5 Support Vector Machine 支持向量机

优化基础 { Lagrangian, KKT condition, Duality

"Convex Optimization"

Steven Boyd

Optimization: $\min f_0(x)$ → objective.

subject to $f_i(x) \le 0$, $g_i(x) = 0$ } constraints → feasible field. (feasible solution)

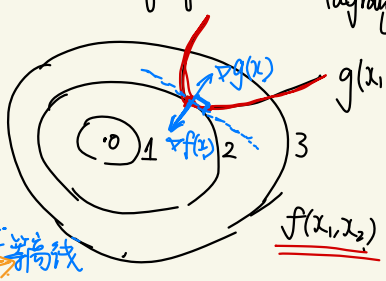
x^*: optimal solution

Lagrangian

① $\min f(x)$ s.t. $g(x)=0$ ⇒ $\underbrace{L(x,\lambda)}_{\text{Lagrangian}} = f(x) + \underbrace{\lambda}_{\text{lagrange multiplier}} g(x)$ ⇒ $\begin{cases} \nabla_x L(x,\lambda) = 0 \Leftarrow \frac{\partial L}{\partial x_i} = 0 \\ g(x) = 0 \end{cases}$

min $f(x_1, x_2)$ → 曲面

s.t. $g(x_1, x_2) = 0$ → 曲线



"tangle"

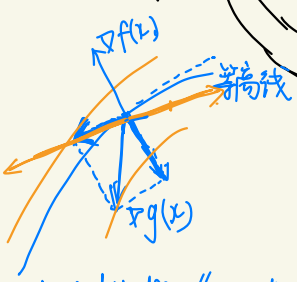
find the lowest point on $g(x_1,x_2)=0$

⇓

$\nabla f(x^*) \parallel \nabla g(x^*)$

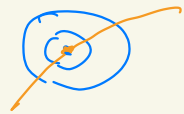
⇒ $\exists \lambda \in \mathbb{R}$, s.t. $\nabla f + \lambda \cdot \nabla g = 0$

(if $\lambda = 0$, $\nabla f(x^*) = 0$)



normal vector ∥ gradient

⇒ not optimal!



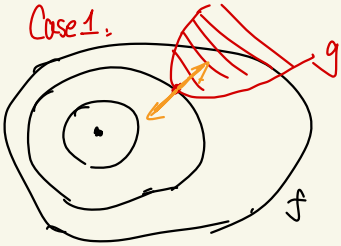
② $\min f(x)$

s.t. $g(x) \le 0$

Case1: Optimal Point x^* s.t. $g(x^*) = 0$

Case2: ～ s.t. $g(x^*) < 0$

Case 1:



Case 2:

gradients $\nabla g(x)$

$$\boxed{\nabla f(x^*) + \lambda \cdot \nabla g(x^*) = 0}$$

$\lambda \geq 0$ (f和g的梯度方向同反)

do not conside $g(x)$ $\quad \lambda = 0$

$\nabla f(x^*) = 0$

$$\Rightarrow \nabla f + \lambda \cdot \nabla g = 0 \quad (\lambda \geq 0)$$

$\lambda \cdot g(x^*) = 0$ for optimal point x^* $\Rightarrow$ Complement Relaxation 互补松弛

$\Downarrow$

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) = 0 \quad i = 1, 2, \cdots, m \\ & h_j(x) \leq 0 \quad j = 1, 2, \cdots, n. \end{aligned}$$

$\Rightarrow$ Lagrangian:

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{n} \mu_j h_j(x)$$

$\lambda = (\lambda_1, \cdots, \lambda_m)^T$. $\quad \mu = (\mu_1, \cdots, \mu_n)^T$.

if x^* is optimal point, then

$$\begin{cases} \nabla_x L(x, \lambda, \mu) = \nabla f + \sum_{i=1}^{m} \lambda_i \cdot \nabla g + \sum_{j=1}^{n} \mu_j \cdot \nabla h = 0 \\ g_i(x^*) = 0 & i \in [m] \\ h_j(x^*) \leq 0 & j \in [n] \\ \lambda_i \in \mathbb{R} & i \in [m] \\ \mu_j \geq 0 & j \in [n] \\ \mu_j h_j(x^*) = 0 & j \in [n]. \end{cases}$$

☆☆ $\boxed{\text{KKT condition}}$ $\Leftarrow$

(Karush-Kuhn-Tucker)

$\nabla L = 0$

constraints

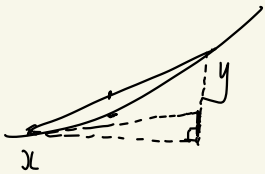
relaxation.

Notice: KKT condition is only necessary, not sufficient (maybe saddle point / local minimum)

But for convex optimization problem, KKT is global optimum!

Convex Function: $f(\lambda x+(1-\lambda)y) \le \lambda f(x)+(1-\lambda)f(y) \quad \forall \lambda\in[0,1]$.

with $\mathrm{dom}(f)$ is convex set



① 1-st: order → $f(y) \ge f(x)+\nabla f(x)^T(y-x)$

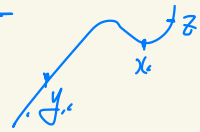
② 2-nd: order → $\nabla^2 f(x) \succeq 0$.

Proposition: local optimum is global optimum.

Proof: x local optimum point i.e. $\forall z\in(x-\delta, x+\delta)$ s.t. $f(z)\ge f(x)$.

Suppose x not global optimum point. i.e. $\exists\, y$ s.t. $f(y)<f(x)$

Let $z=(1-\theta)\cdot x+\theta\cdot y$ ⇒ $\|z-x\|_2=\theta\cdot\|y-x\|_2$, $\theta\in[0,1]$.



By Convexity, $f(z)\le(1-\theta)f(x)+\theta f(y) < f(x)$.

let $\theta=\dfrac{\delta}{2\|y-x\|_2}$ ⇒ $\|z-x\|_2=\dfrac{\delta}{2}<\delta$ ⇒ $f(z)\ge f(x)$ } ⇒ contradictory. □

Duality 对偶性

双双十一的物书 { 固定价格，如何买到尽可能多的高品？ / 固定商品数量，如何花尽可能少的钱？

☆☆

primal 原问题 $\min_x f(x)$ s.t. $g_i(x)=0 \quad i\in[m]$, $h_j(x)\le 0 \quad j\in[n]$ $\quad = p^*$

⇓

dual 对偶问题 $\max_{\lambda,\mu} d(\lambda,\mu)$ s.t. $\mu\succeq 0$ $\quad = d^*$

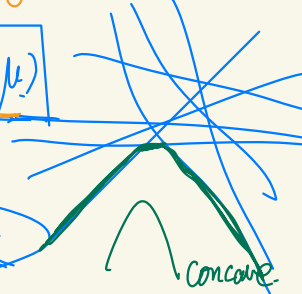
$$L(x,\lambda,\mu)=f(x)+\underbrace{\sum_{i=1}^{m}\lambda_i g_i(x)}_{=0}+\underbrace{\sum_{j=1}^{n}\mu_j h_j(x)}_{\le 0}$$

⇓

$$\boxed{d(\lambda,\mu)=\inf_x L(x,\lambda,\mu)}$$

① concave (凹)

② $\mu\succeq 0$ ⇒ $d(\lambda,\mu)\le p^*$.



① Weak Duality : $d^* \le p^*$ ✓

② Strong Duality : $d^* = p^*$ ✗. "Not always holds".

强对偶性

But, with certain condition. < Slater's condition / KKT condition

primal ⟺ dual

Why duality? < primal : $x \in \mathbb{R}^d$ → 维度. / dual : $\lambda_i \in \mathbb{R}^m$, $\mu \in \mathbb{R}^n$ → 约束数量

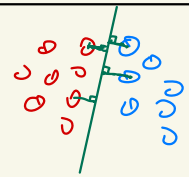
if $d \gg m+n$: 求解对偶问题更有效! (参数量更少)

Linear Classification,

$\mathcal{X}: \mathbb{R}^N$, $N \ge 1$.

