Lecture 7 The VC Dimension

整理者: LobbyBoy* 2020年2月21日

1. Recap

首先我们对上一章的内容进行简要的回顾。上一章我们通过归纳的方法得到了一个非常重要的结论：当某个假说集合存在断点时(假设最小断点为k)，那么该假说集合的增长函数能够被一个$k-1$次多项式bound住，即(To be honest, 上一章我们仅仅得到了$B(N,k)$小于等于组合数的累加，而这里却直接用了等于号。其实两者的确是相等的，只需再证明$B(N,k)$大于等于组合数累加即可)：

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

结合VC bound，当样本数量N非常大时，即使上界中的增长函数项也会变得很大，但是其增长速度也仅仅为polynomial的，而上界中的指数项则以exponential速度趋于0。因此总体上看，当N很大时，上界趋于0，我们便得到了$E_{in} \approx E_{out}$的保证。

由于对组合数进行累加并不是一个特别方便计算且直观的上界，因此我们希望通过继续loose来得到一个更友好的上界。一个简单的猜想是：

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

	$B(N,k)$	k=1	2	3	4	5
	1	1	2	2	2	2
	2	1	3	4	4	4
	3	1	4	7	8	8
N	4	1	5	11	15	16
	5	1	6	16	26	31
	6	1	7	22	42	57

N^{k-1}	k=1	2	3	4	5
1	1	1	1	1	1
2	1	2	4	8	16
3	1	3	9	27	81
4	1	4	16	64	256
5	1	5	25	125	625
6	1	6	36	216	1296

图 1: $B(N,k)$ & N^{k-1}

*本笔记根据台湾大学林轩田教授于线上教育平台Coursera开设的“机器学习基石”课程整理而成(课程内容见: https://www.coursera.org/learn/ntumlone-mathematicalfoundations/home/welcome)。笔记内的大多数图片来自于林老师的课程slides。感谢林老师能够将如此精彩的课程通过线上平台同所有人分享，thanks!

因此，我们不加证明地得到，当$N \geq 2, k \geq 3$时:

$$m_{\mathcal{H}}(N) \leq N^{k-1}$$

将上面的不等式代入下面的VC bound中：

$$P\Big[\exists h \in \mathcal{H} \ s.t. \ |E_{in}(h) - E_{out}(h)| > \epsilon\Big] \leq 4 \cdot m_{\mathcal{H}}(2N) \cdot exp(-\frac{1}{8}\epsilon^2 N)$$

得到：

$$P\Big[\exists h \in \mathcal{H} \ s.t. \ |E_{in}(h) - E_{out}(h)| > \epsilon\Big] \overset{if\ k\ exists}{\leq} 4 \cdot (2N)^{k-1} \cdot exp(-\frac{1}{8}\epsilon^2 N)$$

从上面的bound中我们可以得到关于learnable的几个重要条件：

第一，假说集合$\mathcal{H}$存在断点——我们将其称为“Good $\mathcal{H}$”；

第二，样本量N需要足够大，这样我们就probably能够得到$E_{in} \approx E_{out}$，我们将这个条件总结为“Good $\mathcal{D}$”；

第三，我们需要一个好的演算法$\mathcal{A}$，该演算法能够从假说集合中挑出一个E_{in}较小的hypothesis，我们将这个条件总结为“Good $\mathcal{A}$”；

最后，我们还需要一点点运气：“Good luck”!

