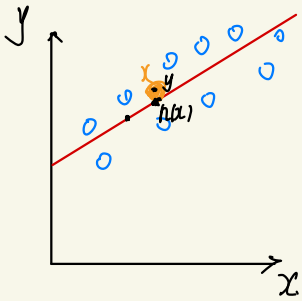
11. Regression 回归

二分类 . 多分类

: 输出为连续值
标签为连续值



The problem of regression

$\mathcal{X}$: input space

$\mathcal{Y}$: label space (continuous)

$f: \mathcal{X} \to \mathcal{Y}$

$S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$

$\forall x \in \mathcal{X}$ 找 $h: \mathcal{X} \to \mathcal{Y}$ s.t. $h(x) \sim y$

$L: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ measure error (difference between prediction and true target)

squared loss: $L(y, y') = |y' - y|^2$

L_p-loss: $L_p(y, y') = |y' - y|^p$

分类: $\mathbb{1}_{h(x) \neq y}$ equality inequality

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

Empirical error: $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(x_i), y_i)$

$\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$ MSE 均方误差

Assume $\forall y, y' \in \mathcal{Y}$. $\boxed{L(y, y') \leq M}$ → bounded regression problem

(* unbounded → { assume "envelope" / assume "bounded moment")

① Finite Hypothesis Space

Thm w.p. $\geq 1-\delta$ $R(h) \leq \hat{R}_S(h) + M \cdot \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}$, $\forall h \in \mathcal{H}$

w.p. $\geq 1-\delta$ $R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}$

Proof: By Hoeffding's inequality, $L(y,y') \in [0,M]$.

$$\Pr\left(R(h)-\hat{R}_S(h) > \varepsilon\right) \le e^{-\frac{2\varepsilon^2}{m\cdot\frac{M^2}{m^2}}} = e^{-\frac{2m\varepsilon^2}{M^2}}$$

Union Bound: $\Pr\left(\exists h\in\mathcal{H}: R(h)-\hat{R}_S(h) > \varepsilon\right) \le \sum_{h\in\mathcal{H}} \Pr\left(R(h)-\hat{R}_S(h) > \varepsilon\right) \le |\mathcal{H}|\cdot e^{-\frac{2m\varepsilon^2}{M^2}} \le \delta$

$$\Rightarrow R(h) \le \hat{R}_S(h) + M\cdot\sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}} \quad \text{w.p.} \ge 1-\delta \quad \forall h\in\mathcal{H} \quad \square$$

Classification: $M=1$.

$|\mathcal{H}|$不能太大 $\Rightarrow$ Occam's Razor.

② Rademacher Complexity

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

数据相关

Lemma for any fixed $y'\in\mathcal{Y}$. $y\mapsto L(y,y')$ μ-Lipschitz.

$$G = \{(x,y)\mapsto L(h(x),y) : h\in\mathcal{H}\} \Rightarrow \boxed{\hat{R}_S(G) \le \mu\cdot\hat{R}_S(\mathcal{H})}$$

(by Talagrand Lemma, $\tilde{\mathcal{H}} = \Phi\circ\mathcal{H}$. $\hat{R}_S(\tilde{\mathcal{H}}) \le l\cdot\hat{R}_S(\mathcal{H})$. Φ Lipschitz)

Thm (Rademacher Complexity Bound) L: μ-Lipschitz

$$\mathbb{E}_{(x,y)\sim D}\left[L(h(x),y)\right] \le \frac{1}{m}\sum_{i=1}^m L(h(x_i),y_i) + 2\mu\cdot\mathcal{R}_m(\mathcal{H}) + M\cdot\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$\mathbb{E}_{(x,y)\sim D}\left[L(h(x),y)\right] \le \frac{1}{m}\sum_{i=1}^m L(h(x_i),y_i) + 2\mu\cdot\hat{R}_S(\mathcal{H}) + 3M\cdot\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

泛化误差　经验误差

McDiarmid

$$\mathbb{E}[g(Z)] \le \frac{1}{m}\sum_{i=1}^m g(Z_i) + 2\mathcal{R}_m(G) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

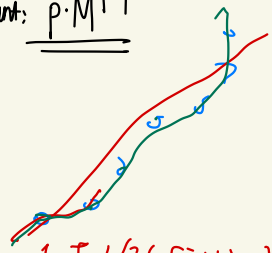
$\forall g\in[0,1]$

ℓ_p-loss: $\mathbb{E}_{(x,y)\sim D}\left[|h(x)-y|^p\right] \le \frac{1}{m}\sum_{i=1}^{m}|h(x_i)-y_i|^p + 2p\cdot M^{p-1}\mathcal{R}_m(\mathcal{H}) + M^p\cdot\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$

assume that $\forall (x,y)$, $|h(x)-y| \le M \Rightarrow \ell(h(x),y) \le M^p$

Lipschitz constant: $p\cdot M^{p-1}$

trade-off: $\begin{cases} \mathcal{H} \text{ rich}: \mathcal{R}_m(\mathcal{H}) \uparrow \\ \mathcal{H} \text{ not rich}: |h(x_i)-y_i| \uparrow \end{cases}$



✓ 拟合(fitting)

✗ 插值(interpolation)

③ VC-Dimension: 最多能把多少个点打散!

shattering → 分类

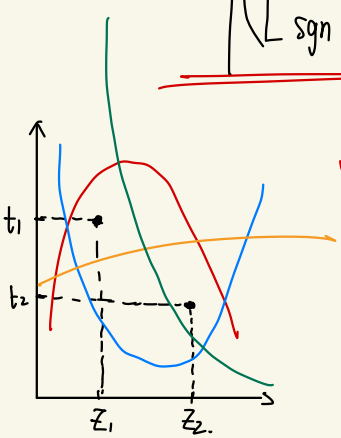
def (Shattering) G is a family of functions from $\mathcal{Z} \to \mathbb{R}$.

A set $\{z_1,\cdots,z_d\} \subseteq \mathcal{Z}$ is shattered by G if $\exists\ t_1,\cdots,t_d \in \mathbb{R}$

"thershold"

s.t. $\left|\left\{\begin{bmatrix} \mathrm{sgn}(g(z_1)-t_1) \\ \mathrm{sgn}(g(z_2)-t_2) \\ \vdots \\ \mathrm{sgn}(g(z_d)-t_d) \end{bmatrix} : g\in G\right\}\right| = 2^d$

witness 见证.



$g(z_i)-t_i$ 的符号任意 < 正 负

找到一组 witnesses $t_1,\cdots,t_d$ s.t. $\forall$ subset $A \subseteq \{z_1,\cdots,z_d\}$

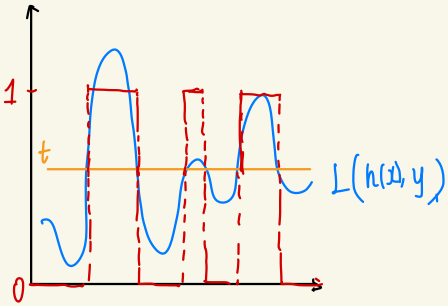
$\exists\ g\in G$ s.t. $g(z_i) \ge t_i$, $z_i \in A$

$g(z_i) \le t_i$, $z_i \in A^c$

def (Pseudo-Dimension) $\mathrm{Pdim}(G)$ is the size of the largest set shattered by G.