$\mathcal{Y}: \{-1, +1\}$

$\mathcal{H} = \{ x \mapsto \mathrm{sgn}(w \cdot x + b) : w \in \mathbb{R}^N, b \in \mathbb{R} \}$



hyperplane

$wx + b = 0$

$\boxed{\text{Separable Case}}$

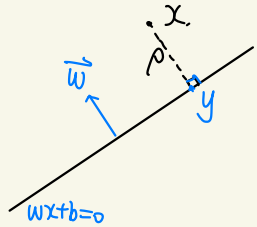
(w, b) s.t. $\forall i \in [m]$ $y_i(w \cdot x_i + b) \ge 0$

Margin

def (Geometric Margin) $\rho_h(x) = \dfrac{|w \cdot x + b|}{\|w\|_2}$

点到超平面的距离

可间隔



Proof: $\vec{w} \parallel \overrightarrow{xy}$

$\Rightarrow |\overrightarrow{xy} \cdot \vec{w}| = \|\overrightarrow{xy}\|_2 \cdot \|\vec{w}\|_2$ $\left(= \|\overrightarrow{xy}\|_p \cdot \|\vec{w}\|_q, \ p, q \text{ conjugate } \frac{1}{p} + \frac{1}{q} = 1 \right)$

$\rho_h(x) = \dfrac{|\overrightarrow{xy} \cdot \vec{w}|}{\|\vec{w}\|_2}$

$\sum_{i=1}^{N} y_i w_i + b = 0$

$|\overrightarrow{xy} \cdot \vec{w}| = |(x_1 - y_1) \cdot w_1 + (x_2 - y_2) \cdot w_2 + \cdots + (x_N - y_N) \cdot w_N|$

$= \left| \sum_{i=1}^{N} x_i w_i - \sum_{i=1}^{N} y_i w_i \right| = \left| \sum_{i=1}^{N} x_i w_i + b \right| = |w \cdot x + b|$ □

Intuition: Maximum-Margin-Hyperplane

$$\max_{w,b}\ \min_{i\in[m]} \rho_h(x_i) = \frac{|w^T x_i + b|}{\|w\|_2} \Rightarrow \max_{w,b}\ \min_{i\in[m]} \frac{y_i(w^T x_i + b)}{\|w\|_2} \rightarrow \min_{i\in[m]} y_i(w^T x_i + b) = 1.$$

不关心 w,b 长度

只关心 w,b 方向 $\frac{w}{n}, \frac{b}{n}$

"缩放"

$$\Rightarrow \max_{w,b} \frac{1}{\|w\|_2}$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1, \quad i\in[m].$$

$$\Rightarrow \boxed{\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|_2^2 \\ &\text{s.t. } y_i(w^T x_i + b) \geq 1, \quad i\in[m] \end{aligned}}$$

primal ☆

参数量 N

SVM: 最大化最小的间隔

QP (quadratic programming) e.g. CVX

Duality 对偶

Lagrangian $L(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i(w^T x_i + b)\right)$

$$g(\alpha) = \inf_{w,b} L(w,b,\alpha)$$

$$\begin{cases} \frac{\partial L}{\partial w} = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \\ \frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

$$\frac{\partial x^T a}{\partial x} = \frac{\partial a^T x}{\partial x} = a$$

$$\frac{\partial x^T x}{\partial x} = 2x.$$

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x$$

$$\Rightarrow \begin{cases} w = \sum_{i=1}^{m} \alpha_i y_i x_i \\ \sum_{i=1}^{m} \alpha_i y_i = 0 \end{cases}$$

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

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\boxed{\begin{aligned} &\text{Dual} \quad \max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &☆ \quad \text{s.t. } \sum_{i=1}^{m} \alpha_i y_i = 0 \\ &\qquad \alpha_i \geq 0 \quad i\in[m] \end{aligned}}$$

参数量 m

QP

$\Rightarrow (\alpha_1, \cdots, \alpha_m)$

Complement Relaxation

KKT condition
$$\begin{cases} \alpha_i \geq 0 \\ y_i f(x_i) \geq 1 \\ \alpha_i(y_i f(x_i) - 1) = 0 \end{cases}$$

$$w=\sum_{i=1}^{m}\alpha_i y_i x_i=\sum_{i\in SV}\alpha_i y_i x_i$$

$\alpha_i=0$ 或 $y_i f(x_i)=1$

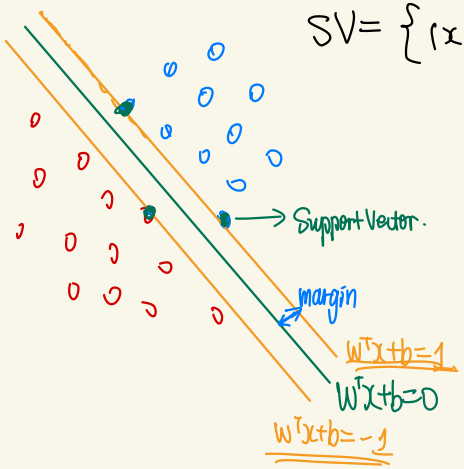
$\Downarrow$

$f(x_i)=\pm 1$

$(w^T x_i+b=\pm 1)$

SV: support vector "支持向量"

$$SV=\{(x_i,y_i):\ y_i f(x_i)=1,\ i\in[m]\}$$



b? 对于所有SV, $y_i f(x_i)=1$,

$$y_i\Big(\underbrace{\sum_{j\in SV}\alpha_j y_j x_j^T}_{w^T} x_i+b\Big)=1$$

$$\Rightarrow\ b_i=\frac{1}{y_i}-\sum_{j\in SV}\alpha_j y_j x_j^T x_i$$

robustness: $b=\frac{1}{|SV|}\sum_{i\in SV}\left(\frac{1}{y_i}-\sum_{j\in SV}\alpha_j y_j x_j^T x_i\right)$

$\Rightarrow$ 超平面(by SVM):

$w^Tx+b=0$: $\boxed{\left(\sum_{i\in SV}\alpha_i y_i x_i^T\right)x+\frac{1}{|SV|}\sum_{i\in SV}\left(\frac{1}{y_i}-\sum_{j\in SV}\alpha_j y_j x_j^T x_i\right)=0}$ $\quad w=\sum_{i\in SV}\alpha_i y_i x_i$

SVM 解的稀疏性: 训练完成后, 最终模型只与支持向量有关. ☆

*怎么解? SMO (Sequential Minimal Optimization)

Step1: 每次选出一对需要更新的样本 i,j $\Rightarrow$ α_i,α_j

Step2: 固定除 α_i,α_j 外的参数, 求解 Dual 更新 α_i,α_j

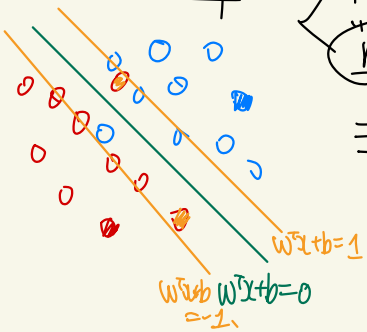
$$0=\sum_{i=1}^{m}\alpha_i y_i\ \Rightarrow\ \alpha_i y_i+\alpha_j y_j=C.\ \alpha_i\geq 0,\ \alpha_j\geq 0\ \Rightarrow\ \alpha_i=\frac{C-\alpha_j y_j}{y_i}$$

$$\max\ \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j\ \Rightarrow\ \max_{\alpha_j} f(\alpha_j)$$ 单变量问题

"block coordinate descent".

