

计算机学习理论

目的：为学习算法提供理论基础

1. 准备知识

样例集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, $x_i \in X$, $y_i \in \{f(x_i), \bar{f}(x_i)\}$

设 X 中所有样本服从分布 π , D 中样本为简单分布

$$E_C(\lambda; D) = P_{x \sim D}(\lambda(x) \neq y) \rightarrow \text{泛化误差} \quad E_C(\lambda) = E_D(\lambda)$$

$$\hat{E}_C(\lambda; D) = \frac{1}{n} \sum_{i=1}^n I(\lambda(x_i) \neq y_i) \rightarrow \text{经验误差}$$

$E_C(\lambda) \leq \varepsilon$, ε 为泛化误差的上界

$$d(\lambda_1, \lambda_2) = P_{x \sim D}(\lambda_1(x) \neq \lambda_2(x)) \rightarrow \text{不一致度量 } \lambda_1, \lambda_2 \text{ 的差别.}$$

2. PAC 学习：概率近似正确

$c(x) = y \rightarrow$ 其中 c 为目标概念, 相应的集合记为 C

H : 假设空间, 是学习算法的集合. \rightarrow 集合大, 包含任意目标的可能性越大, 但从中找到某个具体目标概念难度越大

① $c \in H$, 闭包时上可分

有限假设空间 $\rightarrow |H|$ 有限
无限假设空间

② $c \notin H$, 不可分

由于数据量有限, 样本的偶然性, 无法精确学习到目标概念 c , 以较大概率学习误差满足

召没上限的模型

① $P(E_C(\lambda) \leq \varepsilon) \geq 1 - \delta$, 对于 $0 < \varepsilon, \delta < 1$, 则学习算法能从假设空间中 PAC 鉴别 C

② $m \geq \text{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, \text{size}(x), \text{size}(c))$, 则 C 为 PAC 可学习. \downarrow 艰任务在似乎的条件下可得到较好的模型.

③ 在 ② 下 运行时间也是 $\text{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, \text{size}(x), \text{size}(c))$ 则 C 是高阶 PAC 可学习 \rightarrow 算法在什么条件下可进行有效的学习.

若检测每个样本时间为常数, 时间复杂度 $m \geq \text{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, \text{size}(x), \text{size}(c))$ 中最大的 m

\rightarrow 需要多少训练样本能获得较好的模型.

3. 有限假设空间

① 可分: $c \in H$, 借助 D 剔除不一致的假设

计算达到有效近似的样本数.

$$\text{令 } E_C(\lambda) > \varepsilon, P(\lambda(x) = y) = 1 - E_C(\lambda) < 1 - \varepsilon$$

$$\therefore P(\lambda(x_1) = y_1) \wedge \dots \wedge (\lambda(x_m) = y_m) = (1 - E_C(\lambda))^m < (1 - \varepsilon)^m$$

$$P(\lambda \in H : E_C(\lambda) > \varepsilon \wedge \hat{E}_C(\lambda) = 0) < |H| (1 - \varepsilon)^m < |H| e^{-m\varepsilon} \rightarrow \text{至多不大于}$$

$$\therefore m \geq \frac{1}{\varepsilon} (\ln |H| + \ln \frac{1}{\delta}) \quad , \quad E_C(\lambda) 随 m 增多而收敛到 0, 速率是 O(\frac{1}{m}).$$

② 不可分： $c \in \mathbb{R}$

根据 Hoeffding 不等式有： $P(|E(h) - \hat{E}(h)| > \epsilon) \leq 2 \exp(-2n\epsilon^2)$

进而 $E(h) - \sqrt{\frac{\ln(2/\delta)}{2n}} \leq E(h) \leq \hat{E}(h) + \sqrt{\frac{\ln(2/\delta)}{2n}}$ \rightarrow m 越大时， $E(h) \approx \hat{E}(h)$.

同时有 $P(|E(h) - \hat{E}(h)| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}) \geq 1 - \delta$ \rightarrow

$$\begin{aligned} P(\exists h \in \mathcal{H} : |E(h) - \hat{E}(h)| > \epsilon) &\leq \sum_{h \in \mathcal{H}} P(|E(h) - \hat{E}(h)| > \epsilon) \\ &\leq 2 |\mathcal{H}| \exp(-2n\epsilon^2) \end{aligned}$$

找出泛化误差最小的假设 $\arg \min_{h \in \mathcal{H}} E(h)$

4. 无限假设空间.

标记结果 $H_D = \{h(x_1), h(x_2), \dots, h(x_m)\}$

$$I_H(m) = \max_{\{x_1, \dots, x_m\} \subseteq \mathcal{X}} |\{h(x_1), \dots, h(x_m) | h \in \mathcal{H}\}| \rightarrow$$
 增长函数描述假设空间 \mathcal{H} 的表示能力.

