathbb{E}[R_T] = \mathbb{E}\left[\left|\sum_{t=1}^{T}\frac{1}{2}\sigma_t\right|\right] \triangleq \mathbb{E}\left[\left|\sum_{t=1}^{T}\frac{1}{2}\sigma_t x_t\right|\right]$$

$x_t = \frac{1}{2}$. $\forall t$

$$\ge \sqrt{\frac{1}{2}\sum_{t=1}^{T} x_t^2} = \sqrt{T/8} \quad \square$$

☆-lower bound $R_T = \Omega(\sqrt{T\log N})$

Khinchine-Kahane inequality

$$\frac{1}{2}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{m}\sigma_i x_i\right\|^2\right] \le \left(\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{m}\sigma_i x_i\right\|\right]\right)^2 \le \mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{m}\sigma_i x_i\right\|^2\right]$$

Jensen's inequality

Perceptron Algorithm. 感知机算法

1950s 第一代神经网络

$x \rightarrow \boxed{w^1, w^2, \vdots, w^N}$ $\sum_{i=1}^{N} w^i x^i = w \cdot x \rightarrow \mathrm{sgn}(w \cdot x)$

XOR problem

	0	1
1		
0		

$y_t \cdot h_t(x_t)$: confidence margin

Perceptron (w_0)

1. $w_1 \leftarrow w_0$ (typically $w_0 = 0$)
2. for $t = 1$ to T do
   - Receive (x_t)
   - $\hat{y}_t = \mathrm{sgn}(w_t \cdot x_t)$
   - Receive (y_t)
   - if $\hat{y}_t \neq y_t$ then
     - $w_{t+1} = w_t + \eta \cdot y_t \cdot x_t$
   - else $w_{t+1} = w_t$.
3. return w_{T+1}.

$y_t(w_t \cdot x_t) > 0$ 正确

$y_t(w_t \cdot x_t) < 0$ 错误

⇒ 把 $y_t(w_t \cdot x_t)$ 往正确方向上推.

$y_t(w_{t+1} \cdot x_t) = y_t(w_t \cdot x_t) + \eta \cdot \|x_t\|^2$ > 0

犯错, 则 update.

1970s: 第二代神经网络 (Artificial Neural Network)

(MLP) Multi-layer Perceptron

$\sigma(wx)$

Backpropagation. 梯度

CNN

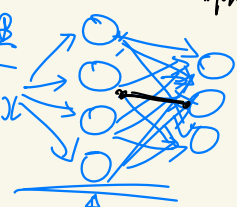
RNN LSTM 可有遗忘机理

Transformer

self-attention

FFN (MLP)

$f(\cdot)$ arbitrary



GAN. VAE. Normalizing Flow. Diffusion.

"第三代神经网络": Spiking Neural Network

ANN2SNN ? → binary activation → thresholding mechanism.

confidence margin

$$F(w) = \frac{1}{T}\sum_{t=1}^{T} \max(0, -y_t(w \cdot x_t))$$

$$= \mathbb{E}_{x \sim \hat{D}}\left[\max(0, -y(w \cdot x))\right]$$

perceptron algorithm minimize

F: convex ✓

differentiable ? ✗

⇓

stochastic subgradient descent

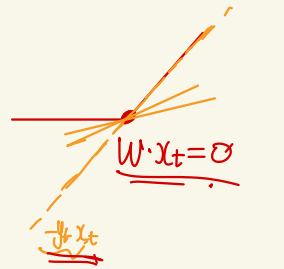
次梯度下降

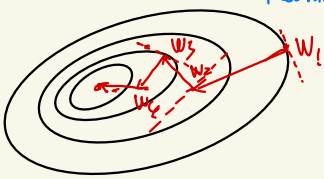
惩罚所有错误分类的项

$$w_{t+1} = w_t - \eta \cdot \delta_w\left(\max(0, -y_t(w\cdot x_t))\right) = \begin{cases} w_t - \eta \cdot \nabla_w\left(\max(0, -y_t(w\cdot x_t))\right), & w\cdot x_t \neq 0 \\ w_t + \eta \cdot y_t \cdot x_t. & w\cdot x_t = 0 \end{cases}$$

$$= \begin{cases} w_t, & y_t(w_t\cdot x_t) > 0 \Rightarrow \text{分类正确, 不变} \\ w_t + \eta \cdot y_t \cdot x_t, & y_t(w_t\cdot x_t) \le 0 \Rightarrow \text{分类错误, 更新(Update)} \end{cases}$$

learning rate step size (gradient)





GD: 每次用所有样本
SGD: 每次用一次样本 (Online learning)

$\Rightarrow$ Perceptron 可以看成对 $F(w)$ 做梯度下降的优化!

Guarantees: $\|x_t\| \le r$, $\eta = 1$

Thm 1 Assume $\exists \rho > 0, v \in \mathbb{R}^N$, s.t. $\forall t \in [T]$, $\rho \le \dfrac{y_t(v\cdot x_t)}{\|v\|}$ "geometric margin"

$\Rightarrow$ # of updates made by Perceptron algorithm processing $x_1, \cdots, x_T$ is bounded by $\dfrac{r^2}{\rho^2}$

i.e. $M(S) \le \dfrac{r^2}{\rho^2}$ update 是错误的

Proof: J is the subset of T at which there is an update. $|J| = M(S)$

$$M\cdot\rho \le \frac{v\cdot\sum_{t\in J} y_t x_t}{\|v\|} \le \left\|\sum_{t\in J} y_t x_t\right\| \quad \text{(Cauchy-Schwarz)}$$

$$= \left\|\sum_{t\in J}(w_{t+1} - w_t)\right\| \quad \text{(definition of updates)}$$

$$= \|w_{T+1}\|$$

$$= \sqrt{\sum_{t\in J}\|w_{t+1}\|^2 - \|w_t\|^2} \quad \text{(telescoping)}$$

$$= \sqrt{\sum_{t\in J}\|w_t + y_t x_t\|^2 - \|w_t\|^2} = \sqrt{\sum_{t\in J} 2y_t w_t x_t + \|x_t\|^2} \le \sqrt{\sum_{t\in J}\|x_t\|^2} \le \sqrt{M r^2}$$

$(y_t^2 = 1)$ ≤ 0 犯错

$\Rightarrow M \le r^2/\rho^2$ □

output $W_{T+1} = \sum_{t \in T} y_t x_t$. 只与"贡献"最大的样本有关 ⇒ Sparsity !

↳ support vectors in Perceptron.

$$\boxed{\text{SVM} \quad w = \sum_{i \in SV} \alpha_i y_i x_i}$$

Remark: ① $M \le \frac{r^2}{\rho^2}$ 与N无关

② $M = \Theta\left(\frac{r^2}{\rho^2}\right)$ 紧!

③ 与数据 $x_1, \cdots, x_T$ 无关

$S = \{(x_1, y_1), \cdots, (x_T, y_T)\}$

↳ linearly separable: Perceptron alg. converges (after a finite # of updates)

linearly non-separable: Perceptron alg. NOT converge.

⇒ stopping time T

Thm② (Perceptron算法的泛化误差界)

classical PAC → i.i.d sampling

h_S: output of Perceptron algorithm ⇒ $\boxed{\mathbb{E}_{S\sim D^m}[R(h_S)] \le \mathbb{E}_{S\sim D^{m+1}}\left[\frac{\min\left(M(S), \frac{r_S^2}{\rho_S^2}\right)}{m+1}\right]}$

$h_S(x) = \mathrm{sgn}(w_S x)$

r_S: radius of smallest origin-centered sphere containing all points in S

ρ_S: largest margin of a separating hyperplane for S.

$M_S \le \frac{r_S^2}{\rho_S^2}$

average-case, "weak"

Proof: LOO (Leave-One-Out)

$$\hat{R}_{LOO}(h_S) = \frac{1}{m+1}\sum_{i=1}^{m+1} \mathbb{1}_{h_{S-\{x_i\}}(x_i) \ne y_i} \le \frac{M(S)}{m+1}$$

if $h_{S-\{x\}}$ misclassifies x then x must be a support vector of S

$$\Rightarrow \mathbb{E}_{S\sim D^m}[R(h_S)] \le \mathbb{E}_{S\sim D^{m+1}}\left[\frac{M(S)}{m+1}\right] \quad \square$$

⇓

SVM assume no offset ($b=0$)

Thm 3 (SVM算法) $\mathbb{E}_{S\sim D^m}[R(h_S)] \le \mathbb{E}_{S\sim D^{m+1}}\left[\frac{\min\left(N_{SV}(S),\ r_S^2/\rho_S^2\right)}{m+1}\right]$

support vectors

Proof: $\mathbb{E}_{S\sim D^m}[R(h_S)] \le \mathbb{E}_{S\sim D^{m+1}}\left[\frac{N_{SV}(S)}{m+1}\right]$ by LOO analysis.

$S=(x_1,\cdots,x_{m+1})$ linearly separable. i.i.d. $\sim D$

$x\in S$ is a misclassified sample by $h_{S-\{x\}}$: $(x=x_{m+1})$ assume

$S'=(x_1,\cdots,x_m)$

$q\in[m+1]$: $G_q(\alpha)=\sum_{i=1}^{q}\alpha_i-\frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{q}\alpha_i\alpha_j y_i y_j x_i^T x_j$

G_{m+1} : SVM's objective function, associated to S.

G_m : SVM's objective function $\sim$ S'.