$\Rightarrow$ Analytical Solution (闭式解)

Non-separable Case



违反 violation (outlier)

projection to higher dimensional space. (升维)

"relaxation" (松弛)

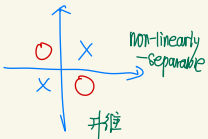
$\exists x_i \in S$ s.t. $y_i(w^Tx_i+b) < 1$

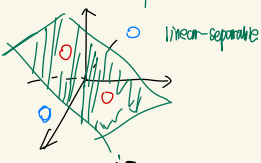
$\Downarrow$

$y_i(w^Tx_i+b) \geq 1-\xi_i$

$(\exists \xi_i > 0)$

slack variable 松弛变量





$x \to \Phi(x)$

kernel function.

核方法, 核技巧

Chapter 6

primal:

$$\boxed{\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}\|w\|_2^2 + C\cdot\sum_{i=1}^{m}\xi_i^p \\ &\text{s.t. } y_i(w^Tx_i+b) \geq 1-\xi_i, \quad i\in[m] \\ &\xi_i \geq 0 \end{aligned}}$$

penalty 惩罚 ξ_i^p (样本违反约束的程度)

① $1-\xi_i$: soft margin (软间隔支持向量机)

② ξ_i^p: penalty term. $\Rightarrow$ regularization term

$\Downarrow$

loss function: $\boxed{l(x_i) = \max(0, 1-y_i(wx_i+b))^p}$

$$\frac{1}{2}\|w\|_2^2 + C\cdot\underbrace{\boxed{\sum_{i=1}^{m}\xi_i^p}}_{\text{Empirical Risk}}$$

SRM (结构风险最小化)

if $p=1$

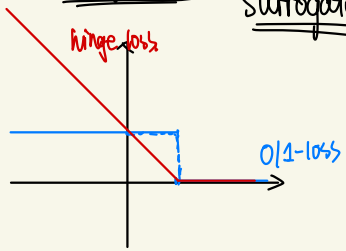
then $l(x_i) = \max(0, 1-y_i\underbrace{(w^Tx_i+b)}_{f(x_i)})$

$$= \begin{cases} 0, & y_i(w^Tx_i+b) \geq 1 \to \text{无事发生} \\ \underbrace{(1-y_i(wx_i+b))^p}_{\xi_i^p}, & y_i(w^Tx_i+b) < 1, \to \begin{cases} (0,1): \text{less confident} \\ <0: \text{mistake} \end{cases} \end{cases}$$

惩罚!

"hinge loss"

surrogate



$\Downarrow$ $p=1$

$$L(w,b,\xi,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\cdot\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left(1-\xi_i-y_i(w^Tx_i+b)\right) - \sum_{i=1}^{m}\beta_i\xi_i$$

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \Rightarrow w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \\ \dfrac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \\ \dfrac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \beta_i = 0 \quad i \in [m] \end{cases}$$

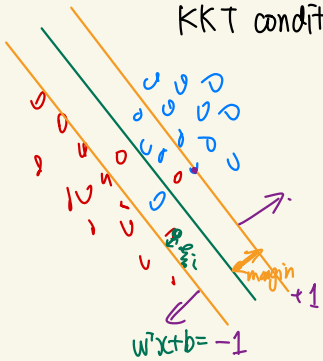
$$\Rightarrow g(\alpha, \beta) = \inf L(w, b, \xi, \alpha, \beta)$$
$$= \underbrace{\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j}_{\text{objective.}}$$

dual problem:

$$\boxed{\begin{aligned} &\max_{\alpha, \beta} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &\text{s.t.} \sum_{i=1}^{m} \alpha_i y_i = 0, \\ &0 \le \alpha_i \le C \quad \text{(Soft Margin)} \end{aligned}}$$

只有内积 (kernel function)

KKT condition: $\begin{cases} \alpha_i \cdot (1 - \xi_i - y_i(w^T x_i + b)) = 0 \\ \beta_i \xi_i = 0 \end{cases} \Rightarrow \begin{cases} \alpha_i = 0 : \xi_i = 0 \\ 0 < \alpha_i < C : \xi_i < 0 \\ \alpha_i = C : \xi_i > 0 \end{cases}$

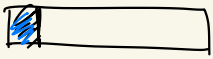


$$\left.\begin{aligned} y_i = 1 &. \ w^T x_i + b \le 1. \\ y_i = -1 &. \ w^T x_i + b \ge -1 \end{aligned}\right\} \text{Support Vector (SV)}$$

松弛

---

maximum-margin-hyperplane

CV: 

① Leave-One-Out (LOO) analysis

def (LOO error) $\hat{R}_{LOO}(A) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{h_{S-\{x_i\}}(x_i) \neq y_i}$

train validate

Lemma $m \ge 2$. $\mathbb{E}_{S \sim \mathcal{D}^m}[\hat{R}_{LOO}(A)] = \mathbb{E}_{S' \sim \mathcal{D}^{m-1}}[R(h_{S'})]$, (unbiased estimator)

Proof: $\mathbb{E}_{S \sim \mathcal{D}^m}[\hat{R}_{LOO}(A)] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[\mathbb{1}_{h_{S-\{x_i\}}(x_i) \neq y_i}]$

i.i.d

$$= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathbb{1}_{h_{S-\{x_1\}}(x_1)\neq y_1}\right]$$

$$= \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}\left[\underline{\mathbb{E}_{x_1\sim\mathcal{D}}\left[\mathbb{1}_{h_{S'}(x_1)\neq y_1}\right]}\right] \overset{\text{decomposition}}{=} \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}\left[R(h_{S'})\right]. \quad \square$$

Sparsity Thm h_S: SVM. $\underline{\mathbb{E}_{S\sim\mathcal{D}^m}\left[R(h_S)\right] \le \mathbb{E}_{S\sim\mathcal{D}^{m+1}}\left[\frac{|SV|}{m+1}\right]}$.

#Support Vectors

average fraction of SV.

Proof: S linearly separable sample set of $m+1$ samples.

$\underline{x \notin SV}$: removing it does not change the SVM solution $\Rightarrow \underline{h_{S-\{x\}} = h_S}$

Contraposition: 若 $h_{S-\{x\}}$ 错误分类 x, 则 $x \in SV$ 且 $h_{S-\{x\}}$ 仍能正确分类 x.

$$\Rightarrow \underline{\hat{R}_{LOO}(SVM)} = \frac{1}{m+1}\sum_{i=1}^{m+1} \underline{\mathbb{1}_{h_{S-\{x_i\}}(x_i)\neq y_i}} \le \frac{|SV|}{m+1}.$$

by Lemma, $\mathbb{E}_{S\sim\mathcal{D}^m}\left[R(h_S)\right] \le \mathbb{E}_{S\sim\mathcal{D}^{m+1}}\left[\frac{|SV|}{m+1}\right]$. $\square$

weak (average) $\mathbb{E}$ ✓ $\underline{Var}$ ✗ $\underline{\text{large-deviation}}$

② VC-dimension generalization bound

$$R(h) \le \hat{R}_S(h_S) + \sqrt{\frac{2(N+1)\cdot\log\frac{em}{N+1}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

$VCdim(\mathcal{H}) = N+1$.

$\mathbb{R}^N$ 维超平面

$\underline{N \text{ large}} \rightarrow$ uninformative

有没有 bound 能够 $\boxed{\text{independent to } N?}$

$\boxed{\text{Margin Theory}}$

间隔理论

③ Margin Theory

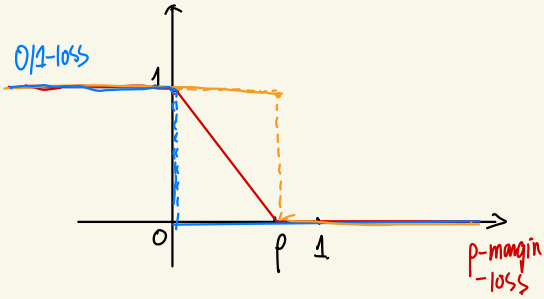
def (confidence margin) h (x,y). $\Rightarrow \underline{yh(x)}$

if $yh(x) > 0 \Rightarrow$ correctly classified.

