Summary 机器学习理论基础

What is machine learning?

离不开经验

experience

data

estimation

inference

prediction ...

Machine Learning

no data → probability

\+

prior knowledge

limited data

m有限

infinite data

performance

泛化误差 $R(h) = \mathbb{E}_{x\sim D}[L(h(x), y)]$

$h \in \mathcal{H}$ ← 假设空间

分类: 0/1

回归: $(y-y')^2$

sampling

"概率"

probablistic / stochastic / agnostic

⇓

PAC (Valient 1984)

$$\Pr(R(h) \le \varepsilon) \ge 1-\delta$$

误差(风险)

置信度

D: 未知 (Bayesian)

$S = \{(x_i, y_i)\}_{i=1}^{m}$

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

经验风险

Law of Large Number: $\hat{R}_S(h) \to R(h)$ $m \to \infty$

$\mathbb{E}_{S\sim D^m}[\hat{R}_S(h)] = R(h)$

有限 m → 一般 generalization 泛化

deduction v.s. induction (inductive reasonable)

exactly "probably"

Hume's problem of induction

Assumption: uniformity principle

↳ test . train i.i.d

minimize $\hat{R}_S(h)$ ⟹ ERM

$$h_S^{ERM} = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \boxed{\hat{R}_S(h)}$$

$\mathcal{H}$

① m至少为多少时，$\hat{R}_S(h)$和$R(h)$相差并不会很大 → 样本复杂性

$m = poly(\frac{1}{\varepsilon}, \frac{1}{\delta}, n, size(c))$

② $\hat{R}_S(h)$和$R(h)$相差不大而概率至少为多少？

⟺ 给定概率，$\hat{R}_S(h)$和$R(h)$会相差多少？

$\boxed{R(h) - \hat{R}_S(h)} \leq$ //// 泛化界

generalization gap (optimism)

Generalization Bound

有限假设空间

$|\mathcal{H}|$

↓

$c(\mathcal{H})$ 复杂度

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

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

Occam's Razor

简单为王！

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

拟合噪声

Growth Function $\Pi_{\mathcal{H}}(m) = \max_S \left|\{(h(x_1), \cdots, h(x_m)) : h\in\mathcal{H}\}\right|$

VC-Dimension: $VCdim(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$

Rademacher Bound: $\forall g \in G$. $g(x) \in [0,1]$

$$\boxed{\mathbb{E}[g(z)] \le \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(G) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}}$$

$$\Rightarrow \quad R(h) \le \hat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

VC dim: $R(h) \le \hat{R}_S(h) + \sqrt{\frac{2d\log\frac{em}{d}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$

$\hookrightarrow$ 大：界松；小：集合简单 } trade-off

d → 大：界松；→ 小：集合简单

Covering Number $\mathcal{X} \to \mathcal{Y} \in \mathbb{R} \to L(h(x), y) = (h(x)-y)^2 \le M^2$

$\mathcal{H}$ bounded

$k = N(\mathcal{H}, \varepsilon)$：最少能用多少个半径为 ε 的球覆盖 $\mathcal{H}$.

$\exists \{h_1, \cdots, h_k\} \subseteq \mathcal{H}$. s.t. $\forall h \in \mathcal{H}$, $\exists\, i \le k$ s.t.

$$\|h - h_i\|_\infty = \max_{x \in \mathcal{X}} |h(x) - h_i(x)| \le \varepsilon.$$

$\mathcal{H}$: compact $\to N(\mathcal{H}, \varepsilon)$ finite

$$\Pr\left(\sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| \ge \varepsilon\right) \le N\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) \cdot 2\exp\left(-\frac{m\varepsilon^2}{2M^4}\right)$$

Proof: ① Optimism $L_S = R(h) - \hat{R}_S(h)$

$\forall h_1, h_2 \in \mathcal{H}$, $\forall S$ $\quad \boxed{|L_S(h_1) - L_S(h_2)| \le 4M \cdot \|h_1 - h_2\|_\infty}$

$$|L_S(h_1) - L_S(h_2)| \le |R(h_1) - R(h_2)| + |\hat{R}_S(h_1) - \hat{R}_S(h_2)|$$