2. Definition of VC Dimension

我们将“maximum non-break point”称为“VC dimension”。也就是说，VC dimension是假说集合能够打散的最多的点的数量，假说集合无论如何也无法打散VC dimension+1个样本，而VC dimension+1也就是该假说集合的最小断点。因此，我们若以k记某假说集合的最小断点，则有：

$$d_{VC}(\mathcal{H}) = k - 1$$

引入了VC dimension后，我们可以将此前公式中的$k-1$全部换成d_{VC}：当$N \geq 2, d_{VC} \geq 2$时:

$$m_{\mathcal{H}}(N) \leq N^{d_{VC}}$$

回忆我们此前讨论过的四个假说集合。Positive rays的最小断点为2，因此$d_{VC} = 1$；Positive intervals的最小断点为3，因此$d_{VC} = 2$；Convex sets的最小断点为无穷大，因此$d_{VC} = \infty$；2D perceptrons的最小断点为4，因此$d_{VC} = 3$。

使用VC dimension的好处是：第一，较增长函数而言很容易计算；第二，和增长函数一样，只与$\mathcal{H}$有关，与演算法$\mathcal{A}$、潜在分布$\mathcal{P}$、目标函数f无关。

Question: If there is a set of N inputs that cannot be shattered by $\mathcal{H}$. Based only on this information, what can we conclude about $d_{VC}(\mathcal{H})$? 无法得到任何关于该假说集合VC dimension的结论。因为题中仅仅告诉我们该假说集合无法打散某一组N个样本，因此该假说集合可能打散另一组N个样本，所以我们不能说该假说集合的VC dimension小于N：可以小于，可以等于，也可以大于。

3. VC Dimension of Perceptrons

本节我们将证明d-D perceptrons的VC dimension为d+1。证明分为两个部分：第一，证明$d_{VC} \geq d+1$；第二，证明$d_{VC} \leq d+1$；两者结合，我们将得到$d_{VC} = d+1$。

要证$d_{VC} \geq d+1$，我们只需证明：存在某$d+1$个样本，d-D perceptrons可以将它们shatter。我们将构建如下的$d+1$个样本：

$$\mathrm{X} = \begin{bmatrix} \text{—}\ \mathbf{x}_1^T\ \text{—} \\ \text{—}\ \mathbf{x}_2^T\ \text{—} \\ \text{—}\ \mathbf{x}_3^T\ \text{—} \\ \vdots \\ \text{—}\ \mathbf{x}_{d+1}^T\ \text{—} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & & \ddots & 0 \\ 1 & 0 & \dots & 0 & 1 \end{bmatrix}$$

图 2: 一个特别的容量为$d+1$的样本集

这$d+1$个样本可以被shatter掉，等价于：$sign(\mathbf{X}\mathbf{w}) = \mathbf{y}$对于任何$\mathbf{y}$(每个位置的元素为0或1)有解。注意到这里的$\mathbf{X}$是可逆的，因此$\mathbf{X}\mathbf{w} = \mathbf{y}$对于任何$\mathbf{y}$有解$\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$。而该解自然也是$sign(\mathbf{X}\mathbf{w}) = \mathbf{y}$的解。综上，我们找到了能被d-D perceptrons shatter掉的一组特殊的d+1个样本，因此我们推断$d_{VC} \geq d+1$。

要证$d_{VC} \leq d+1$，我们需证明：对于任何$d+2$个样本，d-D perceptrons都不可以将它们shatter。假设我们有$d+2$个样本，依然将它们以矩阵形式排列起来：

d-D General Case

$$\mathrm{X} = \begin{bmatrix} \text{—}\ \mathbf{x}_1^T\ \text{—} \\ \text{—}\ \mathbf{x}_2^T\ \text{—} \\ \vdots \\ \text{—}\ \mathbf{x}_{d+1}^T\ \text{—} \\ \text{—}\ \mathbf{x}_{d+2}^T\ \text{—} \end{bmatrix}$$

more rows than columns:

linear dependence (some a_i non-zero)

$$\mathbf{x}_{d+2} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \dots + a_{d+1}\mathbf{x}_{d+1}$$