$|h(x)|$: confidence of the prediction

$$\rho_{geom}(x) = \frac{|w\cdot x+b|}{\|w\|_2} = \frac{|y\,h(x)|}{\|w\|_2} \Rightarrow |y\,h(x)| \geq \rho_{geom}(x)\cdot\|w\|_2$$

def (Margin loss function) $\Phi_\rho(x) = \min\left(1, \max\left(0, 1-\frac{x}{\rho}\right)\right) = \begin{cases} 1, & x\leq 0 \\ 1-\frac{x}{\rho}, & 0\leq x\leq\rho \\ 0, & x\geq\rho \end{cases}$



$y_i h(x_i) \geq \rho$: correct with strong confidence.

$0\leq y_i h(x_i)\leq\rho$: correct with less confidence.

$y_i h(x_i)\leq 0$: incorrect.

$$\boxed{\mathbb{1}_{y_i h(x_i)\leq 0} \leq \Phi_\rho(y_i h(x_i)) \leq \mathbb{1}_{y_i h(x_i)\leq\rho}}$$

def (Empirical Margin loss) $S=\{x_1,\cdots,x_m\}$. $\hat{R}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m}\Phi_\rho(y_i h(x_i))$

confidence margin.

Φ_ρ: $\frac{1}{\rho}$-Lipschitz

(l-Lipschitz function: $\forall x,y\in dom(f)$. $|f(x)-f(y)|\leq l\,|x-y|$)

Lemma (Talagrand's lemma) $\Phi_1,\cdots,\Phi_m$ l-Lipschitz. ; $\sigma_1,\cdots,\sigma_m$ Rademacher Variable

$$\frac{1}{m}\mathbb{E}_\sigma\left[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i(\Phi_i\circ h)(x_i)\right] \leq \frac{l}{m}\cdot\mathbb{E}_\sigma\left[\sup\sum_{i=1}^{m}\sigma_i h(x_i)\right] = l\cdot\hat{R}_S(H)$$

In particular $\forall i\in[m]$. $\Phi_i=\Phi$. $\Rightarrow$ $\hat{R}_S(\Phi\circ H) \leq l\cdot\hat{R}_S(H)$

$\{\Phi\circ h : h\in H\}$ $\quad$ $\circ$: composition

Proof. $S=\{x_1,\cdots,x_m\}$

$$\frac{1}{m}\mathbb{E}_\sigma\left[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i(\Phi_i\circ h)(x_i)\right] = \frac{1}{m}\mathbb{E}_{\sigma_1,\cdots,\sigma_{m-1}}\left[\mathbb{E}_{\sigma_m}\left[\sup_{h\in H} U_{m-1}(h)+\sigma_m(\Phi_m\circ h)(x_m)\right]\right]$$

$U_{m-1}(h) = \sum_{i=1}^{m-1}\sigma_i(\Phi_i\circ h)(x_i)$.

$\sigma_m=1$: $U_{m-1}(h_1)+(\Phi_m\circ h_1)(x_m) \geq (1-\varepsilon)\left(\sup_{h\in H} U_{m-1}(h)+(\Phi_m\circ h)(x_m)\right)$

$\sigma_m=-1$. $U_{m-1}(h_2)-(\Phi_m\circ h_2)(x_m) \geq (1-\varepsilon)\left(\sup_{h\in H} U_{m-1}(h)-(\Phi_m\circ h)(x_m)\right)$ $\quad$ $\forall\varepsilon>0$ $\exists h_1,h_2\in H$

$$(1-\varepsilon)\cdot \mathbb{E}_{\sigma_m}\left[\sup_{h\in H} U_{m-1}(h) + \sigma_m(\Phi_m\circ h)(x_m)\right]$$

sub-additivity

$$\le \frac{1}{2}\cdot\left(U_{m-1}(h_1) + (\Phi_m\circ h_1)(x_m)\right) + \frac{1}{2}\left(U_{m-1}(h_2) - (\Phi_m\circ h_2)(x_m)\right)$$

since $(\Phi_m\circ h_1)(x_m) - (\Phi_m\circ h_2)(x_m) = \Phi_m(h_1(x_m)) - \Phi_m(h_2(x_m)) \le l\cdot|h_1(x_m)-h_2(x_m)|$ Lipschitz.

$= s\cdot l\cdot(h_1(x_m)-h_2(x_m))$

let $s = \mathrm{sgn}(h_1(x_m)-h_2(x_m))$

$$\le \frac{1}{2}\left(U_{m-1}(h_1) + s\,l\,h_1(x_m)\right) + \frac{1}{2}\left(U_{m-1}(h_2) - s\,l\,h_2(x_m)\right)$$

$$\le \frac{1}{2}\left(\sup_{h\in H} U_{m-1}(h) + s\,l\,h(x_m)\right) + \frac{1}{2}\left(\sup_{h\in H} U_{m-1}(h) - s\,l\,h(x_m)\right)$$

$(\sigma_m = +1)$ $\quad$ $(\sigma_m = -1)$

$$= \mathbb{E}_{\sigma_m}\left[\sup_{h\in H} U_{m-1}(h) + \sigma_m\, l\, h(x_m)\right]$$

$\sigma_m \in \{-1,+1\}$ u.a.r.

by arbitrary ε, $\mathbb{E}_{\sigma_m}\left[\sup_{h\in H} U_{m-1}(h) + \sigma_m(\Phi_m\circ h)(x_m)\right] \le \mathbb{E}_{\sigma_m}\left[\sup_{h\in H} U_{m-1}(h) + \sigma_m\, l\, h(x_m)\right]$

same way for all other σ_i: $\mathbb{E}_{\sigma}\left[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i(\Phi_m\circ h)(x_i)\right] \le \mathbb{E}_{\sigma}\left[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i\, l\, h(x_i)\right]$ $\square$

Thm 1 (Margin Bound for Binary Classification) fixed $\rho > 0$

$$\boxed{R(h) \le \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\,\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}}$$

0/1-loss

$$R(h) \le \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\,\hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{3}{\delta}}{2m}}$$

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

ρ-margin-loss

$\mathbb{1}_{h(x)\ne y}$

$\Downarrow$

$\Phi_\rho(y\,h(x))$

Proof: $\mathcal{H} \to \tilde{\mathcal{H}} = \{z=(x,y)\mapsto y\,h(x) : h\in\mathcal{H}\}$

$\to \tilde{\tilde{\mathcal{H}}} = \{\Phi_\rho\circ f : f\in\tilde{\mathcal{H}}\}$

w.p. $\geq 1-\delta$ $\quad \forall g \in \widetilde{\mathcal{H}}$ $\quad \mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(\widetilde{\mathcal{H}}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$

$\Downarrow$

$\forall h \in \mathcal{H}$. $\quad \mathbb{E}[\Phi_\rho(y_i h(x))] \leq \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(y_i h(x_i)) + 2\mathcal{R}_m(\Phi_\rho \circ \mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$

$R(h) = \mathbb{E}[\mathbb{1}_{yh(x)\leq 0}] \leq \mathbb{E}[\Phi_\rho(yh(x))] \leq \widehat{R}_{S,\rho}(h) + 2\mathcal{R}_m(\Phi_\rho \circ \mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$