$\boxed{|h(x) - y| \le M}$

$$|R(h_1) - R(h_2)| = \left|\mathbb{E}_{x\sim D}\left[(h_1(x)-y)^2\right] - \mathbb{E}_{x\sim D}\left[(h_2(x)-y)^2\right]\right|$$

$$= \mathbb{E}_{x\sim D}\left[(h_1(x) - h_2(x))(h_1(x) + h_2(x) - 2y)\right]$$

$$= \sum_{(x,y)\sim \mathcal{X}\times\mathcal{Y}} \Pr(x,y)\cdot\left(h_1(x)-h_2(x)\right)\left(h_1(x)+h_2(x)-2y\right)$$

$$\le \sum_{(x,y)\sim \mathcal{X}\times\mathcal{Y}} \Pr(x,y)\cdot \|h_1-h_2\|_\infty\cdot 2M = 2M\cdot\|h_1-h_2\|_\infty$$

$\{h_1,\cdots,h_K\}$

$\frac{\varepsilon}{8M}$

(2) $\mathcal{H} = \bigcup_{i=1}^{K} B_i$

$$\Pr_{S\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}}|L_S(h)|\ge\varepsilon\right) \le \sum_{i=1}^{K}\Pr\left(\sup_{h\in B_i}|L_S(h)|\ge\varepsilon\right)$$

$$\le \Pr_{S\sim\mathcal{D}^m}\left(\bigcup_{i=1}^{K}\sup_{h\in B_i}|L_S(h)|\ge\varepsilon\right)$$

(3) $\Pr_{S\sim\mathcal{D}^m}\left(\sup_{h\in B_i}|L_S(h)|\ge\varepsilon\right)$

$$\le \Pr_{S\sim\mathcal{D}^m}\left[|L_S(h_i)|\ge\frac{\varepsilon}{2}\right] \quad \forall i\in[K]$$

$$\Longleftrightarrow \sup_{h\in B_i}|L_S(h)|\ge\varepsilon \Rightarrow |L_S(h_i)|\ge\frac{\varepsilon}{2}$$

$$\Longleftrightarrow L_S(h_i)<\frac{\varepsilon}{2} \Rightarrow \sup_{h\in B_i}|L_S(h)|<\varepsilon$$

$$|L_S(h)| \le |L_S(h)-L_S(h_i)| + |L_S(h_i)| \le 4M\cdot\|h-h_i\|_\infty + \frac{\varepsilon}{2}$$

$$\le 4M\cdot\frac{\varepsilon}{8M}+\frac{\varepsilon}{2} < \varepsilon$$

(4) $$\Pr_{S\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}}|L_S(h)|\ge\varepsilon\right) \le \sum_{i=1}^{K}\Pr\left(|L_S(h_i)|\ge\frac{\varepsilon}{2}\right)$$

$R(h_i)-\hat{R}_S(h_i)$

$$\le \sum_{i=1}^{K}\Pr\left(\left|\sum_{i=1}^{m}\frac{(h(x_i)-y_i)^2}{m} - \mathbb{E}\left[\frac{(h(x_i)-y_i)^2}{m}\right]\right|\ge\frac{\varepsilon}{2}\right)$$

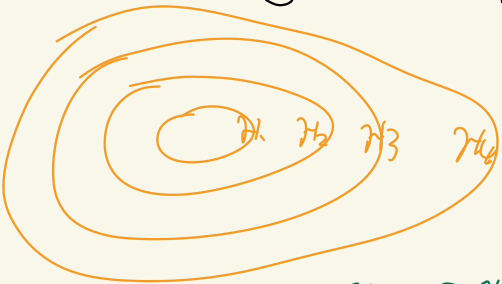
$\left[0,\frac{M^2}{m}\right]$

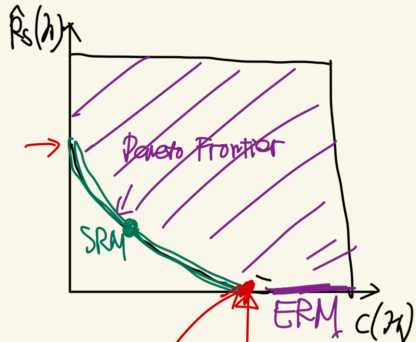
$$\le K\cdot 2\exp\left(-\frac{m\varepsilon^2}{2M^4}\right)$$

Covering number: $N\left(\mathcal{H},\frac{\varepsilon}{8M}\right)$. $\square$

$\mathcal{H}$ 复杂 $\Rightarrow$ SRM (Structural Risk Minimization)

$$h_S^{SRM} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \underbrace{\hat{R}_S(h)}_{\text{ERM}} + \underbrace{C(\mathcal{H})}_{\text{复杂性}}$$





统计学习

结构风险最小化

$$\boxed{h^* = \operatorname{argmin} \underbrace{\hat{R}_S(h)}_{\text{经验风险}} + \underbrace{\lambda \cdot C(h)}_{\text{正则化项}}}$$

trade-off

MDL (Minimal Description Length)