$\alpha\in\mathbb{R}^{m+1}$: solution of $\max_{\alpha\ge 0} G_{m+1}(\alpha)$

$\alpha'\in\mathbb{R}^{m+1}$: $\begin{cases}(\alpha_1',\cdots,\alpha_m')^T \text{ solution of } \max_{\alpha\ge 0} G_m(\alpha)\\ \alpha'_{m+1}=0\end{cases}$

Observe that $\max_{\beta\ge 0} G_{m+1}(\alpha'+\beta\cdot e_{m+1}) \le G_{m+1}(\alpha)$

$G_{m+1}(\alpha-\alpha_{m+1}e_{m+1}) \le G_m(\alpha')$

$A=G_{m+1}(\alpha)-G_m(\alpha')$ $\begin{cases}\le G_{m+1}(\alpha)-G_{m+1}(\alpha-\alpha_{m+1}e_{m+1}) & \text{(Upper Bound)}\\ \ge \max_{\beta\ge 0} G_{m+1}(\alpha'+\beta e_{m+1})-G_m(\alpha') & \text{(Lower Bound)}\end{cases}$

$w=\sum_{i=1}^{m} y_i\alpha_i x_i$ by S

h_S misclassifies x_{m+1}, x_{m+1} must be a support vector for h_S.

$$\Rightarrow \underline{y_{m+1} W \cdot x_{m+1} = 1}$$

$$G_q(\alpha) = \sum_{i=1}^{q} \alpha_i - \frac{1}{2} \sum_{i=1}^{q} \sum_{j=1}^{q} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

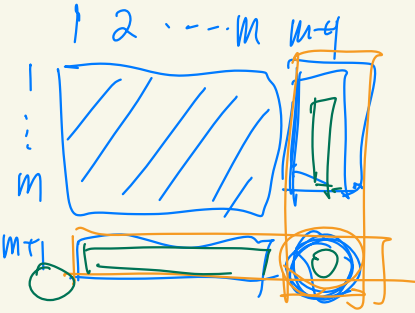
$1 \sim m$

Upper Bound: $G_{m+1}(\alpha) - G_{m+1}(\alpha - \alpha_{m+1} e_{m+1})$

$$= \alpha_{m+1} - \sum_{i=1}^{m+1} (y_i \alpha_i x_i) \cdot (y_{m+1} \alpha_{m+1} x_{m+1}) + \frac{1}{2} \alpha_{m+1}^2 \cdot \|x_{m+1}\|^2$$

$$= \alpha_{m+1} \left(1 - y_{m+1} W \cdot x_{m+1} \right) + \frac{1}{2} \alpha_{m+1}^2 \|x_{m+1}\|^2$$

$$= \frac{1}{2} \alpha_{m+1}^2 \|x_{m+1}\|^2$$



Lower Bound: $\forall \beta > 0$, $G_{m+1}(\alpha' + \beta e_{m+1}) - G_m(\alpha')$

$$= \beta - \sum_{i=1}^{m} (y_i \alpha_i' x_i) \cdot (y_{m+1} \beta x_{m+1}) - \frac{1}{2} \beta^2 \|x_{m+1}\|^2$$

$$= \beta \cdot \left(1 - y_{m+1} \cdot W' \cdot x_{m+1} \right) - \frac{1}{2} \beta^2 \|x_{m+1}\|^2$$

$$\leq \frac{\left(1 - y_{m+1} W' x_{m+1}\right)^2}{2 \|x_{m+1}\|^2}$$

= 以函数 (maximize by $\beta = \dfrac{1 - y_{m+1} W' x_{m+1}}{\|x_{m+1}\|^2}$)

$$\Rightarrow \frac{\left(1 - y_{m+1} W' x_{m+1}\right)^2}{2 \|x_{m+1}\|^2} \leq A \leq \frac{1}{2} \alpha_{m+1}^2 \|x_{m+1}\|^2$$

$$\Rightarrow \alpha_{m+1} \geq \frac{|1 - y_{m+1} W' x_{m+1}|}{\|x_{m+1}\|^2} = \frac{1 - y_{m+1} W' x_{m+1}}{\|x_{m+1}\|^2}$$

$$\geq \frac{1}{\|x_{m+1}\|^2} \geq \frac{1}{r_S^2}$$

W' by $S' = (x_1, \cdots, x_m)$

$\Rightarrow h_{S'} = W' \cdot x$ misclassifies x_{m+1}

$\Rightarrow y_{m+1} \cdot W' \cdot x_{m+1} < 0$

J: $h_{S - \{x_i\}}(x_i) \neq y_i$ 而样本下移 i 的集合 $\Rightarrow \sum_{i \in J} \alpha_i \geq \frac{|J|}{r_S^2}$

$i \in SV$: $\alpha_i \neq 0$. $\quad w x_i + b = y_i \Rightarrow b = y_i - w x_i$

$$\Rightarrow \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{=0} b = \sum_{i=1}^{m} \alpha_i y_i^2 - \sum_{i=1}^{m} \alpha_i y_i w x_i$$

$$\Rightarrow \sum_{i=1}^{m} \alpha_i = w \cdot \sum_{i=1}^{m} \alpha_i y_i x_i = \|w\|^2 \leq \frac{1}{\rho_S^2}$$

$\rho_S = \frac{1}{\|w\|}$ 最大

$$\Rightarrow \frac{|J|}{r_S^2} \le \sum_{t\in J} \alpha_t \le \sum_{t=1}^{m_4} \alpha_t \le \frac{1}{\rho_S^2} \Rightarrow |J| \le \frac{r_S^2}{\rho_S^2}$$ then LOO analysis □

non-linearly separable Perceptron

⇒ "引入松弛项"

Thm 4 J: Perceptron makes mistakes. $M=|J|$

$y_t(w\cdot x_t)$ $\begin{cases} <0: \text{错误分类} \\ 0\sim\rho: \text{不自信} \\ >\rho: \text{正确分类} \end{cases}$

$$M \le \inf_{\substack{\rho>0 \\ \|v\|_2\le 1}} \left(\frac{\frac{r}{\rho} + \sqrt{\frac{r^2}{\rho^2} + 4\|l_\rho\|_1}}{2} \right)^2 \le \inf_{\substack{\rho>0 \\ \|v\|_2\le 1}} \left(\frac{r}{\rho} + \sqrt{\|l_\rho\|_1} \right)^2$$

Jensen's inequality

where $\underbrace{l_\rho = (l_t)_{t\in J}}_{\text{loss}}$. $l_t = \max\left(0, 1 - \frac{y_t(v\cdot x_t)}{\rho}\right)$. "hinge loss"

Proof: $1 - \frac{y_t(v\cdot x_t)}{\rho} \le l_t$. $\Rightarrow \sum_{t\in J} 1 = |J| = M \le \sum_{t\in J} l_t + \sum_{t\in J} \frac{y_t(v\cdot x_t)}{\rho}$

$$\le \|l_\rho\|_1 + \sqrt{Mr^2}/\rho \, \|v\|_2 \le \|l_\rho\|_1 + \sqrt{Mr^2}/\rho$$

$$\Rightarrow M - \frac{r}{\rho}\sqrt{M} - \|l_\rho\|_1 \le 0 \Rightarrow \sqrt{M} \le \frac{1}{2}\left(\frac{r}{\rho} + \sqrt{\frac{r^2}{\rho^2} + 4\|l_\rho\|_1} \right)$$ $\forall v$ □

Thm 5 $$M \le \inf_{\substack{\rho>0 \\ \|v\|_2\le 1}} \left(\frac{r}{\rho} + \|l_\rho\|_2 \right)^2.$$

$$\|x\|_2 \le \|x\|_1 \le \sqrt{N}\cdot\|x\|_2$$ Cauchy-Schwarz.

Proof: $M \le \|l_\rho\|_1 + \frac{\sqrt{Mr^2}}{\rho} \le \sqrt{M}\left(\|l_\rho\|_2 + \frac{r}{\rho}\right)$ □

△ Perceptron algorithm: extension

- PDS kernel → 核方法 ☆
- 对偶问题. ☆
- 加权权重进行学习

$w = \sum_{s=1}^{T} \alpha_s y_s x_s$

分类错误: $w += y_t x_t$.

$(\alpha_t = \alpha_t + 1)$

Winnow Algorithm 扬长避短

$w \mathrel{+}= \eta \cdot y_t x_t$

$w \leftarrow 0$

	Perceptron	Winnow
initialization	0	$1/N$
update	$+$	$\times$
	↓	↓
	SVM	AdaBoost
	(Gradient Descent)	(Coordinate Descent)
	$\|v\|_2$	$\|v\|_1$

Winnow(η)

1. $w_1 \leftarrow 1/N$ (均匀分布)
2. for $t = 1$ to T do

   Receive (x_t)

   $\hat{y}_t = \mathrm{sgn}(w_t x_t)$

   Receive (y_t)

   if $\hat{y}_t \neq y_t$ then

   $Z_t = \sum_{i=1}^{N} w_{t,i} \exp(\eta y_t x_{t,i})$

   for $i = 1$ to N do

   $w_{t+1,i} = \dfrac{w_{t,i} \cdot \exp(\eta y_t x_{t,i})}{Z_t}$

   else $w_{t+1} = w_t$

   normalization

3. return w_{T+1}

若分类错误，则更新各专家的权重

$\rightarrow y_t \cdot x_{t,i} \begin{cases} > 0 : \text{正确} \Rightarrow e^{\eta} \\ < 0 : \text{错误} \Rightarrow e^{-\eta} \end{cases}$

$x_{t,i} \in \{-1, +1\}$: Weighted Majority (WM)

Thm $x_1, \cdots, x_T \in \mathbb{R}^N$. $\|x_t\|_\infty \le r_\infty$.

$$\rho_\infty \le \boxed{\frac{y_t (v \cdot x_t)}{\|v\|_1}}$$ ← l_1-geometric-margin