伪维度

$$\mathrm{Pdim}(\mathcal{G}) = \mathrm{VCdim}\left(\{(x,t) \mapsto \mathbb{1}_{(g(x)-t)>0} : g \in \mathcal{G}\} \right)$$



examples:

1. $\mathrm{Pdim}\left(\{x \mapsto w \cdot x + b : w \in \mathbb{R}^N, b \in \mathbb{R}\}\right) = N+1.$

2. Vector Space $\quad \mathrm{Pdim}(\mathcal{H}) = \dim(\mathcal{H})$

Thm $\quad \mathcal{G} = \{(x,y) \mapsto L(h(x), y),\ h \in \mathcal{H}\} \qquad \mathrm{Pdim}(\mathcal{G}) = d.$

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

$$R(h) \leq \hat{R}_S(h) + M \cdot \sqrt{\frac{2d \log \frac{em}{d}}{m}} + M \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

Proof: 回归→缺

缺态 $c(h,t)$: $(x,y) \mapsto \mathbb{1}_{L(h(x),y)>t}$

$$R(c(h,t)) = \Pr_{(x,y)\sim\mathcal{D}}\left(L(h(x),y) > t\right)$$

$$\hat{R}_S(c(h,t)) = \Pr_{(x,y)\sim\hat{\mathcal{D}}}\left(L(h(x),y) > t\right)$$

$\hat{\mathcal{D}}$: empirical distribution

回归态

$$|R(h) - \hat{R}_S(h)| = \left| \mathbb{E}_{(x,y)\sim\mathcal{D}}[L(h(x),y)] - \mathbb{E}_{(x,y)\sim\hat{\mathcal{D}}}[L(h(x),y)] \right|$$

$$= \left| \int_0^M \left(\Pr_{(x,y)\sim\mathcal{D}}\left(L(h(x),y)>t\right) - \Pr_{(x,y)\sim\hat{\mathcal{D}}}\left(L(h(x),y)>t\right) \right) dt \right|$$

$$\leq M \cdot \sup_{t\in[0,M]} \left| \Pr_{(x,y)\sim\mathcal{D}}\left(L(h(x),y)>t\right) - \Pr_{(x,y)\sim\hat{\mathcal{D}}}\left(L(h(x),y)>t\right) \right|$$

$$= M \cdot \sup_{t\in[0,M]} \left| R(c(h,t)) - \hat{R}_S(c(h,t)) \right|$$

缺态

$\Downarrow$

Double Counting $\quad$ non-negative r.v X.

$$\int_0^\infty \underbrace{\Pr(X>x)}_{1-F(x)}\, dx \qquad f_X(x) \text{ density}$$

$$= \int_0^\infty \left(\int_x^\infty f_X(u)\, du \right) dx.$$

$$= \int_{u=0}^\infty \left(\int_{x=0}^u f_X(u)\, dx \right) du.$$

$$= \int_{u=0}^\infty f_X(u) \cdot u \cdot du = \mathbb{E}[X].$$

$$\Pr\left(\sup_{h\in H} |R(h)-\hat{R}_S(h)| > \varepsilon\right)$$

Change of Variables $\quad \mathbb{E}[f(X)] = \int_0^\infty \Pr(f(x)>x)\,dx.$

LOTUS (Law of the Unconscious Statistics)

$$\le \Pr\left(\sup_{\substack{h\in H\\ t\in[0,M]}} |R(c(h,t)) - \hat{R}_S(c(h,t))| > \frac{\varepsilon}{M}\right)$$

by VC dim bound, $R(h) \le \hat{R}_S(h) + \sqrt{\frac{2d\log\frac{em}{d}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} = \frac{\varepsilon}{M}$

sup → ∃ $t\in[0,M]$: $\quad \mathrm{Pdim}(G) = \mathrm{VCdim}(C)$ $\Downarrow$

Pdim Bound: $\boxed{R(h) \le \hat{R}_S(h) + M\cdot\sqrt{\frac{2d\log\frac{em}{d}}{m}} + M\cdot\sqrt{\frac{\log\frac{1}{\delta}}{2m}}}$ $\square$

However, Pdim is not a scale-sensitive notion 数据无关,

(the measure value changes with the scale of the dataset)

size

*extension: scale-sensitive shattering. Scaling Law

def (γ-shattering) $\exists\, t_1,\cdots,t_d \in \mathbb{R}$,

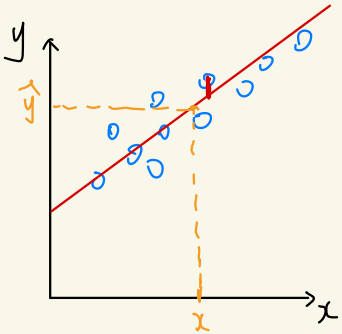
$\forall y\in\{-1,+1\}^m$, $\forall i\in[m]$, $y_i(g(z_i)-t_i) \ge \gamma$

def (γ-fat-dimension) $\mathrm{fat}_\gamma(G)$ is the size of the largest set that can be shattered by G.

Regression Algorithms

① Linear Regression $\quad \mathcal{H} = \{x \mapsto w\cdot x + b,\ w\in\mathbb{R}^N,\ b\in\mathbb{R}\}$

$$S=\{(x_1,y_1),\cdots,(x_m,y_m)\} \Rightarrow \min_{w,b} \frac{1}{m}\sum_{i=1}^{m}\Big(\overbrace{w\cdot x_i+b}^{\Phi(x_i)} - y_i\Big)^2$$



$$X=\begin{pmatrix} x_1 & \cdots & x_m \\ 1 & & 1 \end{pmatrix} \in \mathbb{R}^{(n+1)\times m}$$

Least Square (最小二乘)

$$w=\begin{pmatrix} w_1 \\ \vdots \\ w_N \\ b \end{pmatrix}, \quad y=\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

$\boxed{Ax=b}$ ~~$\Rightarrow x=A^{-1}b$~~

under-determinant 欠定

$y=w^T x.$

$$\boxed{\min_w F(w)=\frac{1}{m}\|X^T w-y\|_2^2,}$$

convex

$$\Rightarrow \nabla_w F(w)=\frac{\partial}{\partial w}\left((X^T w-y)^T(X^T w-y)\right)$$

$$=\frac{\partial}{\partial w}\left(w^T X X^T w - w^T X y - y X^T w + y^2\right)$$

$$= 2XX^T w - 2Xy = 0 \Rightarrow \boxed{XX^T w = Xy}$$

normal equation 正规方程

XX^T: PSD 半正定

若 XX^T 可逆, $W=(XX^T)^{-1}Xy$

若 XX^T 不可逆, pseudo-inverse

A^+ satisfies: $AA^+A=A$

$A^+AA^+=A^+$

$AA^+ \overset{?}{\neq} I$

$A\in\mathbb{R}^{m\times n}$ $\begin{cases} m\geq n, \text{ 列满秩}, & (A^TA)^{-1}A^T=A^L \\ n\geq m, \text{ 行满秩}, & A^T(AA^T)^{-1}=A^R \end{cases}$

直接解方程组!

e.g. LUP 分解

幂迭代

共轭梯度

$(XX^T)^+$.