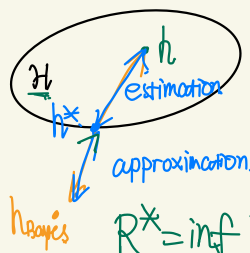
Soft-Margin SVM

L1-Regularized AdaBoost

Ridge Regression

LASSO - -

SRM:



$$R^* = \inf_h R(h)$$

Agnostic PAC $\leftarrow$ stochastic learning scenario

$\downarrow$

bias-variance tradeoff.

$$R = \underbrace{\text{bias}}_{\substack{\downarrow \\ \text{underfitting}}} + \underbrace{\text{variance}}_{\substack{\downarrow \\ \text{overfitting}}} + \text{noise}$$

Distribution P → Data D

Probability

Distribution P fully specified

What can we say about $X\sim P$?

Deductive

Statistics

Inductive

Observe data X from unknown distribution P (i.i.d)

What can we conclude about P?

↓

estimation

经验误差 $\hat{R}_S(h)=\frac{1}{m}\sum_{i=1}^{m}\boxed{\mathbb{1}_{h(x_i)\neq y_i}}$

泛化误差 $R(h)=\mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}_{h(x_i)\neq y_i}\right]$

Generalization bound

$R(h)-\hat{R}_S(h)$ → $\hat{R}_{S,\Phi}(h)$

optimism

ERM: $h_S^{ERM}=\underset{h\in H}{\arg\min}\ \hat{R}_S(h)$

⇓

SRM: $h_S^{SRM}=\boxed{\underset{h\in H}{\arg\min}\ \hat{R}_S(h)+\lambda\cdot C(\mathcal{H})}$ ★

统计机器学习

经验风险 　　正则化项 ← 复杂性 $\begin{cases}\text{Rademacher Complexity}\\ \text{Growth Function}\\ \text{VC-dimension}\\ \text{Covering Number}\end{cases}$ Sauer's Lemma

0/1-loss 　　$\|h\|$

↓

替代损失 surrogate loss $\begin{cases}\text{convex}\\ \text{upper-bound of 0/1-loss}\end{cases}$

Φ-loss

核心：支持向量机(SVM)

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

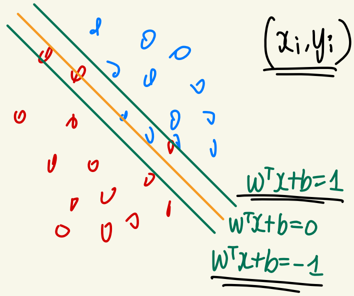
$$\text{s.t.}\quad y_i(w^Tx_i+b)\geq 1-\xi_i$$

$$\xi_i\geq 0$$

$\Longleftrightarrow$

$$\boxed{\begin{aligned}&\max_{\alpha}\ \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j\\ &\text{s.t.}\ \sum_{i=1}^{m}\alpha_i y_i=0\\ &\qquad 0\leq\alpha_i\leq C\end{aligned}}$$

Maximum-Margin



(x_i, y_i) $\quad y_i(w^T x_i + b)$

$\begin{cases} <0: & \text{犯错} \\ (0,1): & \text{正确 confidence} \downarrow \\ >1: & \text{正确 confidence} \uparrow \end{cases}$

confidence margin

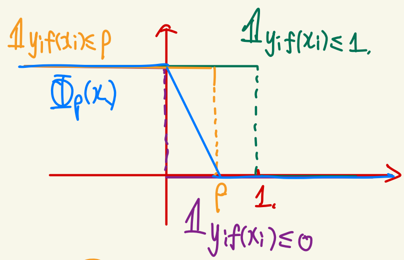
$\Updownarrow$

geometric margin $\rho = \dfrac{|y \cdot h(x)|}{\|w\|}$

margin

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

① $\mathbb{1}_{y_i f(x_i) \le 0} \le \Phi_\rho(x) \le \mathbb{1}_{y_i f(x_i) \le \rho}$



$\Downarrow$ Margin Bound for Binary Classification.

② lipschitz

$$R(h) \le \hat{R}_{S,\rho}(h) + \frac{2}{\rho} \cdot \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \quad \text{w.p.} \ge 1-\delta$$

Hypothesis Space $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$

$\downarrow$ Φ $\boxed{l\text{-lipschitz}}$

Loss Space. $\mathcal{F} = \{f_h(z) = \Phi(h, z),\ h \in \mathcal{H}\}$

$\mathcal{F} = \Phi \circ \mathcal{H}$

Talagrand Contraction Lemma $\mathfrak{R}_m(\Phi \circ \mathcal{H}) \le l \cdot \mathfrak{R}_m(\mathcal{H})$

Bounded

$\|x\|_2 \le r$.