Let $\eta = \dfrac{\rho_\infty}{r_\infty^2}$. # updates by Winnow algorithm $M = |J| \le \dfrac{2 r_\infty^2}{\rho_\infty^2} \cdot \log N$ ☆

Proof: potential function: $\Phi_t = \sum_{i=1}^{N} \dfrac{v_i}{\|v\|_1} \log \dfrac{\frac{v_i}{\|v\|_1}}{w_{t,i}} = D_{KL}\left(\dfrac{v_i}{\|v\|_1} \,\|\, w_t \right)$

$$\Phi_{t+1} - \Phi_t = \sum_{i=1}^{N} \frac{v_i}{\|v\|_1} \cdot \log \frac{w_{t,i}}{w_{t+1,i}} = 0 \qquad (t \notin J : \text{不更新})$$

$$\Phi_{t+1} - \Phi_t = \sum_{i=1}^{N} \frac{v_i}{\|v\|_1} \log \frac{w_{t,i}}{w_{t+1,i}} = \sum_{i=1}^{N} \frac{v_i}{\|v\|_1} \cdot \log \frac{Z_t}{\exp(\eta y_t x_{t,i})} \qquad (t \in J : \text{更新})$$

$$= \log Z_t - \eta \cdot \sum_{i=1}^{N} \frac{v_i}{\|v\|_1} \cdot y_t x_{t,i}$$

$$= \log\left(\sum_{i=1}^{N} w_{t,i} \exp(\eta y_t x_{t,i})\right) - \eta \cdot \frac{y_t (v \cdot x_t)}{\|v\|_1}$$

$$\leq \log \mathbb{E}_{i \sim w_t}\left[\exp(\eta y_t x_{t,i})\right] - \eta \cdot \rho_\infty$$

$$\mathbb{E}_{i \sim w_t}[x_{t,i}] = \sum w_{t,i} x_{t,i} = w_t \cdot x_t$$

$\mathbb{E}[e^X]$ 能母函数

$$= \log \mathbb{E}_{i \sim w_t}\left[\exp\left(\eta y_t x_{t,i} - \eta y_t w_t \cdot x_t + \eta y_t w_t \cdot x_t\right)\right] - \eta \cdot \rho_\infty$$

↳ 均值为0.

$$\leq \log \exp\left(\frac{\eta^2 (2r_\infty)^2}{8}\right) + \eta y_t w_t \cdot x_t - \eta \cdot \rho_\infty$$

$y_t \cdot \hat{y}_t \leq 0$ (更新)

$$\leq \frac{\eta^2 r_\infty^2}{2} - \eta \cdot \rho_\infty$$

$\mathbb{E}[X] = 0,\ X \in [a, b],$

$$\mathbb{E}[e^X] \leq e^{\frac{(b-a)^2}{8}}$$

(Hoeffding's lemma)

telescopic sum: $\Phi_{T+1} - \Phi_1 \leq M \cdot \left(\frac{\eta^2 r_\infty^2}{2} - \eta \cdot \rho_\infty\right)$

lower bound: $\Phi_1 = \sum_{i=1}^{N} \frac{v_i}{\|v\|_1} \cdot \log \frac{v_i / \|v\|_1}{1/N} = \log N + \underbrace{\sum_{i=1}^{N} \frac{v_i}{\|v\|_1} \log \frac{v_i}{\|v\|_1}}_{-H\left(\frac{v_i}{\|v\|_1}\right)} \leq \log N$

Φ_{T+1} is KL-divergence: $\Phi_{T+1} \geq 0$

$$\Rightarrow \Phi_{T+1} - \Phi_1 \geq -\log N$$

$$\Rightarrow -\log N \leq M \cdot \left(\frac{\eta^2 r_\infty^2}{2} - \eta \rho_\infty\right) \overset{\eta = \frac{\rho_\infty}{r_\infty^2}}{=} -\frac{M}{2} \cdot \frac{\rho_\infty^2}{r_\infty^2} \Rightarrow M \leq \frac{2 r_\infty^2}{\rho_\infty^2} \log N \quad \square$$

$\perp 1$ norm $\Rightarrow$ sparsity!

Winnow is favorable

$\|\cdot\|_p$ input vector x.

$\|\cdot\|_q$ margin vector v

$\Rightarrow \frac{1}{p} + \frac{1}{q} = 1$

When a sparse set of the experts $i \in [N]$ can predict well!

$\Rightarrow$ Winnow: $2 \log N$

Perceptron: N

Perceptron: $p = 2,\ q = 2$

Winnow: $p = \infty,\ q = 1$

$e^{\eta y_t x_{t,i}}$

$\perp 1 \Leftrightarrow$ Winnow AdaBoost LASSO

$\perp 2 \Leftrightarrow$ Perceptron SVM Ridge Regression

Online-to-Batch Conversion 在线-批量转换

在传统PAC框架下构建在线学习所服务 ⇒ 泛化界

Generalization

分布假设:

$$S=\{(x_1,y_1),\cdots,(x_T,y_T)\}\in\{\mathcal{X}\times\mathcal{Y}\}^T \text{ drawn i.i.d. according to } \mathcal{D}$$

"Batch"

[fixed unknown

Regret: $R_T=\sum_{t=1}^{T}L(h_t(x_t),y_t)-\min_{h\in\mathcal{H}}\sum_{t=1}^{T}L(h(x_t),y_t)$

Generalization error: $R(h)=\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(h(x),y)]$

Lemma $L\le M$ (bounded loss function), $h_1,\cdots,h_T$ sequence of hypothesis by online alg $\mathcal{A}$ on S

$\forall\delta>0$. w.p. $\ge 1-\delta$

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T}R(h_t)}_{\text{泛化误差}}\le\underbrace{\frac{1}{T}\sum_{t=1}^{T}L(h_t(x_t),y_t)}_{\text{经验误差}}+M\cdot\sqrt{\frac{2\log\frac{1}{\delta}}{T}}$$

Proof: $\forall t\in[T]$ let $V_t=R(h_t)-L(h_t(x_t),y_t)$

$$\mathbb{E}[V_t\mid x_1,\cdots,x_{t-1}]=R(h_t)-\mathbb{E}[L(h_t(x_t),y_t)\mid h_t]=R(h_t)-R(h_t)=0$$

$V_t\in[-M,+M]$

by Azuma's inequality,

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}V_t\ge\varepsilon\right)\le\exp\left(-\frac{2T\varepsilon^2}{(2M)^2}\right)=\delta$$

$$\Rightarrow\frac{1}{T}\sum_{t=1}^{T}V_t\le M\cdot\sqrt{\frac{2\log\frac{1}{\delta}}{T}}\quad \text{w.p.}\ge 1-\delta \qquad \square$$

所有假设泛化误差的平均

⇓

所有假设平均的泛化误差

鞅 Martingale: $\{Y_i\}$

$\mathbb{E}[Y_{t+1}\mid x_1,\cdots,x_t]=Y_t$

且 $|Y_{t+1}-Y_t|\le C_t$

$$\Rightarrow\Pr\left(\left|\sum_{i=1}^{n}Y_i-\mathbb{E}\left[\sum_{i=1}^{n}Y_i\right]\right|\ge\varepsilon\right)\le 2e^{-\frac{2\varepsilon^2}{\sum_{i=1}^{n}C_i^2}}$$

Martingale Difference.

$V_{t+1}=Y_{t+1}-Y_t$

$\Rightarrow\mathbb{E}[V_t\mid x_1,\cdots,x_{t-1}]=0$

Thm $L \le M$. $L: \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}$ convex w.r.t. its first argument

w.p. $\ge 1-\delta$

$$\star \begin{cases} R\left(\frac{1}{T}\sum_{t=1}^{T} h_t\right) \le \frac{1}{T}\sum_{t=1}^{T} L\left(h_t(x_t), y_t\right) + M\cdot\sqrt{\frac{2\log\frac{1}{\delta}}{T}} \\ R\left(\frac{1}{T}\sum_{t=1}^{T} h_t\right) \le \inf_{h\in\mathcal{H}} R(h) + \frac{R_T}{T} + 2M\cdot\sqrt{\frac{2\log\frac{1}{\delta}}{T}} \end{cases}$$

online-learning 不会太差!

regret convergence rate

Proof: L convex to 1st argument

$$\Rightarrow L\left(\frac{1}{T}\sum_{t=1}^{T} h_t(x), y\right) \le \frac{1}{T}\sum_{t=1}^{T} L\left(h_t(x), y\right).$$

$$\Rightarrow R\left(\frac{1}{T}\sum_{t=1}^{T} h_t\right) \le \frac{1}{T}\sum_{t=1}^{T} R(h_t) \le \frac{1}{T}\sum_{t=1}^{T} L\left(h_t(x_t), y_t\right) + M\cdot\sqrt{\frac{2\log\frac{1}{\delta}}{T}}$$

expectation (Lemma)

$$= \min_{h\in\mathcal{H}} \frac{1}{T}\sum_{t=1}^{T} L\left(h(x_t), y_t\right) + \frac{R_T}{T} + M\cdot\sqrt{\frac{2\log\frac{2}{\delta}}{T}}$$

w.p. $\ge 1-\frac{\delta}{2}$

(def of regret)

$$\le \frac{1}{T}\sum_{t=1}^{T} L\left(h^*(x_t), y_t\right) + \frac{R_T}{T} + M\cdot\sqrt{\frac{2\log\frac{1}{\delta}}{T}}$$

Hoeffding's Inequality

w.p. $\ge 1-\frac{\delta}{2}$

$$\le R(h^*) + M\cdot\sqrt{\frac{2\log\frac{2}{\delta}}{T}} + \frac{R_T}{T} + M\cdot\sqrt{\frac{2\log\frac{2}{\delta}}{T}}$$

$$= R(h^*) + \frac{R_T}{T} + 2M\cdot\sqrt{\frac{2\log\frac{2}{\delta}}{T}}$$

$\forall \varepsilon > 0$. $\exists h^* \in \mathcal{H}$.

s.t. $R(h^*) \le \inf_{h} R(h) + \varepsilon$.

$$\le \inf_{h\in\mathcal{H}} R(h) + \varepsilon + \frac{R_T}{T} + 2M\cdot\sqrt{\frac{2\log\frac{2}{\delta}}{T}}$$

w.p. $\ge 1-\delta$

by arbitrary $\varepsilon > 0$: $R\left(\frac{1}{T}\sum_{t=1}^{T} h_t\right) \le \inf_{h\in\mathcal{H}} R(h) + \frac{R_T}{T} + 2M\cdot\sqrt{\frac{2\log\frac{2}{\delta}}{T}}$ □.