Φ_ρ: $\frac{1}{\rho}$-Lipschitz. by Talagrand: $\mathcal{R}_m(\Phi_\rho \circ \widetilde{\mathcal{H}}) \leq \frac{1}{\rho}\mathcal{R}_m(\widetilde{\mathcal{H}})$

$\mathcal{R}_m(\widetilde{\mathcal{H}}) = \frac{1}{m}\mathbb{E}_S\left[\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i y_i h(x_i)\right]\right]$ $\quad$ (all Φ_i identical to Φ_ρ)

$= \frac{1}{m}\mathbb{E}_S\left[\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)\right]\right] = \mathcal{R}_m(\mathcal{H})$

$\Rightarrow R(h) \leq \widehat{R}_{S,\rho}(h) + \frac{2}{\rho}\cdot\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$ $\quad \square$

$\widehat{R}_{S,\rho}(h) \to \frac{1}{m}\sum_{i=1}^{m}\Phi_\rho(y_i h(x_i))$ $\rho\downarrow$

$\frac{2}{\rho} \to \rho\uparrow$

trade-off (与ρ取值有关)

Thm 2 fix $r>0$. $\forall \rho \in (0, r]$: $\quad$ w.p. $\geq 1-\delta$

$$R(h) \leq \widehat{R}_{S,\rho}(h) + \frac{4}{\rho}\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log\log_2\frac{2r}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

$$R(h) \leq \widehat{R}_{S,\rho}(h) + \frac{4}{\rho}\widehat{\mathcal{R}}_S(\mathcal{H}) + \sqrt{\frac{\log\log_2\frac{2r}{\rho}}{m}} + 3\sqrt{\frac{\log\frac{4}{\delta}}{2m}}$$

"modest"

proof in FML 96 Page.

Motivation: bound of SVM $\quad$ N independent

Proposition: $S \subseteq \{x : \|x\|_2 \le r\}$ bounded sample set

$\mathcal{H} = \{x \mapsto w \cdot x : \|w\|_2 \le \Lambda\}$ bounded hypothesis set.

"hyperplane"

$$\Rightarrow \hat{\mathcal{R}}_S(\mathcal{H}) \le \sqrt{\frac{r^2\Lambda^2}{m}}$$

bounded Rademacher Complexity

Proof:

$$\hat{\mathcal{R}}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E}_\sigma\left[\sup_{\|w\|_2 \le \Lambda} \sum_{i=1}^m \sigma_i \, w \cdot x_i\right]$$

$$= \frac{1}{m} \mathbb{E}_\sigma\left[\sup_{\|w\|_2 \le \Lambda} w \cdot \sum_{i=1}^m \sigma_i x_i\right]$$

Cauchy-Schwarz. $\vec{a} \cdot \vec{b} \le \|\vec{a}\|_2 \cdot \|\vec{b}\|_2$

$$\le \frac{\Lambda}{m} \cdot \mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i x_i\right\|_2\right]$$

Jenson $\mathbb{E}[X]^2 \le \mathbb{E}[X^2]$

$$\le \frac{\Lambda}{m} \mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i x_i\right\|_2^2\right]^{1/2}$$

$$= \frac{\Lambda}{m} \mathbb{E}_\sigma\left[\sum_{i=1}^m \sum_{j=1}^m \sigma_i \sigma_j (x_i \cdot x_j)\right]^{1/2}$$

$$\mathbb{E}[\sigma_i \cdot \sigma_j] = \begin{cases} 1, & i = j \\ 0, & i \ne j \end{cases}$$

$$\le \frac{\Lambda}{m} \left(\sum_{i=1}^m \underbrace{\|x_i\|^2}_{\le r}\right)^{\frac{1}{2}} \le \frac{\Lambda}{m} \cdot \sqrt{mr^2} = \sqrt{\frac{r^2\Lambda^2}{m}} \quad \square$$

Cor bounded $\mathcal{X} \subseteq \{x : \|x\| \le r\}$. bounded $\mathcal{H} = \{x \mapsto w \cdot x : \|w\| \le \Lambda\}$

fix $\rho > 0$:

$$R(h) \le \hat{R}_{S,\rho}(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

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

$\forall \rho \in (0, r]$:

$$R(h) \le \hat{R}_{S,\rho}(h) + 4\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log\log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log\frac{3}{\delta}}{2m}}$$

it close not depend directly on the dimension of the feature space.

margin: ρ

if $\rho = \rho_{geom}$: $\hat{R}_{S,\rho}(h) = 0$.

ρ_{geom} large $\rightarrow$ strong guarantee (bound)

bound: distribution-independent

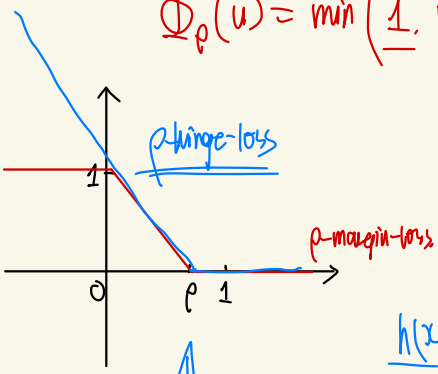
margin: distribution-dependent

Margin Theory

SVM: $h \in \{ x \mapsto w\cdot x : \|w\|_2 \le 1 \}$ $(\Lambda = 1)$

$$\Rightarrow R(h) \le \hat{R}_{S,\rho}(h) + 4\sqrt{\frac{r^2/\rho^2}{m}} + \sqrt{\frac{\log\log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

$$\Phi_\rho(u) = \min\left(1, \max\left(0, 1-\frac{u}{\rho}\right)\right) \le \max\left(0, 1-\frac{u}{\rho}\right)$$



$h(x_i)y_i \le 0$

$$\hat{R}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^m \Phi_\rho(y_i h(x_i))$$

$$\le \frac{1}{m}\sum_{i=1}^m \max\left(0, 1-\frac{y_i(w\cdot x_i)}{\rho}\right)$$

$h(x_i)$

$\|w\|_2 \le 1$

$\frac{h(x)}{\rho}$ 和 $h(x)$ 的经验误差相同

$$\Downarrow$$

$$\sim \frac{1}{m}\sum_{i=1}^m \max(0, 1-y_i h(x_i))$$

$\|w\|_2 \le \frac{1}{\rho}$

bound:

$$R(h) \le \frac{1}{m}\sum_{i=1}^m \max(0, 1-y_i(w\cdot x_i)) + 4\sqrt{\frac{r^2/\rho^2}{m}} + \sqrt{\frac{\log\log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

algorithm $\rightarrow$ $\left\{ \begin{array}{l} w \\ \rho \end{array} \right\}$ minimize RHS

ρ: hyperparameter (by cross-validation)

$$\Downarrow$$

$$\min \frac{1}{m}\sum_{i=1}^m \max(0, 1-y_i(w\cdot x_i)) \quad \text{s.t. } \|w\|_2^2 \le \frac{1}{\rho^2}$$

$$\Rightarrow \min_w \frac{1}{m}\sum_{i=1}^m \max(0, 1-y_i(w\cdot x_i)) + \lambda\cdot\|w\|_2^2$$

6. Kernel Methods 核方法

Recall：SVM与统计学习简史

1963：Vapnik 提出支持向量的概念
1968：Vapnik 和 Chervonenkis 提出 VC 维
1974：提出结构风险最小化原则 (SRM)
… … 苏联解体前一年(1990), Vapnik 来到美国
1995：Support Vector Network 文章发表
1995：《The Nature of Statistical Learning》出版
1998：SVM 在文本分类上取得巨大成功
1998：$\boxed{\text{《Statistical Learning Theory》出版}}$

ERM
$\Omega(f)$

Margin Theory

从样例中学习

$S=\{x_1,\cdots,x_n\}$

① → 赵小学习 西瓜书·Chapter 15

② → 连接主义学习 1980s

黑箱模型 BP

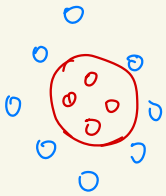
FML

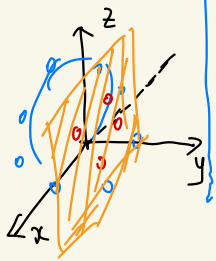
1990s：③ → 统计学习

④ 深度学习 { 数据大 计算快

not linearly separable.

non-linear decision boundary





○: $y\le 0$
•: $y\ge 0$

Principle: $x \to \Phi(x)$

$$K(x,x')=\langle \Phi(x),\Phi(x')\rangle$$

$\langle x,x'\rangle$

x: nonlinearly separable

$\Phi(x)$: linearly separable

升维

$\mathbb{H}$ (Hilbert Space)

① $\Phi(x)$ 维数高 e.g. document classification

vocabulary 10000 $=10^4$ words.