$$W=\boxed{(XX^T)^+XY}+\left(I-(XX^T)^+(XX^T)\right)W_0$$

inductive bias → simple : norm ↓ $W=(XX^T)^{\dagger}XY$ arbitary matrix

$$\Rightarrow W=\begin{cases}(XX^T)^{-1}XY, & XX^T \text{ invertible}\\ (XX^T)^{\dagger}XY, & \text{o.w.}\end{cases}$$

closed-form solution

XX^T: $O(m\cdot N^2)$ $(XX^T)^{-1}\cdot(XX^T)^{\dagger}$: $O(N^3)$

$\Rightarrow O(mN^2+N^3)$

Singular-Value-Decomposition

奇异值分解 $M=U_M\Sigma_M V_M^T$

$M^{\dagger}=U_M\Sigma_M^{-1}V_M^T$ 正交 对角

矩阵求逆 ⇔ 矩阵乘法

Linear Regression:

~~strong generalization bound~~

performance: typically poor

② Ridge Regression 岭回归

$$R(h)\le \hat{R}_S(h)+4M\cdot\sqrt{\frac{r^2\Lambda^2}{m}}+M^2\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \quad \text{w.p.} \ge 1-\delta$$

L2-Regularization: (Tikhonov regularization)

$$\min_w F(w)=\lambda\cdot\|w\|_2^2+\sum_{i=1}^{m}(w\cdot \underbrace{x_i}_{\Phi(x_i)}-y_i)^2$$

$$\begin{cases}|h(x)-y|\le M\\ \|w\|\le\Lambda\\ \|x\|\le r\end{cases}$$

$$\Downarrow$$

$$\boxed{\min_w F(w)=\lambda\cdot\|w\|_2^2+\|X^Tw-y\|_2^2} \quad (*)$$

正则项 经验风险

convex $\lambda>0$

$$\nabla F(w) = 0 \Leftrightarrow 2\lambda \cdot w + 2(XX^T w - Xy) = 0 \Leftrightarrow (XX^T + \lambda I)\cdot w = Xy$$

normal equation

XX^T: PSD 半正定

$XX^T + \lambda I$: PD 正定 $\Rightarrow$ 可逆 $\Rightarrow$ $w = (XX^T + \lambda I)^{-1} \cdot Xy$

closed-form

λ by cross-validation

Another form:

$$\min \|X^T w - y\|_2^2 \quad \text{s.t.} \quad \|w\|_2^2 \le \Lambda^2 \qquad (**)$$

Lagrangian $\mathcal{L} = \|X^T w - y\|_2^2 + \lambda(\|w\|_2^2 - \Lambda^2)$

KKT condition: $\nabla_w \mathcal{L} = 2X(X^T w - y) + 2\lambda \cdot w = 0$

Kernelized ridge regression (KRR)

$$R(h) \le \hat{R}_S(h) + \frac{4M\Lambda}{m}\sqrt{tr(K)} + 3M^2\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

$4M\sqrt{\frac{r^2\Lambda^2}{m}}$

$|h(x) - y| \le M$

$\|w\|_{\mathbb{H}} \le \Lambda$

$\sqrt{K(x,x)} = \|\Phi(x)\|_{\mathbb{H}} \le r$.

$\dim(\mathbb{H}) = N$

$$\min_w F(w) = \lambda \cdot \|w\|^2 + \sum_{i=1}^m (w \cdot \Phi(x_i) - y_i)^2$$

$$\boxed{X = (\Phi(x_1), \cdots, \Phi(x_m)) \in \mathbb{R}^{N \times m}}$$

$$\min_w \sum_{i=1}^m (w \cdot \Phi(x_i) - y_i)^2 \quad \text{s.t.} \quad \|w\|^2 \le \Lambda^2$$

$$\Rightarrow \min_w \sum_{i=1}^m \xi_i^2 \quad \text{s.t.} \quad \|w\|^2 \le \Lambda^2, \quad \xi_i = y_i - w \cdot \Phi(x_i) \quad i \in [m]$$

Lagrangian: $\mathcal{L}(\xi, w, \alpha', \lambda) = \sum_{i=1}^m \xi_i^2 + \sum_{i=1}^m \alpha_i'(y_i - \xi_i - w \cdot \Phi(x_i)) + \lambda(\|w\|^2 - \Lambda^2)$

KKT-condition: $\nabla_w L = -\sum_{i=1}^{m} \alpha_i' \Phi(x_i) + 2\lambda w = 0 \Rightarrow w = \frac{1}{2\lambda}\sum_{i=1}^{m} \alpha_i' \Phi(x_i)$

$$\begin{cases} \nabla_{\xi_i} L = 2\xi_i - \alpha_i' = 0 \Rightarrow \xi_i = \frac{\alpha_i'}{2} \\ \text{Complementary slackness} \quad \forall i \in [m] \quad \alpha_i'\left(y_i - \xi_i - w\cdot\Phi(x_i)\right) = 0 \\ \lambda\cdot\left(\|w\|^2 - \Lambda^2\right) = 0 \end{cases}$$

~~inf~~ maximize

$$\inf L = \sum_{i=1}^{m} \frac{\alpha_i'^2}{4} + \sum_{i=1}^{m} \alpha_i' y_i - \sum_{i=1}^{m} \frac{\alpha_i'^2}{2} - \frac{1}{2\lambda}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i'\alpha_j' \Phi(x_i)^T\Phi(x_j) + \lambda\cdot\left(\frac{1}{4\lambda^2}\cdot\left\|\sum_{i=1}^{m}\alpha_i'\Phi(x_i)\right\|^2 - \Lambda^2\right)$$

$$= -\frac{1}{4}\sum_{i=1}^{m}\alpha_i'^2 + \sum_{i=1}^{m}\alpha_i' y_i - \frac{1}{4\lambda}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i'\alpha_j'\Phi(x_i)^T\Phi(x_j) - \lambda\cdot\Lambda^2$$

$\alpha_i' = 2\sqrt{\lambda}\alpha_i$

$$= -\lambda\cdot\sum_{i=1}^{m}\alpha_i^2 + 2\cdot\sum_{i=1}^{m}\alpha_i y_i - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\Phi(x_i)^T\Phi(x_j) - \lambda\Lambda^2$$

Dual: $\max_{\alpha\in\mathbb{R}^m} -\lambda\cdot\alpha^T\alpha + 2\alpha^T y - \alpha^T(X^TX)\alpha$ ☆

$$\Rightarrow \max_{\alpha} G(\alpha) = -\alpha^T(K+\lambda I)\alpha + 2\alpha^T y$$

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

$K = X^TX \in \mathbb{R}^{m\times m}$

核矩阵

$$\nabla_\alpha G = -2(K+\lambda I)\alpha + 2y = 0$$

$$\Rightarrow \alpha = (K+\lambda I)^{-1}y$$

$$\mathbb{R}^{d\times 1} \ni w = \frac{1}{2\lambda}\cdot\sum_{i=1}^{m}\alpha_i'\Phi(x_i) = \cancel{\frac{1}{\sqrt{\lambda}}}\, X\cdot\alpha = \cancel{\frac{1}{\sqrt{\lambda}}}\, X\cdot(K+\lambda I)^{-1}y$$

学习器:

$$h(x) = w\cdot\Phi(x) = \sum_{i=1}^{m}\alpha_i K(x_i, x)$$ ☆

不具有稀疏性 △

(compared to SVM)

primal: $w = (XX^T + \lambda I)^{-1} X y$

dual: $w = X(X^T X + \lambda I)^{-1} y$

$(XX^T+\lambda I)^{-1} X = X (X^T X + \lambda I)^{-1}$

⇑ 左乘、右乘

$(XX^T+\lambda I) X = X (X^T X + \lambda I)$

$w = X\alpha \Rightarrow \alpha = (K + \lambda I)^{-1} y$

	Solution	Prediction
Primal	$O(mN^2 + N^3)$ ← XX^T	$O(N)$ $\quad h(x) = \sum_{i=1}^{m} w_i x_i$
Dual	$O(\kappa \cdot m^2 + m^3)$ ← $X^T X$	$O(\kappa \cdot m)$ $\quad h(x) = \sum_{i=1}^{m} \alpha_i \cdot K(x_i, x)$

计算核函数值 m 复杂度

e.g. Gaussian/Polynomial $O(N)$

N小 m大: primal

N大 m小: dual.