$\Downarrow$

"Online Regret-Minimization Algorithms"

遗憾最小化

WM $R_T = O(T)$

$$\boxed{\frac{R_T}{T} = O\left(\frac{1}{\sqrt{T}}\right) \xrightarrow{T\to\infty} 0}$$

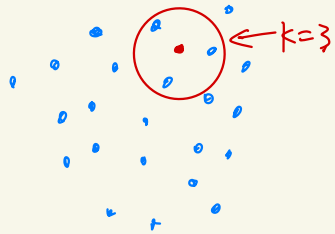
$\hookrightarrow$ OGD

EWA (Exponential Weighted Average)

RWM

15. Dimensionality Reduction

KNN (k-Nearest-Neighbor)



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

分类：投票

回归：平均

① 距离度量 (metric)

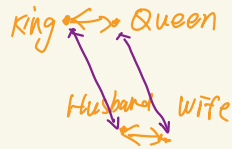
② 无训练过程

⇓

懒惰学习 (lazy learning)

③ k by Cross-Validation

NLP: Word2Vec



④ $k=1$: 测试样本 x, NN: z

error: $R(x) = 1 - \sum_{c \in Y} \Pr(c|x) \Pr(c|z)$

Assume i.i.d. $\forall x, \forall \delta > 0$ $U(x, \delta)$ 邻域内总能找到一个训练样本

密采样 (dense sample)

$$c^* = \underset{c \in Y}{\operatorname{argmax}} \Pr(c|x)$$

$$R(x) = 1 - \sum_{c \in Y} \Pr(c|x) \cdot \Pr(c|z)$$

$$\overset{i.i.d}{=} 1 - \sum_{c \in Y} \Pr(c|x)^2$$

$$\sum_{c \in Y} \Pr(c|x)^2 \geq \Pr(c^*|x)^2$$

$$\leq 1 - \Pr(c^*|x)^2$$

$$= (1 + \Pr(c^*|x))(1 - \Pr(c^*|x))$$

$$\leq 2(1 - \Pr(c^*|x))$$

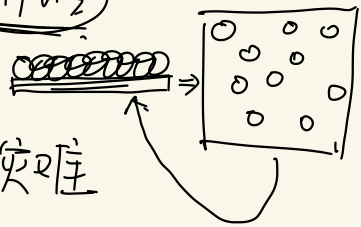
→ KNN的泛化误差不超过 Bayes Optimal 的2倍

$\delta = 0.001$. $N = 1 \Rightarrow 1000$ 个样本

$N = 20 \Rightarrow (10^3)^{20} = 10^{60}$ 个样本

N 成千上万 $\Rightarrow$ 样本稀疏

10^{100}



计算距离不容易

$\Rightarrow$ $\boxed{\text{Curse of Dimensionality}}$ 维数灾难

$\Rightarrow$ 降维 / 维度约简

Dimensionality Reduction

$\mathbb{R}^N \longrightarrow$ "subspace" $\mathbb{R}^K$

embedding 嵌入

- Computational
- Dense $\rightarrow$ feature extraction
- Visualization (tSNE, UMAP)

$S = \{x_1, \cdots, x_m\}$ $x_i \in \mathbb{R}^N$. $X : \mathbb{R}^{N \times m}$ $X = \begin{pmatrix} | & & | \\ x_1 & \cdots & x_m \\ | & & | \end{pmatrix}$

projection $U \in \mathbb{R}^{K \times N}$. $\Rightarrow Y = UX \in \mathbb{R}^{K \times m}$. $Y = \begin{pmatrix} | & & | \\ x_1' & \cdots & x_m' \\ | & & | \end{pmatrix}$

$\boxed{\text{Principal Component Analysis}}$ (PCA 主成分分析)

① X centeralization: $x_i := x_i - \frac{1}{M}\sum_{i=1}^{M} x_i$ $\rightarrow$ 协方差 $C = \frac{1}{m} XX^T$

$P \in \mathcal{P}_K$: set of N-dimensional rank-K orthogonal projection matrices

$$\min_{P \in \mathcal{P}_K} \| P \cdot X - X \|_F^2$$

$P \cdot X$: projected data

X: original data

squared-L2-distance

最边缘构造

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2}$$

Orthogonal Projection Matrix 正交投影矩阵

def V是$\mathbb{R}^n$中的k维线性子空间，正交投影矩阵P将$\mathbb{R}^n$中的向量v投影到V上，即 $Pv = v_p \in V$

"沿着$V^\perp$空间方向进行"

构造 V的标准正交基向量 $u_1, \cdots, u_k$. $\Rightarrow U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix} \in \mathbb{R}^{n\times k}$.

n:

k:

正交投影矩阵 $P = UU^T \in \mathbb{R}^{n\times n}$.

Proof: $v_p = c_1u_1 + \cdots + c_ku_k = U\cdot c$. $c = \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} \in \mathbb{R}^k$.

正交: $\forall u_j, j\in[k]$, $\langle v - v_p, u_j\rangle = 0 \Rightarrow \langle v - Uc, u_j\rangle = 0$

$\Rightarrow \langle v, u_j\rangle = \langle Uc, u_j\rangle = u_j^T Uc$.

而U是由正交基向量组成的矩阵: $u_j^T U = u_j^T \begin{pmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{pmatrix}$

$= (u_j^T u_1 \cdots u_j^T u_k) = (u_j^T u_j)\, e_j^T = e_j^T$.

$U_k^T X$

$\Rightarrow u_j^T v = e_j^T c = c_j$

$\Rightarrow c = U^T v \Rightarrow v_p = Uc = UU^Tv \triangleq Pv \Rightarrow P = UU^T$ □.

正交投影矩阵 $P \in \mathbb{R}^{n\times n}$.

U的列向量是子空间V的标准正交基

- 对称: $P = P^T$. $\langle Pu, v\rangle = \langle u, Pv\rangle$.
- 幂等(idempotent): $P^2 = P$
- 特征值: 0/1. $Px = \lambda x$, $P^2x = Px = \lambda x$.

$U^TU = I$

$UU^T = P$

$U \in \mathbb{R}^{n\times k}$

P本身不是正交的！

$$P^2x = P(Px) = P(\lambda x) = \lambda^2 x.$$

$$\Rightarrow \lambda^2 = \lambda \Rightarrow \begin{cases} \lambda = 0 & x \in V^{\perp} \ (\text{个数 } n-k \text{ 个}) \\ \lambda = 1 & x \in V \ (\text{个数 } k \text{ 个}) \end{cases}$$

为什么需要正交？

① 保持数据方差

② 消除维度之间的相关性

③ 维持矩阵的范数

④ 简化计算

PCA $\boxed{\min_P \|PX - X\|_F^2}$

$X \in \mathbb{R}^{N\times M}$, $P \in \mathbb{R}^{N\times N}$.

$V \to P \to U$ orthogonal projection matrix

$P = UU^T$ $U \in \mathbb{R}^{N\times K}$

Thm $P^* \in \mathcal{P}_k$: PCA solution

$P^* = U_k U_k^T$ where $U_k \in \mathbb{R}^{N\times K}$ is the matrix formed by the top-K eigenvectors of $C = \frac{1}{m} XX^T$

sample covariance matrix

$Y = U_k^T X \in \mathbb{R}^{K\times m}$: k-dimensional representation of X.

Proof: $\|PX - X\|_F^2 = tr(PX-X)^T(PX-X)$

$$= tr(X^T P^2 X - 2X^T P X + X^T X)$$

$$= -tr(X^T P X) + tr(X^T X)$$

$$\Rightarrow \min_{P\in\mathcal{P}_k} \|PX - X\|_F^2 = \max_{P\in\mathcal{P}_k} tr(X^T P X)$$

$$= \max_U tr(X^T U U^T X)$$

$tr(ABC) = tr(CAB)$

$U \in \mathbb{R}^{N\times K}$

$U^T U \in \mathbb{R}^{K\times K}$

$P = UU^T \in \mathbb{R}^{N\times N}$

$$= \boxed{\begin{array}{l} \max_U tr(U^T X X^T U) \\ \text{s.t. } U^T U = I \end{array}}$$

正交 ⇒ 特征正交

$$\boxed{\begin{array}{ll} \max\limits_{U} & tr(U^TXX^TU) \\ \text{s.t.} & U^TU = I \end{array}}$$

$\Rightarrow U_k = \begin{pmatrix} | & & | \\ u_1^k & \cdots & u_k^k \\ | & & | \end{pmatrix}$ 为 XX^T 的最大k个特征向量

$k \times k$

Rayleigh Quotient $\sum_{i=1}^{k} \frac{u_i^T (XX^T) u_i}{u_i^T u_i}$

Courant-Fischer Min-Max

Lagrangian:

$$L(U, \Theta) = -tr(U^TXX^TU) + \langle \Theta, U^TU - I \rangle$$

$$= -tr(U^TXX^TU) + tr(\Theta^T(U^TU - I))$$

$\Theta \in \mathbb{R}^{k \times k}$

仅考虑约束 $u_i^T u_i = 1 \quad (i = 1, \cdots, k)$

正交 $\Rightarrow u_i^T u_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

$\Theta = diag(\lambda_1, \cdots, \lambda_k) \triangleq \Lambda$.

