Southwestern University Of Finance And Economics spring, 2021
School of Finance Lecture 7
Scribe: *Haoming Wang* Lecturer: *Prof. Xuantian Lin* @ NTU

机器学习基石 I: VC Dimension

Key word: VC dimension, VC dimension of Perceptron, Penalty for Model Complexity

1 VC dimension

在上一讲中, 已经证明若 $\mathcal{H}$ 的 break point 为 k, 则其成长函数 $m_{\mathcal{H}}(N)$ 的上界 (的上界) 是 $k-1$ 次多项式

$$m_{\mathcal{H}}(N) \le B(N,k) \le \sum_{i=0}^{k-1} C_N^i \le N^{k-1},$$

因此对于任意 $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$ 有

$$\begin{aligned}\mathbb{P}_{\mathcal{D}}[|E_{in}(g) - E_{out}(g)| > \epsilon] &\le \mathbb{P}_{\mathcal{D}}[\exists h \in \mathcal{H} \text{ s.t. } |E_{in}(h) - E_{out}(h)| > \epsilon] \\ &\le 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} \\ &\le 4(2N)^{k-1}e^{-\frac{1}{8}\epsilon^2 N}.\end{aligned}$$

这说明, 当我们拥有一个好的 hypothesis set $\mathcal{H}$ ($m_{\mathcal{H}}(N)$ breaks at k), 一个好的样本集 $\mathcal{D}$(N 足够大使得 $E_{in}(g) = E_{out}(g)$ p.a.c.), 一个好的算法 $\mathcal{A}$ (可以从 $\mathcal{H}$ 中选出 $E_{in}(h)$ 足够小的 h 作为 g), 则机器学习或许是可行的.

Definition 1.1: VC dimension

假设集 $\mathcal{H}$ 的 VC dimension $d_{vc}(\mathcal{H})$ 是样本可以被 shatter 的最大容量. 即存在 N 个样本, 使得 $m_{\mathcal{H}}(N) = 2^N$, 且任何 $M > N$ 个样本都有 $m_{\mathcal{H}}(M) < 2^M$, 则 $d_{vc}(\mathcal{H}) = N$.

换句话说,$d_{vc}(\mathcal{H}) = k-1$, k 是最小突破点. 如果 $N \le d_{vc}(\mathcal{H})$ 则存在 N 个输入样本, 使得可以被 $\mathcal{H}$ shatter. (注意不是任意 N 个输入样本), 因此我们可以将上面的公式改写为

$$\mathbb{P}_{\mathcal{D}}[|E_{in}(g) - E_{out}(g)| > \epsilon] \le 4(2N)^{d_{vc}(\mathcal{H})}e^{-\frac{1}{8}\epsilon^2 N}.$$

一个好的情况, 是 $d_{vc}(\mathcal{H}) < \infty$, 此时我们说 g 是可以被泛化到样本 $\mathcal{D}$ 之外的, 即 $E_{in}(g) = E_{out}(g)$ p.a.c.

2 VC dimension of Perceptron

回顾二维感知器模型, 对于线性可分的样本集 $\mathcal{D}$, 我们已经证明只要迭代的次数足够 PLA 最终会收敛到 $E_{in}(g) = 0$; 对于 $\mathbf{x}_n$ 服从某一分布 (线性不可分), $y_n = f(\mathbf{x}_n)$ 的数据集, $d_{vc} = 3$, 当样本量 N 足够时, 有 $E_{in}(g) = E_{out}(g)$ p.a.c. 从而二维感知器模型可以学习. 下面考虑 n 维的情况.

一维感知器模型 (Positive/Negative rays) 的 VC dimension $d_{vc} = 2$; 二维感知器模型的 VC dimension $d_{vc} = 3$, 猜测 n 维感知器模型的 $d_{vc} = n+1$.

Theorem 2.1: d_{vc} of n-Perceptron

n 维感知器模型的 VC dimension $d_{vc} = n+1$.

我们分别证明 $d_{vc} \geq n+1$ 以及 $d_{vc} \leq n+1$. 注意, $d_{vc} \geq n+1$ 说明, 存在某 $n+1$ 个输入样本可以被 $\mathcal{H}$ shatter. 下面我们构造 $n+1$ 个输入, 并证明它们是可以被 shatter 的. 令 $\mathbf{x}_0$ 为 origin $\mathbf{0}$; 令 $\mathbf{x}_i$ 为单位向量 $\mathbf{e}_i$:

$$X = \begin{bmatrix} \mathbf{x}_0^T \\ \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ 1 & \vdots & \ddots & 0 \\ 1 & 0 & \cdots & 1 \end{bmatrix}$$

Note 1. 我们统一令每个输入的第一个元素为 1, 因为 $\sum_{i=1} w_i x_i + b = \sum_{i=0} w_i x_i$, 其中 $w_0 = b, x_0 = 1$.