} trade-off.

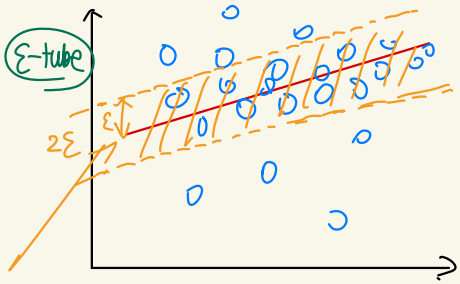
KRR: 强理论保障, 闭式解, 核方法适用自如

---

③ SVR (Support Vector Regression) 支持向量回归

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C \cdot \sum_{i=1}^{m} \xi_i$$

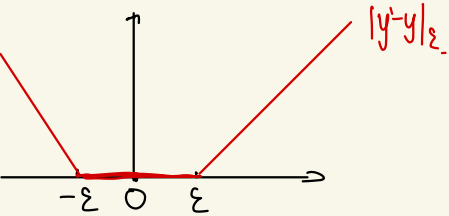
penalty (loss penalty)



$x \to \Phi(x)$: kernel method

def (ε-insensitive loss) ε-不敏感损失

$$\forall y', y \in \mathcal{Y}: \quad |y'-y|_\varepsilon = \begin{cases} 0, & |y'-y| < \varepsilon \\ |y'-y| - \varepsilon, & \text{o.w.} \end{cases}$$



$$\boxed{\begin{aligned} &\min_{w,b,\xi,\xi'} \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{m} (\xi_i + \xi_i') \\ &\text{s.t.} \quad (w^T x_i + b) - y_i \le \varepsilon + \xi_i \qquad \xi_i \ge 0 \\ &\qquad\quad y_i - (w^T x_i + b) \le \varepsilon + \xi_i' \qquad \xi_i' \ge 0 \end{aligned}}$$

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i') - \sum_{i=1}^{m}\mu_i\xi_i - \sum_{i=1}^{m}\mu_i'\xi_i' + \sum_{i=1}^{m}\alpha_i\left(w^Tx_i + b - y_i - \varepsilon - \xi_i\right) + \sum_{i=1}^{m}\alpha_i'\left(y_i - w^Tx_i - b - \varepsilon - \xi_i'\right)$$

两列约束的程度可以不同!

KKT-condition:

$$\begin{cases} \dfrac{\partial L}{\partial w} = w + \sum_{i=1}^{m}\alpha_i x_i - \sum_{i=1}^{m}\alpha_i' x_i = 0 \Rightarrow w = \sum_{i=1}^{m}(\alpha_i' - \alpha_i)\cdot x_i \\ \dfrac{\partial L}{\partial b} = \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i' = 0 \\ \dfrac{\partial L}{\partial \xi_i} = C - \mu_i - \alpha_i = 0 \\ \dfrac{\partial L}{\partial \xi_i'} = C - \mu_i' - \alpha_i' = 0 \end{cases}$$

$$\Downarrow$$

$$\inf L = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)\cdot x_i^T x_j + \sum_{i=1}^{m} y_i(\alpha_i' - \alpha_i) - \varepsilon\cdot\sum_{i=1}^{m}(\alpha_i' + \alpha_i)$$

$$\Rightarrow \text{Dual}: \boxed{\begin{aligned} &\max_{\alpha,\alpha'} \; -\varepsilon\cdot(\alpha+\alpha')^T\cdot 1 + (\alpha'-\alpha)\cdot y - \frac{1}{2}(\alpha'-\alpha)^T(X^TX)(\alpha'-\alpha) \\ &\text{s.t. } (\alpha'-\alpha)\cdot 1 = 0, \quad 0 \le \alpha \le C, \quad 0 \le \alpha' \le C' \end{aligned}}$$

Complementary Slackness:

$$\alpha_i\left(w^Tx_i + b - y_i - \varepsilon - \xi_i\right) = 0$$

$$\alpha_i'\left(y_i - w^Tx_i - b - \varepsilon - \xi_i'\right) = 0$$

$$\Downarrow$$

$$h(x) = \sum_{i=1}^{m}(\alpha_i' - \alpha_i)\, x_i^T x + b$$

$$= \sum_{i\in SV}(\alpha_i' - \alpha_i)\, x_i^T x + b$$

↳ if $\alpha_i \neq 0$ or $\alpha_i' \neq 0$: x_i is a support vector.

稀疏性!

$$\Rightarrow h(x) = \sum_{i\in SV}(\alpha_i' - \alpha_i)\cdot \underbrace{K(x_i, x)}_{\text{核函数}} + b$$

ε 大: sparsity

ε 小: accuracy

trade-off

Thm

$$\mathbb{E}_{(x,y)\sim D}\left[\,|h(x)-y|_\varepsilon\,\right] \le \frac{1}{m}\sum_{i=1}^{m}|h(x_i)-y_i|_\varepsilon + 2\sqrt{\frac{r^2\Lambda^2}{m}} + M\cdot\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

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

$$\mathbb{E}_{(x,y)\sim D}\left[\,|h(x)-y|_\varepsilon\,\right] \le \frac{1}{m}\sum_{i=1}^{m}|h(x_i)-y_i|_\varepsilon + \frac{2\Lambda\sqrt{tr(K)}}{m} + 3M\cdot\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

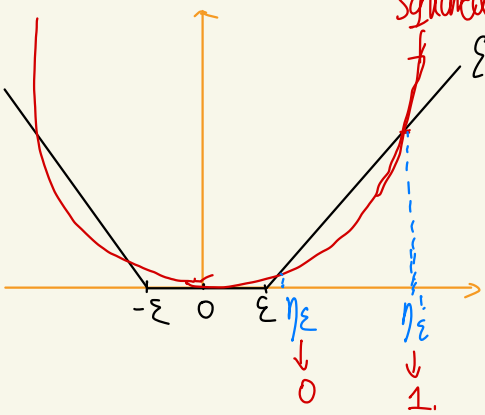
$$\left.\begin{array}{l} |h(x)-y|\le M \\ \|w\|_{\mathbb{H}}\le\Lambda \\ \sqrt{K(x,x)}=\|\Phi(x)\|_{\mathbb{H}}\le r \end{array}\right\}$$