$$L(U, \Lambda) = -tr(U^TXX^TU) + tr(\Lambda^T(U^TU - I)).$$

Matrix Cookbook

$$\frac{\partial L(U, \Lambda)}{\partial U} = -\frac{\partial}{\partial U}\left(tr(U^TXX^TU) + tr(\Lambda^T U^T U)\right)$$

$$= -2XX^TU + 2U\Lambda = 0$$

$\frac{\partial}{\partial X}(tr(X^TBX)) = (B + B^T)X$

$\frac{\partial}{\partial X} tr(BX^TX) = XB^T + XB$

$$\Rightarrow XX^TU = U\Lambda. \Rightarrow XX^Tu_i = \lambda_i u_i \quad (i = 1, \cdots, k)$$

λ_i : eigenvalue

u_i : eigenvector

仅考虑 $u_i^T u_i = 1$ 的约束

由 XX^T 为对称矩阵 $\Rightarrow$ Gram-Schmidt 变换 即可让特征向量正交

$$\max_U tr(U^TXX^TU) = \max_U \sum_{i=1}^{k} u_i^T XX^T u_i$$

$$= \max_U \sum_{i=1}^{k} u_i^T(\lambda u_i) = \max_U \sum_{i=1}^{k} \lambda (u_i^T u_i) = \max_U \sum_{i=1}^{k} \lambda_i$$

取前k个最大的特征值即可!

$U = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{pmatrix}$ $u_1, \dots, u_k$ 为 XX^T 最大的k个特征值对应的特征向量.

$PX = U_k U_k^T X \Rightarrow Y = U_k^T X$ □

特征向量（奇异向量）

$\Updownarrow$

最大k个特征值（奇异值） → variance

最近重构性

$\min \|PX - X\|_F^2$

$\Updownarrow$

最大可分性

$\max \operatorname{tr}(U^T X X^T U)$

project the data onto the subspace with maximal variance

PCA: Input: Sample $S = \{x_1, \dots, x_m\}$, k

1. Center all points $x_i \leftarrow x_i - \frac{1}{m}\sum_{i=1}^{m} x_i$
2. Compute the covariance matrix XX^T of sample.
3. Perform eigenvalue decomposition on the covariance matrix XX^T
4. Take the eigenvectors $U_k = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{pmatrix}$ corresponding to the k-largest eigenvalues.

Output: $Y = U_k^T X$

Note: ① XX^T 特征值分解 $\Rightarrow$ X 奇异值分解（SVD）

$XX^T \in \mathbb{R}^{N\times N}$ $\quad$ $X \in \mathbb{R}^{N\times m}$

$XX^T = U\Sigma V^T V \Sigma^T U^T = U \Sigma^2 U^T$ $\Longleftarrow$ $X = U \Sigma V^T$

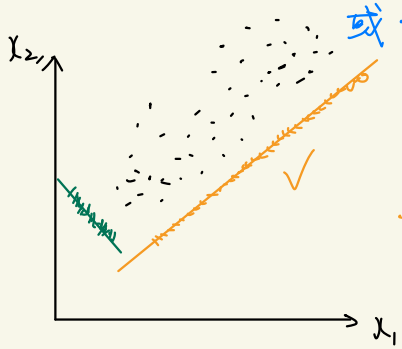
$\Sigma^2 \to \Lambda$ $\quad$ U: 左奇异向量

$U \in \mathbb{R}^{N\times N}$ 正交
$\Sigma \in \mathbb{R}^{N\times m}$ 对角
$V \in \mathbb{R}^{m\times m}$ 正交

对称方阵不区别 Eigenvalue Decomposition 和 SVD.

② 如何确定 k？ 通常用序号确定：

或设置重构阈值 $t=95\%$ → $\dfrac{\sum_{i=1}^{k}\lambda_i}{\underbrace{\sum_{i=1}^{N}\lambda_i}_{=\mathrm{tr}(XX^T)}} \geq t$ 成立的最小k值



e.g. $\mathbb{R}^2 \to \mathbb{R}$!

让样本尽可能分开

⇒ 最大化投影点的方差

KPCA (PCA + Kernel Method)

$$X = \begin{pmatrix} | & & | \\ \Phi(x_1) & \cdots & \Phi(x_m) \\ | & & | \end{pmatrix} \in \mathbb{R}^{N\times m}$$

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

↓

$\Phi \in \mathbb{H} \subseteq \mathbb{R}^N$ (RKHS) unknown.

non-linearly separable

↓

linearly separable

↓

PCA.

$$X = U\Sigma V^T$$

$$(X^TX)_{ij} = \Phi(x_i)^T\Phi(x_j)$$

⇒ covariance $C = \frac{1}{m}XX^T = \frac{1}{m}U\Lambda U^T$

kernel matrix $K = X^TX = V\Lambda V^T$ $\quad (\Lambda = \Sigma^2)$

$$XV\Sigma^{-1} = U\Sigma V^TV\Sigma^{-1} = U \Rightarrow U = XV\Lambda^{-1/2} \Rightarrow u_i = \frac{Xv_i}{\sqrt{\lambda}}$$

KPCA

$$Y = U^TX = (XV\Lambda^{-1/2})^TX$$

$$= \Lambda^{-1/2}V^TX^TX$$

$$= \Lambda^{-1/2}V^TK$$

→ 不需要特征映射 Φ，只需核函数 K 即可！

e.g. 1-dimensional: $P = UU^T \Rightarrow P_u = uu^T$

$$y = u^T x = x^T u = x^T \cdot \frac{Xv}{\sqrt{\lambda}} = \frac{k_x^T v}{\sqrt{\lambda}}. \qquad k_x = \begin{pmatrix} K(x_1, x) \\ \vdots \\ K(x_m, x) \end{pmatrix}$$

$$x^T \begin{pmatrix} | & & | \\ x_1 & \cdots & x_m \\ | & & | \end{pmatrix} = \left(K(x_1, x), \cdots, K(x_m, x) \right)^T$$

If $x = x_i$;

k_x is the ith column of K

$$\Rightarrow y = \frac{k_x^T v}{\sqrt{\lambda}} = \frac{\lambda v_i}{\sqrt{\lambda}} = \sqrt{\lambda}\, v_i$$

i^{th} compment of v.

(KPCA: K 的 top-k singular vectors)

PCA → 最近重构性 $\|P \cdot X - X\|_F^2 \leftarrow \min$

→ 最大可分性 $tr(U^T X X^T U) \leftarrow \max$

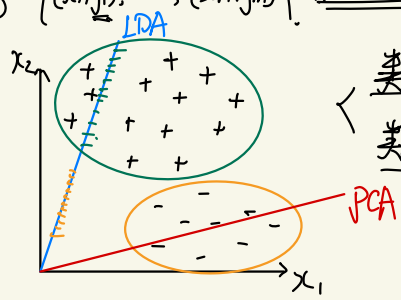
$S = \{x_1, \cdots, x_m\}$ 无标记，无监督

$S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$ 有标记，有监督

PCA : $\max\limits_{x \neq 0} \dfrac{x^T A x}{x^T x}$ → Rayleigh Quotient

LDA: $\max\limits_{x \neq 0} \dfrac{x^T S_b x}{x^T S_w x}$ Generalized Rayleigh Quotient

LDA (Linear Discriminant Analysis)



类内差异尽可能小

类间差异尽可能大

→ 考虑了标记信息.

正类 : $\mu_+ \to w^T \mu_+$ | $\Sigma_+ \to w^T \Sigma_+ w \to$ 小

负类 : $\mu_- \to w^T \mu_-$ | $\Sigma_- \to w^T \Sigma_- w \to$ 小

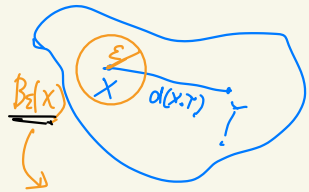
↓ 差异大

$$\Rightarrow \max \frac{\|w^T \mu_+ - w^T \mu_-\|_2^2}{w^T \Sigma_+ w + w^T \Sigma_- w} \quad \begin{matrix} \text{大} \\ \text{小} \end{matrix}$$

scatter matrix $\begin{cases} S_b = (\mu_+ - \mu_-)(\mu_+ - \mu_-)^T \\ S_w = \Sigma_+ + \Sigma_- . \end{cases} \Rightarrow$ LDA: $\max \dfrac{w^T S_b w}{w^T S_w w}$

流形学习 (Manifold Learning)

Metric Space (X, d)



open set

def (topology) X set.