易知 X 可逆, 对于这 $n+1$ 个输入的任意输出 $\mathbf{y}$, 令 $\mathbf{w} = X^{-1}\mathbf{y}$, 则 $X\mathbf{w} = \mathbf{y}$, 从而 $\operatorname{sign}(X\mathbf{w}) = \mathbf{y}$. 因此该 $n+1$ 个输入可以被 shatter, 即 $d_{vc} \geq d+1$.

若 $\mathcal{H}$ 的 $d_{vc} \leq n+1$, 则说明, 对于任意 $n+2$ 个输入, $\mathcal{H}$ 都不可以被 shatter, 即存在一个输出 $\mathbf{y}$, 使得 $\forall h \in \mathcal{H}$ 有 $h(\mathcal{X}) \neq \mathbf{y}$. 其中 $\mathcal{X} = \{\mathbf{x}_0, \cdots, \mathbf{x}_{n+1}\}$. 注意到 $\mathbf{x}_i \in \mathbb{R}^n$, 这么说明 $\mathcal{X}$ 线性相关, 即存在一组不全为零的实数 $\lambda_i, i = 0, \cdots, n+1$, 使得

$$\mathbf{x}_{n+1} = \sum_{i=0}^{n} \lambda_i \mathbf{x}_i.$$

令 $\mathbf{y} = \{\operatorname{sign}(\lambda_1), \cdots, \operatorname{sign}(\lambda_n), -1\}$, 则对于任意满足 $\operatorname{sign}(\mathbf{w}^T\mathbf{x}_i) = \operatorname{sign}(\lambda_i)$ 的 $\mathbf{w}$, 有

$$\mathbf{w}^T\mathbf{x}_{n+1} = \sum_{i=0}^{n} \lambda_i \mathbf{w}^T\mathbf{x}_i > 0.$$

即 $\operatorname{sign}(\mathbf{w}^T\mathbf{x}_{n+1}) \neq -1$, 因此任意 $n+2$ 个输入, 必存在一组输出不能被实现, 即不可被 shatter, 因此 $d_{vc} \leq n+1$. 综上 n 维感知器的 $d_{vc} = n$.

我们来考察以下 VC dimension 的物理含义, 考虑一个 n 维 Perceptron 的公式

$$y = \operatorname{sign}\left(\mathbf{w}^T\mathbf{x}\right) = \operatorname{sign}(w_0x_0 + \cdots + w_nx_n),$$

其中 $w_0 = b$. 即 Perceptron 一共有 $n+1$ 个参数进行调节, 正式的说, Perceptron 共有 $n+1$ 个自由度. d_{vc} 衡量的就是 hypothesis set $\mathcal{H}$ 进行二分类的自由度, 它表示 $\mathcal{H}$ 的分类能力 (灵活程度), 一个较高的 d_{vc} 可以找到 $E_{in}(h)$ 更小的 h, 但是也会增加 BAD ($|E_{in}(h) - E_{out}(h)| > \epsilon$, 过拟合) 发生的概率.

对于 Positive ray $\mathcal{H} = \{h = \operatorname{sign}(x-a) : a \in \mathbb{R}\}$ 而言, 其二分类自由度只有一个参数 a, 因此其 $d_{vc} = 1$. 对于 Positive Interval $\mathcal{H} = \{h = \operatorname{sign}((x-a)(a+h-x) : a \in \mathbb{R}, h \in \mathbb{R}^+)\}$. 其二分类自由度有两个参数 a, 因此其 $d_{vc} = 2$. 对于 n 维 Perceptron 而言, 若令 $b = 0$ (即直线/平面/超平面必须过原点), 则其自由度少一个, 此时 $d_{vc} = n$.

3 Penalty for Model Complexity

回顾霍夫丁不等式, 对于任意 $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$, 有

$$\mathbb{P}_{\mathcal{D}}[\underbrace{|E_{in}(g) - E_{out}(g)| > \epsilon}_{BAD}] \leq \underbrace{4(2N)^{d_{vc}(\mathcal{H})} e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}.$$

如果我们要有 $1-\delta$ 的置信度保证 BAD 不发生, 则置信区间为

$$\epsilon = \sqrt{\frac{8}{N} \ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right)} =: \Omega(N, \mathcal{H}, \delta)$$

定义 $\Omega(N, \mathcal{H}, \delta)$ 为模型复杂度的惩罚项 (penalty for model complexity). 换句话说, 我们有 $1-\delta$ 的概率保证

$$E_{out}(g) \leq E_{in}(g) + \underbrace{\sqrt{\frac{8}{N} \ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right)}}_{\Omega(N,\mathcal{H},\delta)}$$

Note 2. 一般我们不在乎 $E_{out} < E_{in}$ 的情况.

因此当假设集 $\mathcal{H}$ 的 d_{vc} 增加的时候, $\mathcal{H}$ 的解释能力增强, 样本内误差 E_{in} 下降. 当 d_{vc} 增加的时候, 模型复杂度的惩罚项增加, 模型置信区间变宽, E_{out} 增加.