Proof: ε-insensitive-loss:

1-Lipschitz

$\mathbb{E}_{(x,y)\sim D}\left[\,|h(x)-y|_\varepsilon\,\right]$ 和 $R(h)=\mathbb{E}_{(x,y)\sim D}\left[\,L(h(x),y)\,\right]$

$|h(x)-y|^2$. (Square Loss)



$$x-\varepsilon = x^2 \Rightarrow x=\frac{1\pm\sqrt{1-4\varepsilon}}{2}$$

若 $0<\varepsilon<\frac{1}{4}$, 在 $[-\eta'_\varepsilon,-\eta_\varepsilon]\cup[\eta_\varepsilon,\eta'_\varepsilon]$ 区间内

$$|h(x)-y|_\varepsilon \ge |h(x)-y|^2$$

$|h(x)-y|\le M = 2r\Lambda \le 1$

almost $[-1,1]$ useful generalization bound

*extensions:

- quadratic ε-insensitive: $x\mapsto \max(0,|x|-\varepsilon)^2$
- Huber: $x\mapsto x^2\mathbb{1}_{|x|\le c}+(2c|x|-c^2)\cdot\mathbb{1}_{|x|>c}$

$$\frac{1}{2}\|w\|^2 + C\cdot\sum_{i=1}^{m}(\xi_i^2+\xi_i'^2)$$

$|x|\le c$: quadratic 二次损失

$|x|>c$: linear 一次损失

* $|x|\le c$: linear

$|x|>c$: quadratic → 不希望有outlier

SVR 强理论保障 PDS核方法 稳定 (Stability) 稀疏

↓

KRR: 不稀疏

④ LASSO (Least Absolute Shrinkage and Selection Operator)

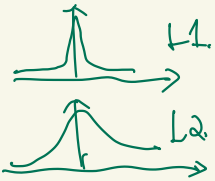
$\mathcal{H} = \{ x \mapsto w \cdot x + b ;\ w \in \mathbb{R}^N,\ b \in \mathbb{R} \}$

不好用核方法

L1-norm

$$\boxed{\min_{w,b} F(w,b) = \lambda \cdot \|w\|_1 + \sum_{i=1}^{m} (w \cdot x_i + b - y_i)^2}$$

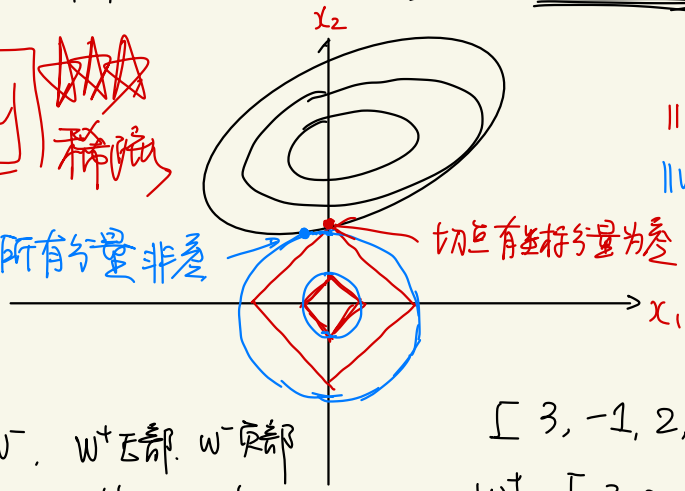
$$\Leftrightarrow \min_{w,b} \sum_{i=1}^{m} (w \cdot x_i + b - y_i)^2 \quad \text{subject to: } \|w\|_1 \le \Lambda_1$$



$\|w\|_1 \le \Lambda_1 \Leftrightarrow d_\infty$

$\|w\|_2 \le \Lambda_2 \Leftrightarrow d_2$

Sparsity (零分量更多) 稀疏



求解 Lasso:

$w = w^+ - w^-$, w^+ 左部, w^- 右部

$w^+ \ge 0$, $w^- \ge 0$

$[3, -1, 2, 7, -3, -4]$

$w^+ : [3, 0, 2, 7, 0, 0]$

$w^- : [0, 1, 0, 0, 3, 4]$

$$\Downarrow$$

$$\|w\|_1 = \sum_{j=1}^{N} w_j^+ + w_j^-$$

$$\Rightarrow \min_{w^+, w^-, b} \lambda \cdot \sum_{j=1}^{N} (w_j^+ + w_j^-) + \sum_{i=1}^{m} \left((w^+ - w^-) \cdot x_i + b - y_i \right)^2$$

QP

solution: $w_*^+, w_*^- \Rightarrow w_* = w_*^+ - w_*^-$

不可能某个分量两个都非零 (o.w. 差不变而减小)

* Proximal Gradient Descent 近端梯度下降

$\min_x f(x) + \lambda \cdot \|x\|_1$ $\quad \nabla f$ 满足 l-Lipschitz

$$f(x) \doteq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{l}{2} \cdot \|x - x_k\|^2$$

$$= \frac{l}{2} \left\| x - x_k + \frac{1}{l} \nabla f(x_k) \right\|_2^2 + \text{const}$$

$$\text{const} = f(x_k) - \frac{1}{2l} \nabla f(x_k)^T f(x_k)$$

minimized by $x = x_k - \frac{1}{l} \cdot \nabla f(x_k)$

GD: $$x_{k+1} = \underset{x}{\operatorname{argmin}} \left(\frac{l}{2} \cdot \left\| x - \underbrace{\left(x_k - \frac{1}{l} \nabla f(x_k)\right)}_{z} \right\|_2^2 + \lambda \cdot \|x\|_1 \right)$$

对 f 进行梯度下降 in 同时让 L1-norm 最小化.

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \left(\frac{l}{2} \cdot \|x - z\|_2^2 + \lambda \cdot \|x\|_1 \right)$$

$$= \sum_{i=1}^{N} \left(\frac{l}{2} \cdot (x^i - z^i)^2 + \lambda \cdot |x^i| \right)$$

$$= \sum_{i=1}^{N} g(x^i)$$

各个分量相互独立.

$$\nabla_{x^i} g(x^i) = l \cdot (x^i - z^i) + \lambda \cdot \operatorname{sgn}(x^i) = 0$$

$$\Rightarrow x^i = z^i - \frac{\lambda}{l} \cdot \operatorname{sgn}(x^i)$$

1) $z^i > \frac{\lambda}{l}$: $x^i > 0$. $\nabla_{x^i} g(x^i) = l(x^i - z^i) + \lambda = 0 \Rightarrow x^i = z^i - \frac{\lambda}{l}$ ✓

2) $z^i < -\frac{\lambda}{l}$: $x^i < 0$ $\nabla_{x^i} g(x^i) = l(x^i - z^i) - \lambda = 0 \Rightarrow x^i = z^i + \frac{\lambda}{l}$ ✓

3) $z^i \in \left[-\frac{\lambda}{l}, \frac{\lambda}{l}\right]$: $x^i > 0 \Rightarrow x^i = z^i - \frac{\lambda}{l} \le 0$ 矛盾!

$x^i < 0 \Rightarrow x^i = z^i + \frac{\lambda}{l} \ge 0$ 矛盾!

