

可学习性分析

► 一个简单的二分类问题，给定如下已知条件，可否学到一个函数 g 接近于目标函数 f ？

x_n	$y_n = f(x_n)$
0 0 0	o
0 0 1	x
0 1 0	x
0 1 1	x
1 0 0	x
1 0 1	?
1 1 0	o
1 1 1	x

x	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	o	o	o	o	o	o	o	o	o	o
0 0 1	x	x	x	x	x	x	x	x	x	x
0 1 0	x	x	x	x	x	x	x	x	x	x
0 1 1	x	x	x	x	x	x	x	x	x	x
1 0 0	x	x	x	x	x	x	x	x	x	x
1 0 1	?	o	o	o	o	o	o	o	o	o
1 1 0	o	o	o	x	x	x	x	x	x	x
1 1 1	x	x	x	x	x	x	x	x	x	x

目标函数 f 未知，可能的假设空间 \mathcal{H} 有 2^3 种可能，从中选择前5个结果与给定条件一致的 2^3 个假设

抽样得到 真实分布 使用~~已知量~~估计~~未知量~~ Hoeffding不等式 $P(|u-v| > \epsilon) \leq 2\exp(-2\epsilon^2 N)$ N 足够大时， $u \approx v$ 概率近似正确

任何 \hat{h} 预测函数无法在所有的训练样本上表现的一样好

$|u-v| > 0.3$ 测验： $u=0.4$, 采样10个样本得到的比例 <0.1 , 那 u 和 v 差距较大的上限是多少?

$$2\exp(-2\epsilon^2 N)$$

A. 0.67 B. 0.4 C. 0.33 D. 0.05 ✓

抽样模型及霍夫丁不等式与学习的联系

罐子抽样模型		学习模型	
需确定的量：	橙球比例 u	需输出的量：	给定假设 $h(\vec{x}) = f(\vec{x})$ 是否成立
球的空间：	罐子	输入空间：	$\vec{x} \in \mathcal{X}$
橙球		正确的： $h(\vec{x}) = f(\vec{x})$	
从罐子里抽 n 个样本		错误的： $h(\vec{x}) \neq f(\vec{x})$	

备注：统计分析要求样本是均匀采样的
机器学习要求样本满足独立同分布 (independent and identically distributed, 简写 i.i.d.)

数据集规模足够大时，可通过数据集上 $h(\vec{x}) \neq f(\vec{x})$ 概率推断输入空间上 $h(\vec{x}) \neq f(\vec{x})$ 的概率

当存在多个假设时 $\mathcal{H} = \{h_1, h_2, \dots, h_M\}$

依据墨菲定律，虽然某事件发生几率极低，但如果重复次数达到大规模的时候，则小概率的事件必然发生。即 $E_{in}(h)$ 与 $E_{out}(h)$ 相差甚远的情况（称为BAD样本）可能发生

$P_D[\text{BAD}] \leq 2M\exp(-2\epsilon^2 N)$

依据霍夫丁不等式，BAD样本产生的概率低

$P_D[\text{BAD}] = \sum_{\text{all possible } D} P(D) \cdot P[\text{BAD } D]$

D_1	D_2	\dots	D_{120}	\dots	D_{320}	Hoeffding
h_1	BAD					$P_D[\text{BAD } D \text{ for } h_1] \leq \dots$
h_2		BAD				$P_D[\text{BAD } D \text{ for } h_2] \leq \dots$
h_3			BAD			$P_D[\text{BAD } D \text{ for } h_3] \leq \dots$
h_4				BAD		$P_D[\text{BAD } D \text{ for } h_4] \leq \dots$
all	BAD	BAD	BAD	BAD	BAD	$P_D[\text{BAD } D \text{ for } h_5] \leq \dots$

算法在假设空间

$\mathcal{H} = \{h_1, h_2, \dots, h_M\}$

由中自由选择的对BAD样本的选则

$= 2\exp(-2\epsilon^2 N) + 2\exp(-2\epsilon^2 N) + \dots + 2\exp(-2\epsilon^2 N)$

$= 2M\exp(-2\epsilon^2 N)$

若 $|\mathcal{H}|=M$ 有限， N 足够大，任选假设 h 为学习函数 g

若 $E_{in}(g) \approx E_{out}(g)$

若 $E_{in}(g) \approx 0$ 则 $E_{out}(g) \approx 0 \Rightarrow$ 数据集外 $g \approx f$ \Rightarrow 机器可学习

Hoeffding不等式与模式二分性

存在 M 个假设时， $P_D[\text{BAD } D] \leq 2M\exp(-2\epsilon^2 N)$

仅使用有限值 M_H 代替无限大的 M

本 N 个样本的PLA最多有 2^N 种二分类，即有效直线数量 effective(N) $\leq 2^N$ 保证其 $\leq 2^N$ 时接近 0 , $E_{in}(g) \approx E_{out}(g)$ 可学习

使用PLA有效直线数量代替 M 有Hoeffding不等式 $P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2\text{effective}(N)\exp(-2\epsilon^2 N)$