$P(X)$ power set

$\mathcal{T} \subseteq P(X)$ collection of subsets

if $\mathcal{T}$ satisfies : ① $\phi, X \in \mathcal{T}$

② $A, B \in \mathcal{T} \Rightarrow A \cap B \in \mathcal{T}$ 有限交

③ $(A_i)_{i \in I}$ with $A_i \in \mathcal{T} \Rightarrow \bigcup_{i \in I} A_i \in \mathcal{T}$ 可列并

$\Rightarrow$ $\mathcal{T}$ is called a topology of X

$t \in \mathcal{T}$ is called an open set

$(X, \mathcal{T})$ is called topology space

流形

def (manifold) n-dimensional topological manifold

$X = \mathbb{R}$ $\mathcal{T} = \{\phi, \mathbb{R}\} \cup \{(a, +\infty) \mid a \in \mathbb{R}\}$

topological space $(X, \mathcal{T})$ with
- Hausdorff Space
- Second Countable
- Locally Euclidean of dimension n

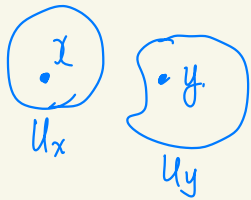
①

$a_n = \frac{1}{n}$, $n \in \mathbb{N}$

- converges to 0
- converges to $\forall x < 0$

x $\boxed{(x - \varepsilon, +\infty)}$

$\Rightarrow$ Hausdorff Space



$U_x \cap U_y = \phi$

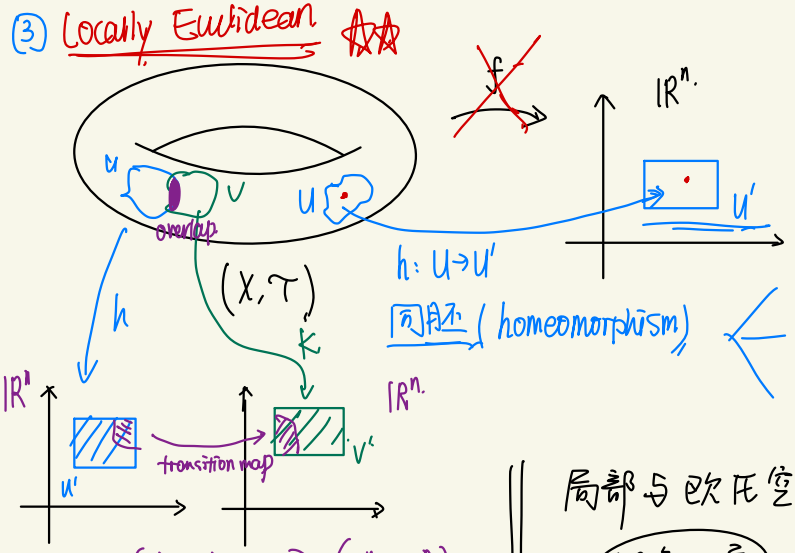
② $\mathcal{T}$: $\mathcal{B} \subseteq \mathcal{T}$ 基 $\quad \forall U \in \mathcal{T}.\ \exists (B_i)_{i\in I}.\ \bigcup_{i\in I} B_i = U$

$\mathcal{B}$ countable $\Rightarrow$ $\mathcal{T}$ second-countable

$\mathbb{R}^n$ with standard topology (Euclidean metric)

$\mathcal{B} = \{ B_\varepsilon(x) \mid x \in \mathbb{Q}^n,\ \varepsilon \in \mathbb{Q}^+ \}$ → countable.

③ Locally Euclidean ☆☆



$\forall x \in X.\ \exists$ open neighborhood $U \in \mathcal{T}$

homeomorphism $h: U \to U'$

(U, h) chart ↳ atlas $\quad U' \subseteq \mathbb{R}^n$ open

$h: U \to U'$

同胚 (homeomorphism)

- h: bijective
- h: continuous
- h^{-1}: continuous.

ω: $\mathbb{R}^n \to \mathbb{R}^n$ $(\mathbb{R}^n \to \mathbb{R}^n)$

differentiable ?

局部与欧氏空间同胚

$\Rightarrow$ 距离计算

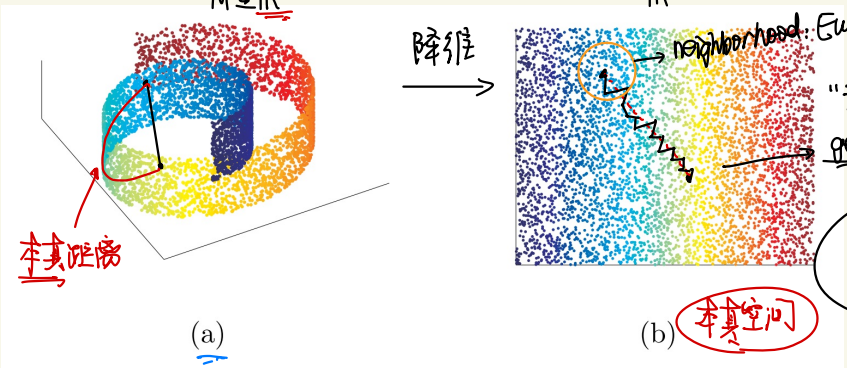
$S^2 \subseteq \mathbb{R}^3$

↓

$\mathbb{R}^3$

(2-dimensional manifold)

manifold



local structure

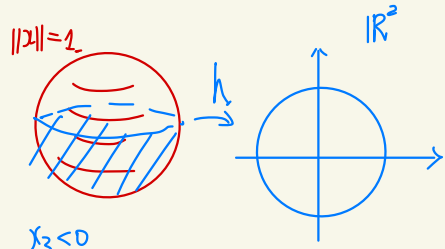
流形学习是一类借鉴了拓扑流形概念的降维方式

① Isomap (Isometric Map)

geodesic distance

↓

intrinsic distance



$x_3<0$

$$h: \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$h^{-1}: \begin{pmatrix} x_1' \\ x_2' \end{pmatrix} \mapsto \begin{pmatrix} x_1' \\ x_2' \\ -\sqrt{1-x_1'^2-x_2'^2} \end{pmatrix}$$

1. Find the t nearest neighbors for each data point based on L2-distance

undirected neighborhood graph (近邻图)

x_i < 近邻点距离: $\|x_i - x_n\|_2$
其他点距离: ∞

2. Shortest Path between all pairs of data points

→ geodesic distance matrix Δ Floyd-Warshall, Johnson

3. Find the k-dimensional representation:

低维空间中的距离

高维空间中的边的近邻距离

$$Y^* = \underset{Y}{\operatorname{argmin}} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\|y_i - y_j\|_2^2 - \Delta_{ij}^2 \right)$$

$$Y^* = \left(\Sigma_{Iso,k}\right)^{1/2} U_{Iso,k}^T \leftarrow K_{Iso} \text{ "核矩阵"}$$

top-k singular values

$$K_{Iso} = -\frac{1}{2} H \Delta^2 H \quad \left(H = I - \frac{1}{m} 1 1^T\right)$$

double centering

KPCA! K: PSD $\Rightarrow$ smooth manifold

关键是构造近邻图学距离!

↳ < t-NN (个数)
ε-NN (距离)

$G = (V, E)$

V: 样本点

E: 近邻距离

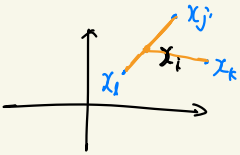
② Locally Linear Embedding（LLE 局部线性嵌入）

保持样本邻域间的线性关系

Step1. 学习线性关系

$$x_i = w_{ij} x_j + w_{ik} x_k + w_{il} x_l$$

1. LLE every $x_i \to Q_i$: t-neighbors of x_i



计算出基于 Q_i 中的样本点对 x_i 进行线性重构的系数 $\vec{w}_i$.

Step2 这样的关系可以在低维空间中得到保持

$$\boxed{\begin{aligned} &\min_{w_1,\cdots,w_m} \sum_{i=1}^{m} \Big\| x_i - \sum_{j\in Q_i} w_{ij} x_j \Big\|_2^2 \\ &\text{s.t.} \quad \sum_{j\in Q_i} w_{ij} = 1 \end{aligned}}$$

$x_i \in \mathbb{R}^N$

$Q_i = \{ q_i^1, \cdots, q_i^t \}$

$$\sum_{i=1}^{m} \Big\| x_i - \sum_{j\in Q_i} w_{ij} x_j \Big\|_2^2 = \sum_{i=1}^{m} \Big\| \underbrace{\sum_{j\in Q_j} w_{ij}}_{1} \cdot x_i - \sum_{j\in Q_i} w_{ij} x_j \Big\|_2^2$$

$$= \sum_{i=1}^{m} \Big\| \sum_{j\in Q_j} w_{ij} (x_i - x_j) \Big\|_2^2$$

$$= \sum_{i=1}^{m} \| X_i w_i \|_2^2$$

$$= \sum_{i=1}^{m} w_i^T X_i^T X_i w_i$$

$$w_i = \begin{pmatrix} w_{iq_i^1} \\ w_{iq_i^2} \\ \vdots \\ w_{iq_i^t} \end{pmatrix} \in \mathbb{R}^t \to \mathbb{R}^m$$

$$X_i = \begin{pmatrix} | & | & & | \\ x_i - x_{q_i^1} & x_i - x_{q_i^2} & \cdots & x_i - x_{q_i^t} \\ | & | & & | \end{pmatrix} \in \mathbb{R}^{N\times t}$$

$$\Rightarrow \boxed{\begin{aligned} &\min_{w_1,\cdots,w_m} \sum_{i=1}^{m} w_i^T X_i^T X_i w_i \\ &\text{s.t.} \quad w_i^T \vec{1} = 1 \qquad i \in [m] \end{aligned}}$$

$$\Rightarrow L(w_1,\cdots,w_m,\lambda) = \sum_{i=1}^{m} w_i^T X_i^T X_i w_i + \lambda (w_i^T \vec{1} - 1)$$

$$\frac{\partial L}{\partial w_i} = \frac{\partial}{\partial w_i} \left(w_i^T X_i^T X_i w_i + \lambda w_i^T \vec{1} \right) = 2 X_i^T X_i w_i + \lambda \vec{1} = 0$$

$$\Rightarrow X_i^T X_i w_i = -\frac{1}{2} \lambda \vec{1} \Rightarrow w_i = -\frac{1}{2} \lambda (X_i^T X_i)^{-1} \cdot 1$$