$\Rightarrow x^i = 0$

$\Rightarrow$ PGD: $x_{k+1}^i = \begin{cases} z^i - \frac{\lambda}{\ell}, & z^i > \frac{\lambda}{\ell} \\ 0, & -\frac{\lambda}{\ell} \le z^i \le \frac{\lambda}{\ell} \\ z^i + \frac{\lambda}{\ell}, & z^i < -\frac{\lambda}{\ell} \end{cases}$

Generalization Bound:

Lemma $\hat{R}_S(\mathcal{H}) \le \sqrt{\frac{2 r_\infty^2 \Lambda_1^2 \log(2N)}{m}}$.

$\|x_i\|_\infty \le r_\infty$

$\|w\|_1 \le \Lambda_1$

$$\mathcal{H} = \{ x \mapsto w \cdot x : \|w\|_1 \le \Lambda_1 \}$$

Proof: $\hat{R}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{\|w\|_1 \le \Lambda_1} \sum_{i=1}^m \sigma_i w \cdot x_i \right]$

$= \frac{\Lambda_1}{m} \mathbb{E}_\sigma \left[\max_{j \in [N]} \left| \sum_{i=1}^m \sigma_i x_{ij} \right| \right]$ 所有权重给x值最大的一个特征

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

$\le \frac{\Lambda_1}{m} \mathbb{E}_\sigma \left[\max_{j \in [N]} \max_{s \in \{-1,+1\}} s \cdot \sum_{i=1}^m \sigma_i x_{ij} \right]$

dual norm: $L_1 \Leftrightarrow L_\infty$

$= \frac{\Lambda_1}{m} \cdot \mathbb{E}_\sigma \left[\sup_{z \in A} \sum_{i=1}^m \sigma_i z_i \right]$

$A = \left\{ s \cdot \begin{pmatrix} x_{1j} \\ \vdots \\ x_{mj} \end{pmatrix} : j \in [N], s \in \{-1,+1\} \right\}$

$\forall z \in A$. $\|z\|_2^2 = \sum_{i=1}^m z_i^2 \le m \cdot \cancel{s^2} \cdot r_\infty^2$

$\Rightarrow \|z\|_2 \le r_\infty \sqrt{m}$

Massart's Lemma: $\mathbb{E}_\sigma \left[\frac{1}{m} \sum_{i=1}^m \sigma_i x_i \right] \le r \cdot \frac{\sqrt{2 \log |A|}}{m}$

2-norm bound

$|A| = 2N$.

$r = r_\infty \sqrt{m}$.

$\Rightarrow \hat{R}_S(\mathcal{H}) \le \sqrt{\frac{2 r_\infty^2 \Lambda_1^2 \log(2N)}{m}}$ $\square$.

Thm (Generalization Bound for LASSO).

$|h(x) - y| \le M$

$$R(h) \le \hat{R}_S(h) + 2 r_\infty \Lambda_1 M \sqrt{\frac{2 \log(2N)}{m}} + M^2 \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

14. Algorithmic Stability

Rademacher Complexity
Growth Function
VC-Dimension
$\Rightarrow$ algorithm-independent

e.g. $R(h) \le \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$ $\forall h \in \mathcal{H}$

Specific Algorithm: $\rightarrow$ 更紧的界? 更精细的保障?
$\hookrightarrow h_S$

algorithm-dependent

$z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ hypothesis $h: \mathcal{X} \to \mathcal{Y}' = \mathcal{Y}$ (most cases)

$L: \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$ $L_z(h) = L(h(x), y)$

泛化误差: $R(h) = \mathbb{E}_{z \sim \mathcal{D}_m}[L_z(h)]$

经验误差: $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1} L_{z_i}(h)$

Given an algorithm $\mathcal{A}$. sample set S. $\Rightarrow h_S$ < 算法 $\mathcal{A}$, 样本 S

def (uniform stability) 均匀稳定性. S 和 S' 在一个点上不同, 其余均相同

数据集

uniformly-β-stable: $\forall z \in \mathcal{Z}$: $\boxed{|L_z(h_S) - L_z(h_{S'})| \le \beta.}$ stability coefficient.

数据稍微小扰动, 损失变化不大!

Assume bound-loss-function $\forall h.z.$ $L_z(h) \le M$

通常 $\beta = O(\frac{1}{m})$ ...

Thm (Generalization Guarantee) A: β-stable learning algorithm $S = ((x_1, y_1), \cdots, (x_m, y_m))$

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

$$\boxed{R(h_S) \le \hat{R}_S(h_S) + \beta + (2m\beta + M)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}}$$

(Rademacher Complexity: $R(h) \le \hat{R}_S(h) + R_m(\mathcal{H}) + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$) $\forall h \in \mathcal{H}$

$\forall g \in [0,1]$ $\mathbb{E}[g(z)] \le \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2R_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$

Proof: $\Phi(S) = R(h_S) - \hat{R}_S(h_S)$

$S = (z_1, \cdots, z_{m-1}, z_m)$

$S' = (z_1, \cdots, z_{m-1}, z_m')$

$$|\Phi(S') - \Phi(S)|$$

$$= |R(h_{S'}) - \hat{R}_{S'}(h_{S'}) - R(h_S) + \hat{R}_S(h_S)|$$

$$\le \underbrace{|R(h_{S'}) - R(h_S)|}_{\le M} + \underbrace{|\hat{R}_{S'}(h_{S'}) - \hat{R}_S(h_S)|}_{\le M}$$

~~Chapter 3. $= 0$~~ ~~$\le M$~~

h_S 与 S 有关的!

$$|R(h_{S'}) - R(h_S)| = |\mathbb{E}_z[L_z(h_{S'})] - \mathbb{E}_z[L_z(h_S)]| \le \mathbb{E}_z\left[|L_z(h_{S'}) - L_z(h_S)|\right] \le \beta$$

$$|\hat{R}_{S'}(h_{S'}) - \hat{R}_S(h_S)| = \frac{1}{m}\left|\sum_{i=1}^{m} L_{z_i'}(h_{S'}) - \sum_{i=1}^{m} L_{z_i}(h_S)\right|$$

$$= \frac{1}{m}\left|\sum_{i=1}^{m-1}\left(L_{z_i}(h_{S'}) - L_{z_i}(h_S)\right) + \underbrace{L_{z_m'}(h_{S'}) - L_{z_m}(h_S)}_{\text{样本不同}}\right|$$

$$\le \frac{m-1}{m}\cdot\beta + \frac{M}{m} \le \beta + \frac{M}{m}$$

$$\Rightarrow |\Phi(S') - \Phi(S)| \le 2\beta + \frac{M}{m} \ll 2M$$

by McDiarmid Inequality: $\Pr\left(\Phi(S) - \mathbb{E}_S[\Phi(S)] \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{m\left(2\beta+\frac{M}{m}\right)^2}\right)$

$$= \exp\left(-\frac{2m\varepsilon^2}{(2m\beta+M)^2}\right) = \delta$$

$$\Rightarrow \Phi(S) \leq \mathbb{E}_S[\Phi(S)] + (2m\beta+M)\cdot\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \quad \text{w.p.} \geq 1-\delta$$

引入稳定性后, Lipschitz常数更小!