$\|w\|_2 \le \Lambda$

$\mathfrak{R}_m(\mathcal{H}) \le \sqrt{\dfrac{r^2 \Lambda^2}{m}}$

$\max(1 - y h(x_i), 0) = \xi_i$

Margin Theory

$\downarrow$

Kernel Method

PDS kernel $\Longleftrightarrow$ K SPSD

低维 $\xrightarrow{\Phi}$ 高维 $\mathbb{H}$ (本) K ✓ → { polynomial, Gaussian (RBF), Laplace, Sigmoid }

feature map   RKHS

表示定理 ⇓

$$\arg\min_{h\in\mathbb{H}} F(h) = \arg\min_{h\in\mathbb{H}} G(\|h\|_{\mathbb{H}}) + L(h(x_1),\dots,h(x_m))$$

不减→增

$$\Rightarrow h^* = \sum_{i=1}^{m} \alpha_i K(x_i, \cdot)$$

$\|\Phi\|_{\mathbb{H}} \le r \iff K(x_i, x_i) \le r^2.$

$$\mathfrak{R}_m(\mathcal{H}) \le \frac{r\Lambda}{m} \Rightarrow \mathfrak{R}_m(\mathcal{H}) \le \frac{\Lambda\sqrt{\mathrm{tr}(K)}}{m}$$

$$R(h) \le \hat{R}_{S,\rho}(h) + 2\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \quad \text{w.p.} \ge 1-\delta$$

$\boxed{\text{SVM + Kernel Method}}$ 核技巧

因地制宜

→ 多分类 (Multi-class classification)

$y = \{1, \dots, k\}$ $1_{h(x)\ne y} \Rightarrow d_H(h(x), y)$

$\rho_h(x,y) = h(x,y) - \max_{y'\ne y} h(x,y')$

Multi-Class SVM, Decision Tree → impurity

OVO. OVR. MVM.

→ 回归 (Regression)

$L(y,y') = |y'-y|^2$ $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m}(h(x_i)-y_i)^2$ → MSE

Generalization Bound

Rademacher Complexity

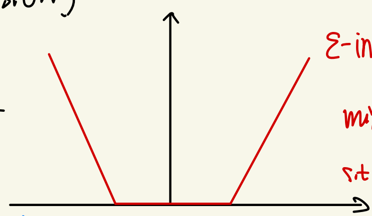
VC-dimension

↓

pseudo-dimension

fat-dimension (数据相关)

fitting



ε-insensitive loss

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \hat{\xi}_i)$$

$$\text{s.t. } (w^T x_i + b) - y_i \le \varepsilon + \xi_i, \quad \xi_i \ge 0$$

$$y_i - (w^T x_i + b) \le \varepsilon + \hat{\xi}_i, \quad \hat{\xi}_i \ge 0$$

Linear Regression

→ L2: Ridge Regression

→ L1: LASSO

Ranking

$$S = \{(x_i, x_i', y)\}_{i=1}^{m} \in \mathcal{X} \times \mathcal{X} \times \{-1, +1\}$$

$\rightarrow$ preference $\{-1, 0, +1\}$

$\downarrow$

scoring

$h \in H$: $h(x_i) > h(x_i')$

$h(x_i) < h(x_i')$

confidence margin: $y_i\big(h(x_i') - h(x_i)\big) \le 0$

$x, x' : x$

$x', x'' : x'$

$x, x'' : x''$

\$100 \$0.1

antisymmetric

transductive

$$\Rightarrow R(h) = \mathbb{E}_{(x,x') \sim D}\left[\mathbb{1}_{y_i(h(x_i') - h(x_i)) \le 0} \right]$$

$$\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{(y_i(h(x_i') - h(x_i)) \le 0 \,\wedge\, (y_i \ne 0))}$$

$$\widehat{R}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho\Big(y_i\big(h(x_i') - h(x_i)\big)\Big)$$

Thm. (Margin Bound for ranking)

$$R(h) \le \widehat{R}_{S,\rho}(h) + \frac{2}{\rho}\left(\mathfrak{R}_m^{D_1}(\mathcal{H}) + \mathfrak{R}_m^{D_2}(\mathcal{H})\right) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

fixed ρ

Proof:

(x, x', y)