定义成长函数 $m_{\mathcal{H}}(N) = \max_{h \in \mathcal{H}} |\{x_1, \dots, x_N\}|$ 消除对输入的依赖，值与 N 无关

用成长函数 $m_{\mathcal{H}}(N)$ 代替 M ，霍夫丁不等式可写成

$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2m_{\mathcal{H}}(N)\exp(-2\epsilon^2 N)$

若 $m_{\mathcal{H}}(N)$ 为多项式，则 $E_{in}(h) \approx E_{out}(h)$ 成立

若 $m_{\mathcal{H}}(N)$ 为指数函数，则 $E_{in}(h) \approx E_{out}(h)$ 不成立

2维平面上感知器学习模型

的成长函数 $m_{\mathcal{H}}(N)$ 为：

\dots

N

$< 2^N$

1维平面上感知器学习模型的成长函数 $m_{\mathcal{H}}(N)$

Positive Rays (正射线)

$h(x) = \text{sign}(x - a)$

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

Positive Intervals (正间隔)

$h(x) = \begin{cases} +1, & \text{if } x \in [l, r] \\ -1, & \text{otherwise} \end{cases}$

$m_{\mathcal{H}}(N) = C_{N+1}^2 + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

Convex set

$m_{\mathcal{H}}(N) = 2^N$

二维平面上的感知器模型

的成长函数 $m_{\mathcal{H}}(N)$ 为：

\dots

N

$< 2^N$

Convex set的成长函数 $m_{\mathcal{H}}(N)$

考虑一种输入：所有样本点分布在一个圆上

假设空间 \mathcal{H} 为任意凸多边形，该多边形所覆盖到的点为+1，其他点为-1

成长函数为：

$m_{\mathcal{H}}(N) = 2^N$

→ 这 N 个样本可被 \mathcal{H} 所打散 (shattered)

二维平面感知器

1) $m_{\mathcal{H}}(N) = 2$

2) $m_{\mathcal{H}}(N) = 4$

3) $m_{\mathcal{H}}(N) = 8$

4) $m_{\mathcal{H}}(N) = 14$

(共4类情况) (XOR不可分)

VC维

前面的论述用 $m_{\mathcal{H}}(N)$ 代替 M ，从 \mathcal{H} 中选取学习函数 g ，有

$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2m_{\mathcal{H}}(N)\exp(-2\epsilon^2 N)$

但直接替换是有问题的，Vapnik 和 Chervonenkis 提出并完整地证明准确的式子如下 (具体证明参考论文[Vapnik et al. 1971])

$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2 \cdot \frac{1}{16} \epsilon^{-2}$

上式右边是VC Bound，若存在break point $k \geq 3$ ，则

$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 4(2N)^{k-1}\exp(-\frac{1}{8}\epsilon^2 N)$

M 与VC维 d_{VC} 对机器可学习性的影响，机器可学习说明

1) 预测误差接近于训练误差 $E_{out}(h) \approx E_{in}(h)$ ；

2) 存在可选的某假设 g 使得训练误差足够小 $E_{in}(g) \approx 0$

M 小

1) 满足， $P(BAD) \leq 2M\exp(\dots)$

2) 不一定满足，假设空间选择太少

M 大

1) 不满足， $P(BAD) \leq 2M\exp(\dots)$

2) 满足，假设空间选择很多

d_{VC} 小

1) 不满足， $P(BAD) \leq 4(2N)^{d_{VC}}\exp(\dots)$

2) 满足，假设空间有效模式少

d_{VC} 大

1) 不满足， $P(BAD) \leq 4(2N)^{d_{VC}}\exp(\dots)$

2) 满足，假设空间有效模式多

对VC维进行重写 $P(|E_{in}(g) - E_{out}(g)| < \epsilon) \leq 4(2N)^{d_{VC}}\exp(-\frac{1}{8}\epsilon^2 N) \leq \epsilon$

$\Rightarrow \epsilon = \sqrt{\frac{8}{N} \ln \frac{4(2N)^{d_{VC}}}{\epsilon}}$

$\Rightarrow E_{in}(g) - \sqrt{\frac{8}{N} \ln \frac{4(2N)^{d_{VC}}}{\epsilon}} \leq E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4(2N)^{d_{VC}}}{\epsilon}}$

预测误差 逆为 d_{VC} 模型复杂度

对VC维进行重写 $P(|E_{in}(g) - E_{out}(g)| < \epsilon) \leq 4(2N)^{d_{VC}}\exp(-\frac{1}{8}\epsilon^2 N) \leq \epsilon$

$\Rightarrow \epsilon = \sqrt{\frac{8}{N} \ln \frac{4(2N)^{d_{VC}}}{\epsilon}}$

$\Rightarrow E_{in}(g) - \sqrt{\frac{8}{N} \ln \frac{4(2N)^{d_{VC}}}{\epsilon}} \leq E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4(2N)^{d_{VC}}}{\epsilon}}$

预测误差 逆为 d_{VC} 模型复杂度

机器可学习性