$\downarrow$

10^{12}

trigram.
(n-gram)

$\Downarrow$

But: N-independent

ρ: margin (泛化界不受影响)

② $\Phi(x)$、$\langle \Phi(x),\Phi(x')\rangle$ 计算复杂 (costly).

def (kernels) A function $K: \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is called a kernel over $\mathcal{X}$. "核"

$\forall x, x' \in \mathcal{X}$ $\exists \Phi$ s.t. $\boxed{K(x,x') = \langle \Phi(x), \Phi(x') \rangle}$ ☆☆

$\Phi: \mathcal{X}\to\mathbb{H}$ (feature mapping) $\mathbb{H}$: feature space. 特征空间

不需要显式知道 Φ 是什么，只需要知道 K 即可！

efficient

inner product: similarity measure

K

Thm (Mercer's Theorem) ☆☆

$K: \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ continuous, symmetric.

K is a kernel function.

K admits a uniformly convergent expansion $K(x,x') = \sum_{n=0}^{\infty} a_n \phi_n(x)\phi_n(x')$ with $a_n > 0$

iff $\forall$ square integrable function c s.t. $\iint_{\mathcal{X}\times\mathcal{X}'} c(x)\cdot c(x') K(x,x')\,dx\,dx' \geq 0$

$\Updownarrow$

K: PDS (positive definite symmetric)

def (PDS kernels): $K: \mathcal{X}\times\mathcal{X}\to\mathbb{R}$

if $\forall \{x_1, \cdots, x_m\} \subseteq \mathcal{X}$,

the matrix $\mathbf{K} = (K(x_i, x_j))_{m\times m}$ is symmetric and positive semi-definite. ☆☆☆

$$\mathbf{K} = \begin{pmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_m) \\ K(x_2,x_1) & K(x_2,x_2) & \cdots & K(x_2,x_m) \\ \vdots & \vdots & & \vdots \\ K(x_m,x_1) & K(x_m,x_2) & \cdots & K(x_m,x_m) \end{pmatrix}$$

核矩阵 (kernel matrix / Gram matrix)

$\to$ 对称

半正定 ① or ②

① $\mathbf{K}$ 特征值非负

② $\forall c = (c_1, \cdots, c_m)^T$

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

K 是核函数 $\Leftrightarrow$ K 是 PDS 核

$\Leftrightarrow$ $\mathbf{K}$ 是 SPSD 矩阵

Examples:

① Polynomial Kernel: $\forall x, x' \in \mathbb{R}^N$. $\boxed{K(x,x') = (x \cdot x' + c)^d}$ PDS核

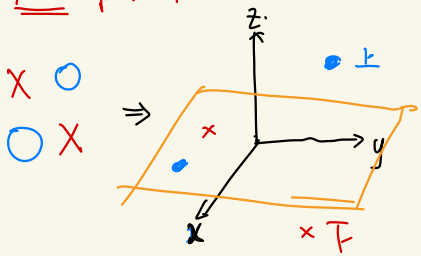
$$\Phi: \in \mathbb{H} \in \mathbb{R}^{\binom{N+d}{d}}$$

e.g. $N=2$. $d=2$. $K(x,x') = (x_1x_1' + x_2x_2' + c)^2$ 展开后有n项.

$$= x_1^2x_1'^2 + x_2^2x_2'^2 + 2x_1x_1'x_2x_2' + 2cx_1x_1' + 2cx_2x_2' + c^2$$

$$= \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2c}\,x_1 \\ \sqrt{2c}\,x_2 \\ c \end{pmatrix} \cdot \begin{pmatrix} x_1'^2 \\ x_2'^2 \\ \sqrt{2}x_1'x_2' \\ \sqrt{2c}\,x_1' \\ \sqrt{2c}\,x_2' \\ c \end{pmatrix} \qquad \binom{4}{2} = 6$$

XOR problem



② Gaussian Kernel (RBF: Radial Basis Function) 径向基函数

$$\boxed{K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)}$$

normalization.

$$\sum_{n=0}^{\infty} \frac{(x \cdot x')^n}{n!\,\sigma^{2n}} = \exp\left(\frac{x \cdot x'}{\sigma^2}\right)$$

③ Laplacian Kernel

$$K(x,x') = \exp\left(-\frac{\|x-x'\|}{\sigma}\right)$$

(L_1-regularization → sparsity)

④ Sigmoid Kernel $K(x,x') = \tanh(a(x \cdot x') + b)$ $\langle x, x' \rangle \to K(x,x')$

$K(x,x') = \langle \Phi(x), \Phi(x') \rangle$ $\Phi(x) \in \mathbb{H}$

Lemma (Cauchy-Schwarz inequality for PDS kernel)

$\forall x, x' \in \mathcal{X}$. $K(x,x')^2 \le K(x,x) \cdot K(x',x')$

如何选择核函数?

e.g. 文本分类 n-gram → polynomial.

像素不明 → Gaussian.

kernel tricks

↓

kernel methods

Proof: Kernel matrix $\mathbb{K} = \begin{pmatrix} K(x,x) & K(x,x') \\ K(x',x) & K(x',x') \end{pmatrix}$ $\quad K: \text{PDS} \Rightarrow \mathbb{K}: \text{SPSD}$

eigenvalue $\lambda_1, \lambda_2 \geq 0 \Rightarrow \lambda_1 \lambda_2 = \det(\mathbb{K}) \geq 0 \Rightarrow K(x,x)K(x',x') \geq K(x,x')^2$ $\square$

Thm $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is PDS kernel $\Rightarrow \exists$ Hilbert Space $\mathbb{H}$ mapping $\Phi: \mathcal{X} \to \mathbb{H}$

s.t. $\forall x, x' \in \mathcal{X}$. $K(x,x') = \langle \Phi(x), \Phi(x') \rangle$

☆☆☆

$\mathbb{H}$ follows "reproducing property"

$\forall h \in \mathbb{H}$. $\forall x \in \mathcal{X}$. $h(x) = \langle h, K(x, \cdot) \rangle$

值 函数 函数

Reproducing Kernel Hilbert Space (RKHS) 再生核希尔伯特空间

PDS 核 $\Rightarrow$ RKHS

Proof: 证明 Hilbert Space
- 内积空间: $\langle \cdot, \cdot \rangle$ (正定, 对称, 线性)
- 完备 (Complete): all cauchy sequence converges.