$\downarrow \quad \downarrow$

marginal: $D_1 \quad D_2$

$$\mathbb{E}\Big[\Phi_\rho\big(y(h(x') - h(x))\big)\Big] \le \widehat{R}_{S,\rho}(h) + 2\mathfrak{R}_m(\Phi_\rho \circ \mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$\text{||}\wedge$

$R(h)$

$\mathcal{F}$: loss-space ($\widetilde{\mathcal{H}}$)

$\mathcal{H}$: hypothesis-space

Talagrand's Lemma: $\mathfrak{R}_m(\Phi_\rho \circ \mathcal{F}) \le \frac{1}{\rho}\mathfrak{R}_m(\mathcal{F})$

$$\mathfrak{R}_m(\mathcal{F}) = \frac{1}{m}\mathbb{E}_{S,\sigma}\left[\sup_{h \in H}\sum_{i=1}^{m} \sigma_i y_i\big(h(x_i') - h(x_i)\big)\right]$$

$\uparrow$ Rademacher Variable

$$\le \frac{1}{m}\mathbb{E}_{S,\sigma}\left[\sup_{h \in H}\sum_{i=1}^{m} \sigma_i h(x_i') + \sup_{h \in H}\sum_{i=1}^{m} \sigma_i h(x_i)\right]$$

$$= \mathfrak{R}_m^{D_2}(\mathcal{H}) + \mathfrak{R}_m^{D_1}(\mathcal{H})$$

$$\Rightarrow R(h) \le \hat{R}_{S,\rho}(h) + 2\left(\hat{\mathfrak{R}}_m^{D_1}(\mathcal{H}) + \hat{\mathfrak{R}}_m^{D_2}(\mathcal{H})\right) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \quad \square$$

$\hookrightarrow \forall \rho \quad + \sqrt{(\log\log_2 \frac{2}{\rho})/m}$.

Kernel-based $\mathfrak{R}_m(\mathcal{H}) \le \sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} \Rightarrow R(h) \le \hat{R}_{S,\rho}(h) + 4\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$.

SVM with Ranking: $y_i\, w\left(\Phi(x_i') - \Phi(x_i)\right) \ge 1 - \xi_i$

$$\Rightarrow \xi_i = \max\left(1 - y_i\left(w\cdot\left(\Phi(x_i') - \Phi(x_i)\right)\right), 0\right)$$

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}\xi_i$$

$$\text{s.t. } y_i\left(w\left(\Phi(x_i') - \Phi(x_i)\right)\right) \ge 1 - \xi_i$$

$$\xi_i \ge 0. \qquad \Phi(x_i') - \Phi(x_i) = \Psi(x_i', x_i) \qquad \forall i \in [M]$$

$w \cdot \Psi(x_i', x_i)$

③ Margin Bound

与 N 无关的 Bound $\Rightarrow$ Margin Maximization

AdaBoost ⟨ L1-margin 坐标下降 → Winnow

L2-margin 梯度下降 → Perceptron.

$\downarrow$

$$\gamma = \frac{\rho_{max}}{2},$$

$\uparrow$

edge

Online Learning

Regret Bound 遗憾系.

$$R_T = \sum_{t=1}^{T} L(\hat{y}_t, y_t) - \min_{i=1}^{N}\sum_{t=1}^{T} L(\hat{y}_{t,i}, y_t)$$

Stability $|L_z(h_S) - L_z(h_{S'})| \le \beta$.

Statistical Learning Theory ☆☆

统计学习理论

分析问题

↓

损失函数

↓

泛化界

Bound

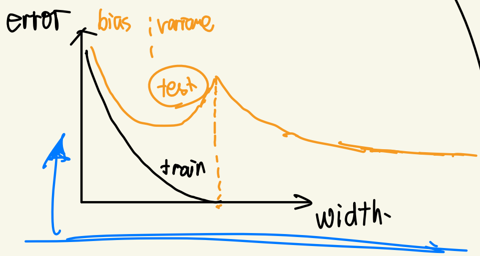
↓

复杂性

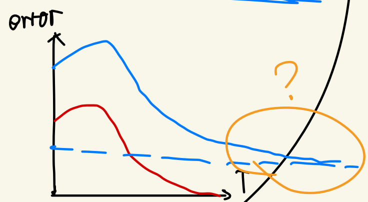
↓

算法（优化问题）

Double Descent



AdaBoost:



模型

1. 神经网络

Neural Network Learning: Theoretical Foundation 1999

2. 概率图模型

有向图: Bayesian Network

无向图: Markov Network

推断

采样 Sampling

MCMC

Gibbs (MH)

Causal Inference