($X_i^T X_i$ 可逆时)

$$1 = w_i^T \vec{1} = \vec{1}^T w_i = -\frac{1}{2}\lambda \cdot \vec{1}^T (X_i^T X_i)^{-1} \vec{1}. \Rightarrow -\frac{1}{2}\lambda = \frac{1}{\vec{1}^T (X_i^T X_i)^{-1} \vec{1}}.$$

$$\Rightarrow w_i = \frac{(X_i^T X_i)^{-1} \vec{1}}{\vec{1}^T (X_i^T X_i)^{-1} \vec{1}}.$$

矩阵 $C^{(i)}$: $C_{jk} = (x_i - x_j)^T (x_i - x_k)$

则 C_{jk}^{-1} 是 $(X_i^T X_i)^{-1}$ 第 j 行第 k 列的元素

$$w_{ij} = w_{iq_i^j} = \frac{\sum_{k \in Q_i} C_{jk}^{-1}}{\sum_{l \in Q_i} \sum_{s \in Q_i} C_{ls}^{-1}} \leftarrow \text{analytical solution}$$

(2)

2. LLE 在低维空间中保持 w_i 不变, $x_i \in \mathbb{R}^N \to z_i \in \mathbb{R}^K$

$$\boxed{\min_{z_1, \cdots, z_M} \sum_{i=1}^{M} \left\| z_i - \sum_{j \in Q_i} w_{ij} z_j \right\|_2^2}$$

形式相同 $\begin{cases} \text{Step 1: } w \\ \text{Step 2: } z. \end{cases}$

let $Z = (z_1, \cdots, z_m) \in \mathbb{R}^{k \times m}$. $z_i = Z \cdot e_i$ $e_i \in \mathbb{R}^m$

$$\sum_{j \in Q_i} w_{ij} z_j = Z \cdot w_i$$

对 $w_i \in \mathbb{R}^m$ 而言 只有 x_i 的邻域内样本对应的下标 $w_{ij} \neq 0$ $(j \in Q_i)$ 且和为 1.

$$\Rightarrow z_i - \sum_{j \in Q_i} w_{ij} z_j = Z e_i - Z w_i = Z(e_i - w_i)$$

$$\Rightarrow \sum_{i=1}^{m} \left\| z_i - \sum_{j \in Q_i} w_{ij} z_j \right\|_2^2 \triangleq \|V\|_F^2 \qquad V = (v_1 \cdots v_m) \in \mathbb{R}^{k \times m}.$$

minimize

$$v_i = Z(e_i - w_i) \Rightarrow \underset{k \times m}{V} = \underset{k \times m}{Z} \underset{m \times m}{(I - W)}^T$$

$$W = \begin{pmatrix} w_1^T \\ \vdots \\ w_m^T \end{pmatrix} \qquad W^T = (w_1 \cdots w_m)$$

$$\|V\|_F^2 = tr(VV^T) = tr\left(Z(I-W)^T\right)\left(Z(I-W)^T\right)^T$$

$$= tr\left(Z(I-W)^T(I-W)Z^T\right)$$

$$\triangleq tr(Z M Z^T)$$

$M = (I-W)^T(I-W)$ $\Rightarrow$ $\boxed{\begin{array}{ll} \min\limits_{Z} & tr(ZMZ^T) \\ s.t. & ZZ^T = I \end{array}}$ $Z \in \mathbb{R}^{k \times m}$.

正交化

对M特征值分解 $\Rightarrow$ 最小的k个特征值对应的特征向量即为 Z^T.

LLE: 学习并维持局部线性关系

$x_i \simeq \sum_{j \in Q_i} w_{ij} x_j$

$z_i \simeq \sum_{j \in Q_i} w_{ij} z_j$

Input: $S = \{x_1, \cdots, x_m\}$ k.

1. for $i = 1, \cdots, m$, do

   find the t-NN of $x_i \to Q_i$

   w_i < $j \in Q_i$: w_{ij} by (1)

   $j \notin Q_i$: $w_{ij} = 0$ ← 将变动限制在局部 (manifold 的性质)

2. $M = (I-W)^T(I-W)$
3. M: eigenvalue decomposition
4. k eigenvectors corresponding to the smallest k eigenvalues of M

Output: Z.

③ Laplacian Eigenmaps 拉普拉斯特征映射

Similiar to LLE. $\Rightarrow$ preserve neighborhood relations by measure matrix W

local Euclidean.

1. Find t-NN for each point.
2. Construct W: sparse, symmetric $m \times m$ matrix

$$W_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|_2^2}{\sigma^2}\right), & (x_i, x_j) \text{ neighbors} \\ 0, & \text{otherwise.} \end{cases}$$

3. Find the k-dimensional representation

$$\boxed{Y^* = \underset{Y}{\operatorname{argmin}} \sum_{i=1}^{M} \sum_{j=1}^{M} W_{ij} \|y_i - y_j\|_2^2}$$

minimizing the weighted distance between neighbors

"W" to measure nearness!

x_i, x_j $\begin{cases} \text{近} \to W \text{接近 } 1 \\ \text{远} \to W \text{接近 } 0 \end{cases}$

非邻居 $\to W=0$

若W大，则 $\|y_i - y_j\|$ 需要小

$\Downarrow$

penalize nearby inputs for being mapped to faraway outputs

$Y^* = U_{L,k}^T$. $L = D - W$ graph Laplacian

diagnal $D_{ii} = \sum_{j=1}^{M} W_{ij}$ ← 第i个点到所有点的权重和

Neighborhood Graph 近邻图

连通(connected) $\Rightarrow$ L 有特征值 0.

对应特征向量 $\vec{1}$.

谱图论

Spectral Graph Theory

用线性代数的方式研究图的性质.

$\Downarrow$

$U_{L,k}^T$: bottom k singular vectors of L (excluding $\vec{1}$)

$\Longleftrightarrow$ Finding the largest singular value/vector of $L^\dagger$.

manifold learning

$\Updownarrow$

KPCA

$K_L = L^\dagger$: KPCA !

↳ commute times of graph random walk (Markov chain)

- stationary distribution
- mixing time
- hitting time
- commute time
- cover time

$K \to L \to W$: $W_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$

核函数什么时候可以成为一种距离度量？

metric

Gaussian kernel

$\exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$

$\Downarrow$

NDS kernels

def (NDS kernel) A kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a <u>negative-definite-symmetric</u> kernel if it is symmetric, and $\forall \{x_1, \cdots, x_m\} \subseteq \mathcal{X}$, $c \in \mathbb{R}^m$ with $\vec{1}^T c = 0$

$$c^T K c = \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j K(x_i, x_j) \leq 0$$

$\sum_{i=1}^{m} c_i = 0$

not any c !

(PDS: $\forall c \in \mathbb{R}^m$, $c^T K c \geq 0$)

K PDS $\Longrightarrow$ $-K$ NDS

e.g. (Squared-distance) $K: (x, x') \mapsto \|x - x'\|^2$ is an NDS kernel !

$$\sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \|x_i - x_j\|^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \left(\|x_i\|^2 + \|x_j\|^2 - 2 x_i \cdot x_j \right)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \left(\|x_i\|^2 + \|x_j\|^2 \right) - 2 \sum_{i=1}^{m} c_i x_i \sum_{j=1}^{m} c_j x_j$$

$$2 \left\| \sum_{i=1}^{m} c_i x_i \right\|^2 \geq 0$$

$$\leq \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \left(\|x_i\|^2 + \|x_j\|^2 \right)$$

$$= \underbrace{\left(\sum_{j=1}^{m} c_j \right)}_{0} \left(\sum_{i=1}^{m} c_i \|x_i\|^2 \right) + \underbrace{\left(\sum_{i=1}^{m} c_i \right)}_{0} \left(\sum_{j=1}^{m} c_j \|x_j\|^2 \right) = 0$$

Thm (NDS $\Leftrightarrow$ PDS) Let K' defined for any x_0:

$$K'(x, x') = K(x, x_0) + K(x', x_0) - K(x, x') - K(x_0, x_0), \quad \forall x, x' \in \mathcal{X}.$$

then, K is NDS $\Leftrightarrow$ K' is PDS

Proof: "$\Leftarrow$" Assume K' is PDS. for any c s.t. $\sum_{i=1}^{m} c_i = 1$

$$\sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j K(x_i, x_j) = \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j \left(K(x_i, x_0) + K(x_j, x_0) - K'(x_i, x_j) - K(x_0, x_0) \right)$$

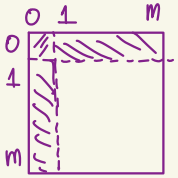
$$= \underbrace{\left(\sum_{j=1}^{m} c_j \right)}_{0} \left(\sum_{i=1}^{m} c_i K(x_i, x_0) \right) + \underbrace{\left(\sum_{i=1}^{m} c_i \right)}_{0} \left(\sum_{j=1}^{m} c_j K(x_j, x_0) \right) - \sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j K'(x_i, x_j) - \underbrace{\left(\sum_{i=1}^{m} c_i \right)^2}_{0} K(x_0, x_0)$$

$$= - \underbrace{\sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j K'(x_i, x_j)}_{K' \text{ PDS}} \leq 0 \Longrightarrow K \text{ NDS}$$

"$\Longrightarrow$" Assume K is NDS. for any $c \in \mathbb{R}^m$ $\rightarrow$ $\vec{1}^T c = 0$?

let $c_0 = -\vec{1}^T c$. $\Rightarrow \sum_{i=0}^{m} c_i = 0$ (增加一组变量(x_0, c_0)来凑出和为0)

$$\Rightarrow \sum_{i=0}^{m}\sum_{j=0}^{m} c_i c_j K(x_i, x_j) = \left(\sum_{i=0}^{m} c_i K(x_i, x_0)\right)\underbrace{\left(\sum_{j=0}^{m} c_j\right)}_{0} + \underbrace{\left(\sum_{i=0}^{m} c_i\right)}_{0}\left(\sum_{j=0}^{m} c_j K(x_j, x_0)\right)$$



$$-\underbrace{\left(\sum_{i=0}^{m} c_i\right)^2}_{0} K(x_0, x_0) - \sum_{i=0}^{m}\sum_{j=0}^{m} c_i c_j K(x_i, x_j) = -\sum_{i=0}^{m}\sum_{j=0}^{m} c_i c_j K'(x_i, x_j) \leq 0$$

(K NDS)

$$\Rightarrow \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j K'(x_i, x_j) + 2 \underbrace{c_0}_{-\sum_{i=1}^{m} c_i} \sum_{i=1}^{m} c_i \underbrace{K'(x_i, x_0)}_{0} + c_0^2 \underbrace{K'(x_0, x_0)}_{0} \geq 0$$

$\Rightarrow \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j K'(x_i, x_j) \geq 0$

$\forall x \in \mathcal{X}.\ K'(x, x_0) = 0$ $\triangle$

$\Rightarrow K'$ PDS. $\square$

Thm 1 Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric kernel. K is NDS $\Leftrightarrow$ $\exp(-tK)$ is PDS $t \geq 0$

$\Downarrow$

Thm 2 (NDS kernel as a metric)

e.g. $(x, x') \mapsto \|x - x'\|^2$ is NDS

$\Rightarrow (x, x') \mapsto \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$ is PDS !

Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be an NDS kernel such that $\forall x, x' \in \mathcal{X}$. $K(x, x') = 0$ iff $x = x'$.

then, there exist a Hilbert Space $\mathbb{H}$ and a mapping $\Phi: \mathcal{X} \to \mathbb{H}$ such that

$$\forall x, x' \in \mathcal{X}. \quad K(x, x') = \|\Phi(x) - \Phi(x')\|^2. \quad ☆☆$$

$\Longrightarrow$ $\sqrt{K} = \|\Phi(x) - \Phi(x')\|$ provides a metric in $\mathbb{H}$.

e.g. $(x, x') \mapsto \exp(-|x - x'|^p)$ is NOT PDS for $p > 2$

否则. $\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j e^{-t|x_i - x_j|^p} = \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j e^{-|t^{\frac{1}{p}} x_i - t^{\frac{1}{p}} x_j|^p} \geq 0$

no metric here!

$\Rightarrow (x, x') \mapsto |x - x'|^p$ is NDS for $p > 2$? X

Theoretical Guarantee for Dimensionality Reduction

↳ Johnson-Lindenstrauss lemma.

任意高维空间中的m个样本点都可以被映射到某个低维空间$\mathbb{R}^k$.

with $k \geq O\left(\frac{\log m}{\varepsilon^2}\right)$ 且任意两个点之间的距离改变不超过$(1\pm\varepsilon)$.

by Randomized polynomial time algorithm.
(randomly choose k-dimensional linear subspace)

Lemma 1 Random variable $Q \sim \chi^2(k)$

$\forall\, 0<\varepsilon<\frac{1}{2}$,

$$\Pr\left((1-\varepsilon)k \leq Q \leq (1+\varepsilon)k\right) \geq 1-2e^{-(\varepsilon^2-\varepsilon^3)k/4}$$

Proof: By Markov's inequality, $\Pr(Q \geq (1+\varepsilon)k) = \Pr\left(e^{\lambda Q} \geq e^{\lambda(1+\varepsilon)k}\right) \leq \frac{\mathbb{E}[e^{\lambda Q}]}{e^{\lambda(1+\varepsilon)k}}$

MGF

MGF for χ^2 distribution: $\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t\sum_{i=1}^{k} X_i^2}]$

$$= \mathbb{E}\left[\prod_{i=1}^{k} e^{tX_i^2}\right] = \prod_{i=1}^{k} \mathbb{E}[e^{tX_i^2}] = \mathbb{E}[e^{tX_1^2}]^k.$$

$X_i \sim N(0,1)$ $\Rightarrow \mathbb{E}[e^{tX_i^2}] = \int_{-\infty}^{+\infty} e^{tx^2} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{(1-2t)\frac{-x^2}{2}} dx.$

$$\Rightarrow \Pr(Q \geq (1+\varepsilon)k) \leq \boxed{\frac{(1-2\lambda)^{-k/2}}{\exp(\lambda(1+\varepsilon)k)}} = \frac{1}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{+\infty} \frac{e^{-\frac{u^2}{2}}}{\sqrt{1-2t}} du}_{\sqrt{2\pi}\ (\text{Gaussian Integral})} = \frac{1}{\sqrt{1-2t}}$$

$u = \sqrt{1-2t}\, x.$ $(t<\frac{1}{2})$

log: $\left(-\frac{k}{2}\log(1-2\lambda) - \lambda(1+\varepsilon)k\right)' = \frac{k}{1-2\lambda} - (1+\varepsilon)k = 0 \Rightarrow \lambda = \frac{\varepsilon}{2(1+\varepsilon)} < \frac{1}{2}$

$$\Pr(Q \geq (1+\varepsilon)k) \leq \left(\frac{1+\varepsilon}{e^{\varepsilon}}\right)^{k/2} \leq \left(\frac{e^{\varepsilon-\frac{\varepsilon^2-\varepsilon^3}{2}}}{e^{\varepsilon}}\right)^{k/2} = \exp\left(-\frac{k}{4}(\varepsilon^2-\varepsilon^3)\right)$$

Similiarly, $\Pr(Q \leq (1-\varepsilon)k) \leq \exp\left(-\frac{k}{4}(\varepsilon^2-\varepsilon^3)\right)$

By Union Bound, $\Pr\left((1-\varepsilon)k \leq Q \leq (1+\varepsilon)k\right) \geq 1-2\exp\left(-\frac{k}{4}(\varepsilon^2-\varepsilon^3)\right)$

Lemma 2 $x \in \mathbb{R}^N$. $A \in \mathbb{R}^{k\times N}$. $A_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $i \in [k]$. $j \in [N]$.

$$\forall\, 0<\varepsilon<\tfrac{1}{2}.\quad \Pr\left((1-\varepsilon)\|x\|^2 \le \left\|\tfrac{1}{\sqrt{k}}Ax\right\|^2 \le (1+\varepsilon)\|x\|^2\right) \ge 1-2\exp\left(-\tfrac{k}{4}(\varepsilon^2-\varepsilon^3)\right)$$

Proof: $\hat{x} = Ax \in \mathbb{R}^k$.

$\mathbb{E}=0$ $\quad$ $\mathbb{E}^2=1$

$$\mathbb{E}[\hat{x}_j^2] = \mathbb{E}\left[\left(\sum_{i=1}^N A_{ji}x_i\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^N A_{ji}^2 x_i^2\right] = \sum_{i=1}^N x_i^2 = \|x\|^2$$

随机性 来源于 A.

$$T_j = \frac{\hat{x}_j}{\|x\|} \quad (\text{normalization}) \qquad T_j = \frac{\sum_{i=1}^N A_{ji}x_i}{\|x\|} \sim \mathcal{N}(0,1)$$

$N\uparrow$ $\quad$ $\in \mathbb{R}^N$

$$Q = \sum_{j=1}^k T_j^2 \sim \chi^2(k)$$

$$\Rightarrow \Pr\left((1-\varepsilon)\|x\|^2 \le \frac{\|\hat{x}\|^2}{k} \le (1+\varepsilon)\|x\|^2\right) = \Pr\left((1-\varepsilon)k \le \sum_{j=1}^k T_j^2 \le (1+\varepsilon)k\right)$$

$$= \Pr\left((1-\varepsilon)k \le Q \le (1+\varepsilon)k\right) \ge 1-2\exp\left(-\tfrac{k}{4}(\varepsilon^2-\varepsilon^3)\right) \quad \text{by Lemma 1.} \quad \square$$

Thm (Johnson-Lindenstrauss Lemma) $\forall\, 0<\varepsilon<\frac{1}{2}$. $\forall\, m>4$. let $k = \frac{20\log m}{\varepsilon^2} = O\left(\frac{\log m}{\varepsilon^2}\right)$

与N无关!

for any set V of m points in $\mathbb{R}^N$. there exists a map $f: \mathbb{R}^N \to \mathbb{R}^k$

such that $\forall\, u,v \in V$: $\boxed{(1-\varepsilon)\|u-v\|^2 \le \|f(u)-f(v)\|^2 \le (1+\varepsilon)\cdot\|u-v\|^2}$.

低维空间中的距离 $\quad$ 高维空间中的距离

Proof: let $f = \frac{1}{\sqrt{k}}A$. $A \in \mathbb{R}^{k\times N}$. with $A_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$

randomly sample!

$(1\pm\varepsilon)$

with fixed $u,v \in V$: $x = u-v$. by Lemma 2.

$$\Pr\left((1-\varepsilon)\|u-v\|^2 \le \|f(u)-f(v)\|^2 \le (1+\varepsilon)\|u-v\|^2\right)$$

$$\ge 1-2m^2\exp\left(-\tfrac{k}{4}(\varepsilon^2-\varepsilon^3)\right) \overset{k=\frac{20\log m}{\varepsilon^2}}{=} 1-2m^2\exp\left(-5\log m(1-\varepsilon)\right) = 1-2m^2\cdot m^{5\varepsilon-5}$$

for all $O(m^2)$ pairs of points (u,v)

$$= 1-2m^{5\varepsilon-3} > 1-2m^{-1/2} > 0$$

$\varepsilon<\frac{1}{2}$. $\quad$ $m>4$

$\Rightarrow$ there exist such map $f: \mathbb{R}^N \to \mathbb{R}^k$! $\quad \square$