定理： $\forall n \in \mathbb{N}, 0 < \delta < 1$ 有 $P(|E(h) - \hat{E}(h)| > \epsilon) \leq 4 I_H(m) \exp\left(\frac{-n\epsilon^2}{8}\right)$

VC 维：能被升幂集的最大示例集大小 \rightarrow 与数据分布无关.

$$VC(H) = \max_{m \in \mathbb{N}} I_H(m) = 2^n$$

若假设空间 \mathcal{H} 的 VC 维为 d , 则 $\begin{cases} \text{对任意 } m \in \mathbb{N} \text{ 有 } I_H(m) \leq \sum_{i=0}^d \binom{m}{i} \\ \text{对任意整数 } m \geq d \text{ 有 } I_H(m) \leq \left(\frac{e \cdot m}{d}\right)^d \\ \text{对任意 } m > d, 0 < \delta < 1, h \in \mathcal{H} \end{cases}$

当 VC 维有限的假设空间 \mathcal{H} 都是不可知学习

$$P(|E(h) - \hat{E}(h)| \leq \sqrt{\frac{8d \ln \frac{2m}{\delta}}{m}}) \geq 1 - \delta,$$

ϵ 只与 m 有关, 速率 $O(\frac{1}{\sqrt{m}})$, 与数据分布, 样本集无关.

(2) Rademacher 复杂度：刻画假设空间复杂度, 不看数据分布.

经验误差量： $\arg \max_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n y_i h(x_i)$

高斯噪声, 引入随机变量 $g_i: \{-1, +1\} \rightarrow \mathbb{R}$, $\rightarrow E_g[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n g_i h(x_i)] = \begin{cases} 0, & |\mathcal{H}| = 1 \\ 1, & |\mathcal{H}| = 2^m, \text{ 完全拟合.} \end{cases}$

$$f: \mathcal{Z} \rightarrow \mathbb{R}, \quad \begin{cases} x \mapsto z \\ h \mapsto f_h \end{cases}$$

$$\rightarrow R_2(f) = E_g[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n g_i f(x_i)] \rightarrow$$
 函数空间 \mathcal{F} 上的经验 Rademacher 复杂度.

$$R_m(F) = E_Z[\sup_{|Z|=m} [R_2^*(F)]]$$

至少有 $1 - \delta$ 概率有 \leftarrow 有时回归问题.

$$E[Z] \leq \frac{1}{n} \sum_{i=1}^n f(x_i) + 2 R_m(F) + \sqrt{\frac{\ln(1/\delta)}{2n}}$$

实值函数空间 $F: \mathcal{Z} \rightarrow [0, 1]$, 数据分布 D 从 \mathcal{Z} 中独立

同分布采样得 $Z = \{z_1, z_2, \dots, z_n\}, z_i \in \mathcal{Z}, 0 < \delta < 1$, 对任意 $f \in F$

$$E[f(Z)] \leq \frac{1}{n} \sum_{i=1}^n f(z_i) + 2 R_2^*(F) + 3 \sqrt{\frac{\ln(1/\delta)}{2n}}$$

假设二分类：

假设空间 $H: X \rightarrow \{-1, +1\}$, 根据分布 D 从 X 中独立同分布

每样需到示例是 $D = \{x_1, x_2, \dots, x_n\}, x_i \in X, 0 < \delta < 1$, 且存在 $\lambda \in H$

若有 $1-\delta$ 概率有

$$E(\lambda) \leq \hat{E}(\lambda) + R_m(H) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

$$E(\lambda) \leq \hat{E}(\lambda) + R_D(H) + 3\sqrt{\frac{\ln(1/\delta)}{2m}}$$

Rademacher 复杂度与分布无关, 均有差.

若 $R_m(H) \leq \sqrt{\frac{2\ln(2m)}{m}}$

$$\rightarrow E(\lambda) \leq \hat{E}(\lambda) + \sqrt{\frac{2\ln(2m)}{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

(3) 稳定性：高斯且样本学习算法

β -均匀稳定性:

$D^{(i)}$: 特殊 D 中第 i 个样例

D^i : 普通 D 中第 i 个样例

泛化误差: $l(\lambda, D) = E_{x \in X, z = (x, y)} [l(\lambda_D, z)]$

经验误差: $\hat{l}(\lambda, D) = \frac{1}{n} \sum_{i=1}^n l(\lambda_{D^i}, z_i)$

留一误差: $l_{\text{loo}}(\lambda, D) = \frac{1}{n} \sum_{i=1}^n l(\lambda_{D^{\setminus i}}, z_i)$

若学习算法是 ERM 且稳定的, 则假设空间可学习

$$|l(\lambda_D, z) - l(\lambda_{D^{(i)}}, z)| \leq \beta, i = 1, 2, \dots, n$$

若学习算法满足关于损失函数 l 的 β 稳定性, 其上界为 M ,

$0 < \delta < 1$, 对任意 $m \geq 1$, 至少 $1-\delta$ 概率有

$$l(\lambda, D) \leq \hat{l}(\lambda, D) + 2\beta + C(4m\beta + M) \sqrt{\frac{\ln(1/\delta)}{2m}}$$

$$l(\lambda, D) \leq l_{\text{loo}}(\lambda, D) + \beta + C(4m\beta + M) \sqrt{\frac{\ln(1/\delta)}{2m}}$$