只需证: $\mathbb{E}_S[\Phi(S)] \leq \beta$

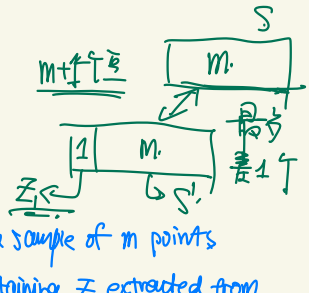
$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S[R(h_S)] - \mathbb{E}_S[\hat{R}_S(h_S)]$$

$$= \mathbb{E}_{S\sim D^m}\left[\mathbb{E}_{Z\sim D}[L_Z(h_S)]\right] - \mathbb{E}_S\left[\frac{1}{m}\sum_{i=1}^{m} L_{Z_i}(h_S)\right]$$

i.i.d

$$= \mathbb{E}_{S,Z\sim D^{m+1}}[L_Z(h_S)] - \mathbb{E}_{S\sim D^m}[L_{Z_1}(h_S)] \quad Z_1\sim D$$

m+1个点



$$= \mathbb{E}_{S,Z\sim D^{m+1}}[L_Z(h_S)] - \mathbb{E}_{S,Z\sim D^{m+1}}[L_Z(h_{S'})]$$

$$= \mathbb{E}_{S,Z\sim D^{m+1}}\left[L_Z(h_S) - L_Z(h_{S'})\right]$$

$$\leq \mathbb{E}_{S,Z}[\beta] = \beta.$$

S' is a sample of m points containing Z extracted from the m+1 points formed by S, Z.

$$\Rightarrow R(h_S) \leq \hat{R}_S(h_S) + \beta + (2m\beta+M)\cdot\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \qquad \square.$$

h_S < Algorithm A / Sample S

algorithmic-dependent!

convergence: 收敛

$$\frac{m\beta}{\sqrt{m}} = O(1) \Rightarrow \boxed{\beta = O\left(\frac{1}{\sqrt{m}}\right)}.$$

particular: $\beta = O\left(\frac{1}{m}\right) \Rightarrow R(h_S) - \hat{R}_S(h_S) = O\left(\frac{1}{\sqrt{m}}\right)$ w.h.p.

Kernel-based Regularization Algorithm. $\beta = O(\frac{1}{m})$

$\mathbb{K}$: PDS kernel.

$\mathbb{H}$: RKHS $\rightarrow$ norm: $\|x\|_{\mathbb{K}} = \sqrt{K(x,x)}$ $\qquad S = (z_1, \cdots, z_m)$

Objective: $\boxed{\text{minimize} \quad F_S(h) = \underbrace{\hat{R}_S(h)}_{\text{经验风险}} + \underbrace{\lambda \cdot \|h\|_{\mathbb{K}}^2}_{\text{正则化项}}}$ ☆

def (σ-admissibility) $L: \mathcal{Y}' \times \mathcal{Y} \rightarrow \mathbb{R}_+$

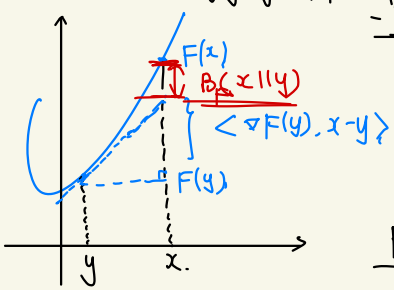
$\forall h, h' \in \mathbb{H}, \ \forall (x,y) \in \mathcal{X} \times \mathcal{Y}, \quad |L(h'(x), y) - L(h(x), y)| \leq \sigma |h'(x) - h(x)|$

↳ Lipschitz. for 1st argument.

def (Bregman Divergence) $F: \mathbb{H} \rightarrow \mathbb{R}$. convex differentiable.

$\forall f, g \in \mathbb{H}, \quad B_F(f \| g) = F(f) - F(g) - \langle f - g, \nabla F(g) \rangle \geq 0$

(convexity)



F 不一定可微! $\nabla F(g)$ gradient. ↓ sub-gradient. 次梯度

F convex!

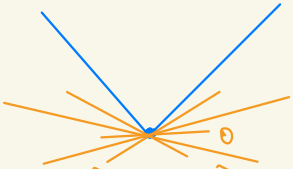
(sub-differential) 次微分 $\partial F(h) = \{ g \in \mathbb{H} : \forall h' \in \mathbb{H}, \ F(h') - F(h) \geq \langle h' - h, g \rangle \}$

(sub-gradient) 次梯度 $\delta F(h) \in \partial F(h)$

① 若F可微 at h, $\partial F(h) = \{\nabla F(h)\}$

② 若F在h处取到最小值 $0 \in \partial F(h)$

③ 可加性: $\partial(F_1 + F_2)(h) = \{ g_1 + g_2 : g_1 \in \partial F_1(h), \ g_2 \in \partial F_2(h) \}$



直线在曲线下方即可

def (generalized Bregman Divergence) F convex

$\forall h, h' \in \mathbb{H}$. $B_F(h' \| h) = F(h') - F(h) - \langle h' - h, \delta F(h) \rangle$

minimize $\boxed{F_S(h) = \hat{R}_S(h) + \lambda \cdot \|h\|_K^2}$ $\rightarrow N(h)$

norm is differential: $\delta \|h\|_K^2 = \nabla \|h\|_K^2$

$$B_{F_S}(h' \| h) = B_{\hat{R}_S}(h' \| h) + \lambda \cdot B_N(h' \| h)$$

$$\Rightarrow \delta \hat{R}_S(h) = \delta F_S(h) - \lambda \cdot \nabla \|h\|_K^2$$

h: minimizer of $F_S(h) \Rightarrow \delta F_S(h) = 0$

o.w.: 随便取即可!

$\beta = O\left(\frac{1}{\sqrt{m}}\right)$

β界与λ选取有关

$\lambda = \omega\left(\frac{1}{\sqrt{m}}\right)$

$\Downarrow$

Simpler

$\Downarrow$

Occam Razor

Thm (β-bound for kernalized regularization algorithms)

☆☆ $\boxed{\beta \leq \frac{\sigma^2 r^2}{m \cdot \lambda}}$ $\Rightarrow \beta = O\left(\frac{1}{m}\right)$

σ: convex $\perp$ σ-admissible

r: $K(x,x) \leq r^2$

$\Rightarrow \|\Phi(x)\|_K \leq r$.

M: $L(h(x), y) \leq M$.