$\forall x \in \mathcal{X}$ $\Phi(x): \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$: $\forall x' \in \mathcal{X}$ $\boxed{\Phi(x)(x') = K(x,x')}$

函数

$\mathbb{H}_0 = \left\{ \sum_{i \in I} a_i \Phi(x_i) : a_i \in \mathbb{R}, x_i \in \mathcal{X}, |I| < \infty \right\}$

(linear combination of all functions $\Phi(x)$)

define inner product

$f = \sum_{i \in I} a_i \Phi(x_i)$, $g = \sum_{j \in J} b_j \Phi(x_j')$, $\Rightarrow$ ☆ $\boxed{\langle f, g \rangle = \sum_{i \in I} \sum_{j \in J} a_i b_j K(x_i, x_j')}$

$= \sum_{j \in J} b_j \cdot \sum_{i \in I} a_i K(x_i, x_j')$

$= \sum_{j \in J} b_j \cdot \sum_{i \in I} a_i \Phi(x_i)(x_j')$

$= \sum_{j \in J} b_j \cdot f(x_j')$

$\left(= \sum_{i \in I} a_i \cdot g(x_i) \right)$

① definiteness: K PDS

$\Rightarrow \langle f, f \rangle = \sum_{i,j \in I} a_i a_j K(x_i, x_j) \geq 0$

② symmetry: K PDS

③ bi-linearity $\langle \cdot, \cdot \rangle$

e.g. $\langle f+f', g\rangle = \sum_{j\in J} b_j\big(f(x_j)+f'(x_j)\big) = \sum_{j\in J} b_j f(x_j) + \sum_{j\in J} b_j f'(x_j)$

$\left(\forall f_1,\cdots,f_m\in \mathbb{H},\ c_1,\cdots,c_m\in\mathbb{R},\ \sum_{i=1}^m\sum_{j=1}^m c_i c_j\langle f_i, f_j\rangle = \left\langle \sum_{i=1}^m c_i f_i, \sum_{i=1}^m c_j f_j\right\rangle \geq 0\right)$

$\Rightarrow \langle\cdot,\cdot\rangle$ is an inner product. $\Rightarrow$ $\mathbb{H}_0$ is an inner-product space.

(pre-Hilbert)

Completion Theorem (完备性):

☆ 任意度量空间 $\mathbb{H}_0$ 都存在一个完备的度量空间 $\mathbb{H}$. s.t. $\mathbb{H}_0$ 是 $\mathbb{H}$ 的一个稠密子空间

$\mathbb{H}_0 \subseteq \mathbb{H}$

$\mathbb{H}_0 \xrightarrow{\text{completion}} \mathbb{H}$: Hilbert Space.

Reproducing property: $h = \sum_{i\in I} a_i \Phi(x_i) \in \mathbb{H}$.

$$\forall x\in\mathcal{X}:\ h(x) = \sum_{i\in I} a_i \Phi(x_i)(x) = \sum_{i\in I} a_i K(x_i, x) = \langle h, \underbrace{\Phi(x)}_{K(x,\cdot)}\rangle \qquad \square$$

PDS kernel K $\Rightarrow$ RKHS $\mathbb{H}$

$\mathbb{H}$: feature space

Φ: feature mapping

$\|\cdot\|_{\mathbb{H}} = \sqrt{\langle\cdot,\cdot\rangle_{\mathbb{H}}}$

不需找 Φ. 只需找 K

不一定唯一.

Properties of kernel:

① Normalization $K(x,x') \xrightarrow{\text{normalize}} \tilde{K}(x,x') = \begin{cases} 0, & \text{if } K(x,x)=0 \text{ or } K(x',x')=0 \\ \dfrac{K(x,x')}{\sqrt{K(x,x)\cdot K(x',x')}} & \text{otherwise.} \end{cases}$

$\forall x\in\mathcal{X}$, $K(x,x)=1$ (归一)

$\exp\left(-\dfrac{\|x-x'\|^2}{2\sigma^2}\right) \leftarrow \exp\left(\dfrac{x\cdot x'}{\sigma^2}\right) = K(x,x')$

"a similarity measure"

$$\frac{K(x,x')}{\sqrt{K(x,x)\cdot K(x',x')}} = \frac{e^{\frac{x\cdot x'}{\sigma^2}}}{e^{\frac{\|x\|^2}{2\sigma^2}}\cdot e^{\frac{\|x'\|^2}{2\sigma^2}}} = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$$

Proposition: K PDS kernel $\xrightarrow{\text{normalization}}$ $\tilde{K}$ PDS kernel

Proof: $\{x_1, \cdots, x_m\} \subseteq \mathcal{X}$. $c \in \mathbb{R}^m$.

PSD
$$\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \tilde{K}(x_i, x_j) = \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{c_i c_j K(x_i, x_j)}{\sqrt{K(x_i, x_i)\cdot K(x_j, x_j)}} \quad (K(x_i,x_i),\ K(x_j,x_j) > 0)$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{c_i c_j \langle \Phi(x_i), \Phi(x_j)\rangle}{\|\Phi(x_i)\|_{\mathbb{H}} \cdot \|\Phi(x_j)\|_{\mathbb{H}}} = \left\| \sum_{i=1}^{m} \frac{c_i \Phi(x_i)}{\|\Phi(x_i)\|_{\mathbb{H}}} \right\|_{\mathbb{H}}^2 \geq 0. \quad \square$$

② **def** (empirical feature map) given $\{x_1, \cdots, x_m\} \subseteq \mathcal{X}$.

已知K 寻找Φ

$$\Phi: \mathcal{X} \to \mathbb{R}^m. \qquad \forall x \in \mathcal{X}.\ \Phi(x) = \begin{bmatrix} K(x, x_1) \\ K(x, x_2) \\ \vdots \\ K(x, x_m) \end{bmatrix}$$

"K- similarity" $(\Phi(x)(x_i) = x_i)$

$$\Phi(x_i) = \mathbb{K} \cdot e_i \quad (\text{核矩阵m第i列})$$

$$\langle \Phi(x_i), \Phi(x_j)\rangle = (\mathbb{K} e_i)^T \cdot (\mathbb{K} e_j)$$
$$= e_i^T \mathbb{K}^2 e_j = (\mathbb{K}^2)_{ij}$$

$$K = \begin{pmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_m) \\ K(x_2,x_1) & K(x_2,x_2) & \cdots & K(x_2,x_m) \\ \vdots & \vdots & & \vdots \\ K(x_m,x_1) & K(x_m,x_2) & \cdots & K(x_m,x_m) \end{pmatrix}$$

$\Rightarrow$ the kernel matrix associated with Φ is K^2

Consider $\underbrace{(\mathbb{K}^+)^{\frac{1}{2}}}_{\text{by SVD of } K^+}$, where $\mathbb{K}^+$ is pseudo-inverse of $\mathbb{K}$

$$\begin{cases} K \text{ invertible}: & K^+ = K^{-1} \\ K \text{ non-invertible}: & K^+ = K^T \cdot (KK^T)^{-1} \end{cases}$$

define $\Psi(x) = (\mathbb{K}^+)^{\frac{1}{2}} \Phi(x) \quad \forall x \in \mathcal{X}$

$$\Rightarrow \langle \Psi(x_i), \Psi(x_j)\rangle = \left((\mathbb{K}^+)^{\frac{1}{2}} \mathbb{K} e_i\right)^T \left((\mathbb{K}^+)^{\frac{1}{2}} \mathbb{K} e_j\right)$$

$\mathbb{K}\mathbb{K}^+\mathbb{K} = \mathbb{K}$

$$= e_i^T \underbrace{\mathbb{K}^T \left((\mathbb{K}^+)^{\frac{1}{2}}\right)^T}_{\mathbb{K}} \underbrace{(\mathbb{K}^+)^{\frac{1}{2}} \mathbb{K}}_{\mathbb{K}^+} e_j = e_i^T \mathbb{K} e_j$$

$\Rightarrow$ the kernel matrix associated with Ψ is $\mathbb{K}$

define $\Omega(x) = \mathbb{K}^{\dagger}\Phi(x) \quad \forall x \in \mathcal{X}$

$$\langle \Omega(x_i), \Omega(x_j) \rangle = e_i^T \mathbb{K}^T (\mathbb{K}^{\dagger})^T (\mathbb{K}^{\dagger}\mathbb{K} e_j) = e_i^T e_j = \begin{cases} 1, & i=j \\ 0, & i \neq j \end{cases}$$

$\Rightarrow$ the kernel matrix associated with Ω is $\mathbb{I}$.