Proof: $S = (z_1, \cdots, z_m)$ minimizer of F_S: h_S

$S' = (z_1, \cdots, z_{m-1}, z_m')$: minimizer of $F_{S'}$: $h_{S'}$

$B_{F_S} = \underbrace{B_{\hat{R}_S}}_{\geq 0} + \lambda \cdot B_N$ $\quad B_{F_{S'}} = \underbrace{B_{\hat{R}_{S'}}}_{\geq 0} + \lambda \cdot B_N$

$$\Rightarrow B_{F_S}(h_{S'} \| h_S) + B_{F_{S'}}(h_S \| h_{S'}) \geq \lambda \cdot \left(B_N(h_{S'} \| h_S) + B_N(h_S \| h_{S'}) \right) \geq 2\lambda \cdot \|h_{S'} - h_S\|_K^2$$

$B_F(x \| y) = F(x) - F(y) - \langle \nabla F(y), x - y \rangle$ $\quad \nabla \|h\|_K^2$

$$B_N(h_{S'} \| h_S) + B_N(h_S \| h_{S'}) = -\langle h_S' - h_S, 2h_S \rangle - \langle h_S - h_S', 2h_S' \rangle = 2 \cdot \|h_{S'} - h_S\|_K^2$$

$$F_S(h_S') - F_S(h_S) - \langle h_S' - h_S, \underbrace{\delta F_S(h_S)}_{\to 0} \rangle + F_{S'}(h_S) - F_{S'}(h_{S'}) - \langle h_S - h_S', \underbrace{\delta F_{S'}(h_{S'})}_{\to 0} \rangle$$

$$= F_S(h_S') - F_S(h_S) + F_{S'}(h_S) - F_{S'}(h_{S'})$$

$F_S(h) = \hat{R}_S(h) + \lambda \cdot \|h\|_K^2$

$$= \hat{R}_S(h_S') + \lambda\|h_S'\|_K^2 - \hat{R}_S(h_S) - \lambda\|h_S\|_K^2 + \hat{R}_{S'}(h_S) + \lambda\|h_S\|_K^2 - \hat{R}_{S'}(h_S') - \lambda\|h_S'\|_K^2$$

$$= \hat{R}_S(h_S') - \hat{R}_S(h_S) + \hat{R}_{S'}(h_S) - \hat{R}_{S'}(h_S') \geq 2\lambda\|h_S' - h_S\|_K^2 = 2\lambda\|\Delta h\|_K^2$$

$$\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} L_{z_i}(h)$$

$$2\lambda\|h_S' - h_S\|_K^2 \leq \frac{1}{m}\left(L_{z_m'}(h_S) - L_{z_m}(h_S) + L_{z_m}(h_S') - L_{z_m'}(h_S')\right)$$

$$\leq \frac{\sigma}{m}\left(|h_S(x_m) - h_S'(x_m)| + |h_S(x_m') - h_S'(x_m')|\right)$$

$$= \frac{\sigma}{m}\left(|\Delta h(x_m)| + |\Delta h(x_m')|\right)$$

σ-admissibility (Lipschitz)

By Reproducing Kernel Property.

$$\Delta h(x) = \langle \Delta h, K(x,\cdot)\rangle \leq \|\Delta h\|_K \cdot \|K(x,\cdot)\|_K = \sqrt{K(x,x)}\cdot\|\Delta h\|_K \leq r\cdot\|\Delta h\|_K$$

Cauchy-Schwarz

bounded

$$\Rightarrow 2\lambda\|\Delta h\|_K^2 \leq \frac{\sigma}{m}\cdot 2\cdot r\cdot\|\Delta h\|_K \Rightarrow \|\Delta h\|_K \leq \frac{\sigma\cdot r}{m\cdot\lambda}$$

不同的算法不同

$$\beta = |L_z(h_S') - L_z(h_S)| \leq \sigma\cdot|\Delta h(x)| \leq \sigma\cdot r\cdot\|\Delta h\|_K = \frac{\sigma^2 r^2}{m\lambda} \quad \square$$

核方法优化 $F_S(h) = \hat{R}_S(h) + \lambda\|h\|_K^2 \rightarrow$ 稳定性系数 $\beta = O\left(\frac{1}{m}\right)$ (for fixed λ)

e.g.1. SVR 支持向量回归 : ε-insensitive loss

$$L_\varepsilon(y', y) = \begin{cases} 0 & \text{if } |y'-y| < \varepsilon \\ |y'-y| - \varepsilon & \text{o.w.} \end{cases}$$

$$\sigma = 1 \Rightarrow R(h_S) \leq \hat{R}_S(h_S) + \frac{r^2}{m\lambda} + \left(\frac{2r^2}{\lambda} + M\right)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

tighter bound by stability

$$* \quad \mathbb{E}_{(x,y)\sim D}\left[|h(x) - y|_\varepsilon\right] \leq \frac{1}{m}\sum_{i=1}^{m}|h(x_i) - y_i|_\varepsilon + 2\sqrt{\frac{r^2\Lambda^2}{m}} + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$\begin{cases} \|w\|_K \leq \Lambda \\ K(x,x) \leq r \end{cases}$$

e.g.2. KRR 岭回归(核化) $L_2(y', y) = (y'-y)^2 \le M$

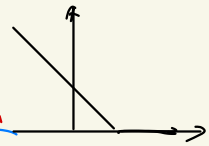
$$\left| L_2(h'(x), y) - L_2(h(x), y) \right| = \left| (h'(x)-y)^2 - (h(x)-y)^2 \right|$$
$$= \left| \left(h'(x)-h(x)\right) \cdot \left(h'(x)-y+h(x)-y \right) \right|$$
$$\le \left|h'(x)-h(x)\right| \cdot \left| h'(x)-y+h(x)-y \right| \le 2\sqrt{M} \cdot \left|h'(x)-h(x)\right|$$

$$\Rightarrow \sigma = 2\sqrt{M} \quad \Rightarrow R(h_S) \le \hat{R}_S(\hat{h}_S) + \frac{4M r^2}{\lambda m} + \left(\frac{8M r^2}{\lambda} + M\right) \cdot \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

* Linear Regression: $\lambda = 0$. $\Rightarrow$ uninformative! unstable.

e.g.3. SVM 支持向量机 $y \in \mathcal{Y} = \{-1, +1\}$

hinge-loss: $L_{hinge}(y', y) = \begin{cases} 0, & \text{if } yy' \ge 1 \\ 1 - yy', & 0.w. \end{cases}$ $= 1 - y \cdot h_S(x)$



无论有多
且非常有信心

$\sigma = 1$

SVM stability bound

$$\boxed{R(h_S) \le \hat{R}_S(h_S) + \frac{r^2}{m\lambda} + \left(\frac{2r^2}{\lambda} + 1\right) \cdot \sqrt{\frac{\log \frac{1}{\delta}}{2m}}}$$

$\frac{r}{\sqrt{\lambda}} + 1$

$K(x,x) \le r^2$

$L \le M$

Lemma bounded-label: $\|y\| = \sqrt{B}$ ($\Leftrightarrow L(0, y) \le B$)

$\Rightarrow \forall x \in \mathcal{X}$: $|h_S(x)| \le r \cdot \sqrt{B/\lambda}$. (minimization hypothesis bound)

Proof: Reproducing Property, $\forall x \in \mathcal{X}$. $|h_S(x)| = \langle h_S, K(x, \cdot) \rangle \le \|h_S\|_K \cdot \sqrt{K(x,x)} \le r \cdot \|h_S\|_K$.

$0 \in \mathbb{H}$. minimization $F_S(h_S) \le F_S(0) = \frac{1}{m} \sum_{i=1}^{m} L(0, y_i) \le B$

$F_S(h_S) = \underbrace{\hat{R}_S(h_S)}_{\ge 0} + \lambda \cdot \|h_S\|_K^2 \Rightarrow \lambda \cdot \|h_S\|_K^2 \le F_S(h_S) \le B \Rightarrow |h_S(x)| \le r \sqrt{\frac{B}{\lambda}}$ □

$|L_{hinge}(0, y)| \le 1. \Rightarrow |h_S(x)| \le \frac{r}{\sqrt{\lambda}} \Rightarrow L_{hinge}(h_S(x), y) \le 1 + \frac{r}{\sqrt{\lambda}} = M$

$y' = -1$.