③ Closure under: 求和 sum. 求积 product. 直积 Cartesian product. 极限 pointwise limit. 幂级数 power series

闭包性质

$\sum_{n=0}^{N} a_n K^n \quad \forall N.$

$(K_n)_{n\geq 1}$

$K(x_1, x_1') \cdot K'(x_2, x_2')$

$\parallel$

$(K \otimes K')(x_1, x_1', x_2, x_2')$

PDS kernel

Proof: e.g. $c^T K c \geq 0$, $c^T K' c \geq 0$ $\Rightarrow c^T(K+K')c \geq 0$

others similar. □.

Kernel-Based Algorithms

SVM with Kernel Methods

dual:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0 \quad i \in [m]$

$0 \leq \alpha_i \leq C \quad i \in [m]$.

element-wise (Hadamard product)

$$\max_{\alpha} \; 2 \cdot \mathbb{1}^T \alpha - (\alpha * y)^T \mathbb{K} (\alpha * y)$$

s.t. $0 \preceq \alpha \preceq C$

$\alpha^T y = 0$

$$h(x) = \operatorname{sgn}\left(\sum_{i \in SV} \alpha_i y_i K(x_i, x) + b\right).$$

"核化支持向量机"

extension: $\mathcal{X} \to \mathbb{H}$

large-margin separation

如何使用SVM?

入门: 使用各种版本SVM

专业: 尝试, 组合核函数

☆专家: 根据问题设计目标函数, 算法搜索

度身定制

Vapnik: "Nothing is more practical than a good theory."

Thm (Representer Theorem) 表示定理

$K: \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ PDS kernel,

$\mathbb{H}$: RKHS associated with K

$\forall$ non-decreasing function $G:\mathbb{R}\to\mathbb{R}$,

$\forall$ loss-function $\mathcal{L}: \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$.

带结构风险

结构风险 ↑ 正则化项

经验风险 ↑ 损失项

$\Rightarrow$ the optimization problem

$$\boxed{\underset{h\in\mathbb{H}}{\operatorname{argmin}}\, F(h) = \underset{h\in\mathbb{H}}{\operatorname{argmin}}\, G(\|h\|_{\mathbb{H}}) + \mathcal{L}(h(x_1),\cdots,h(x_m))}$$

admits the solution of the form $\boxed{h^* = \sum_{i=1}^{m}\alpha_i K(x_i,\cdot)}$

Furthermore, G increasing, then any solution has this form.

解是核函数的线性组合

Proof: $\mathbb{H}_1 = \operatorname{span}(\{K(x_1,\cdot),\cdots,K(x_m,\cdot)\})$ ← $\left(h_1\in\mathbb{H}: h_1=\sum_{i=1}^{m}\alpha_i K(x_i,\cdot)\right)$

$\Phi(x_1),\cdots,\Phi(x_m)$

$\mathbb{H}$ 正交分解: $\mathbb{H} = \mathbb{H}_1 \oplus \mathbb{H}_1^{\perp}$. $\oplus$: direct sum. $\forall x\in V\oplus W$

$\forall h\in\mathbb{H} \Rightarrow h = h_1 \oplus h_1^{\perp}$. $h_1\in\mathbb{H}_1$

$\exists!\ v\in V,\ w\in W$ s.t. $x = v+w$

G: non-decreasing,

$$G(\|h_1\|_{\mathbb{H}}) \le G\left(\sqrt{\|h_1\|_{\mathbb{H}}^2 + \|h_1^{\perp}\|_{\mathbb{H}}^2}\right) = G(\|h\|_{\mathbb{H}}) \quad ①$$

By producing property. $h(x_i) = \langle h, K(x_i,\cdot)\rangle = h_1(x_i)$

$h_1\oplus h_1^{\perp}$ $\quad$ ↳ $\in\mathbb{H}_1$

$\Rightarrow \mathcal{L}(h(x_1),\cdots,h(x_m)) = \mathcal{L}(h_1(x_1),\cdots,h_1(x_m))$ ②

①.② $\Rightarrow F(h_1)\le F(h)$. Furthermore, G increasing, $F(h_1)<F(h)$ □

Learning Guarantee

$\mathcal{H} = \{ h \in \mathbb{H} : \|h\|_{\mathbb{H}} \le \Lambda \}$ $\quad h(x) = \langle h, K(x,\cdot) \rangle = \langle h, \Phi(x) \rangle$

$S = \{ x_1, \cdots, x_m : \underbrace{K(x_i, x_i) \le r^2}_{\|\Phi(x_i)\|_{\mathbb{H}}^2 \le r^2} \}$

Thm (Rademacher Complexity of Kernel-based algorithm)

$\mathcal{H} = \{ x \mapsto \langle w, \Phi(x) \rangle : \|w\|_{\mathbb{H}} \le \Lambda \}$ $\Rightarrow$ $\boxed{\hat{R}_S(\mathcal{H}) \le \frac{\Lambda \cdot \sqrt{tr(K)}}{m} \le \sqrt{\frac{r^2 \Lambda^2}{m}}}$

Proof,

$$\hat{R}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E}_\sigma \Big[\sup_{\|w\| \le \Lambda} \Big\langle w, \sum_{i=1}^m \sigma_i \Phi(x_i) \Big\rangle \Big]$$

$$\le \frac{\Lambda}{m} \mathbb{E}_\sigma \Big[\Big\| \sum_{i=1}^m \sigma_i \Phi(x_i) \Big\|_{\mathbb{H}} \Big] \qquad \text{Cauchy-Schwarz}$$

$$\le \frac{\Lambda}{m} \mathbb{E}_\sigma \Big[\Big\| \sum_{i=1}^m \sigma_i \Phi(x_i) \Big\|_{\mathbb{H}}^2 \Big]^{1/2} \qquad \text{Jenson.}$$

$$= \frac{\Lambda}{m} \Big(\sum_{i=1}^m \|\Phi(x_i)\|_{\mathbb{H}}^2 \Big)^{1/2} \qquad \begin{cases} i \ne j : \mathbb{E}[\sigma_i \sigma_j] = 0 \\ i = j \quad \mathbb{E}[\sigma_i \sigma_j] = 1 \end{cases}$$

$$= \frac{\Lambda}{m} \Big(\sum_{i=1}^m K(x_i, x_i) \Big)^{1/2} = \frac{\Lambda}{m} \cdot \sqrt{tr(K)} \le \sqrt{\frac{r^2 \Lambda^2}{m}} \qquad \square$$

Khintchine-Kahane inequality: $\frac{1}{2} \sum_{i=1}^m \|x_i\|^2 \le \mathbb{E}_\sigma \Big[\Big\| \sum_{i=1}^m \sigma_i x_i \Big\| \Big]^2 \le \sum_{i=1}^m \|x_i\|^2$

$$\Rightarrow \frac{1}{\sqrt{2}} \frac{\Lambda}{m} \cdot \sqrt{tr(K)} \le \hat{R}_S(\mathcal{H}) \le \frac{\Lambda}{m} \cdot \sqrt{tr(K)} \qquad \star$$

Kernel Matrix very important!

Cor (Margin Bounds for kernel-based hypothesis) $\quad$ fixed $\rho > 0$.

$$R(h) \le \hat{R}_{S,\rho}(h) + 2 \sqrt{\frac{r^2 \Lambda^2 / \rho^2}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

$$R(h) \le \hat{R}_{S,\rho}(h) + 2 \cdot \sqrt{\frac{tr(K) \Lambda^2 / \rho^2}{m}} + 3 \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$