图 3: $d+2$个样本

此时$\mathbf{X}$是$d+2$行、$d+1$列的矩阵，因此其行是线性相关的。因此存在不全为0的系数，使得样本的加权和为0。不妨设$\mathbf{x}_{d+2}$的系数不为0，并假设$a_1,...,a_{d+1}$中至少有一个不为0(都为0的话，则$\mathbf{x_{d+2}}$为$\mathbf{0}$，这样的样本集肯定无法被shatter，因为$\mathbf{w^T x_{d+2}}$恒为0，在规定了$sign(0)$的规则后，对立的一面永远无法通过改变$\mathbf{w^T}$而得到)，则：

$$\mathbf{x}_{d+2}=a_1\mathbf{x}_1+a_2\mathbf{x}_2+...+a_{d+1}\mathbf{x}_{d+1}$$

两边同时乘$\mathbf{w}$并取$sign$得：

$$sign\Big(\mathbf{w^T}\mathbf{x}_{d+2}\Big)=sign\Big(a_1\Big)sign\Big(\mathbf{w^T}\mathbf{x}_1\Big)+...+sign\Big(a_{d+1}\Big)sign\Big(\mathbf{w^T}\mathbf{x}_{d+1}\Big)$$

可见，有一种dichotomy是无论如何也做不出来的：

$$(sign\Big(a_1\Big),sign\Big(a_2\Big),...,sign\Big(a_{d+1}\Big),-1)$$

因为此时：

$$sign\Big(\mathbf{w^T}\mathbf{x}_{d+2}\Big)=sign\Big(a_1\Big)sign\Big(a_1\Big)+...+sign\Big(a_{d+1}\Big)sign\Big(a_{d+1}\Big)>0$$

总结一下：任意$d+2$个样本必线性相关→必存在某个样本能被其他样本线性表示→那么当其他样本都被“预测”为$sign(a_i)$时，这个样本则“必须”被预测为+1，这是由线性相关性所决定的→任意$d+2$个样本都无法shatter掉→$d_{VC}\leq d+1$。

综上所述，因为$d_{VC}\geq d+1$且$d_{VC}\leq d+1$，所以$d_{VC}=d+1$。

4. Physical Intuition of VC Dimension

通过上一节的推导，我们知道了d-D perceptrons的VC dimension为$d+1$，恰好是$\mathbf{w}=(w_0,w_1,...,w_d)$的维度数。

我们最初单纯地用M来表示一个假说集合的大小。因为perceptron $\mathbf{w}$中的每个维度都可以从负无穷变到正无穷，因此我们认为perceptron的$M=\infty$。而现在我们建立的VC dimension“恰好”等于perceptron $\mathbf{w}$的维度，正说明了VC dimension从另一个方面描绘了假说集合的“大小”：$\mathbf{w}$的维度越大，则假说集合越“复杂”，相应的VC dimension越大。此外，VC dimension等于perceptron $\mathbf{w}$的维度，也使得我们有理由推测VC dimension的大小等于假说集合的自由度(degrees of freedom)，即能够自由变化的参数的数目。这一点在“很大程度”上是对的，可以用positive rays、positive intervals来验证：

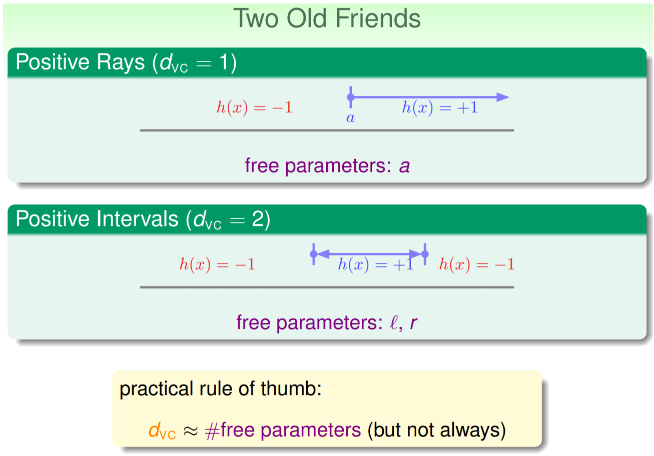

图 4: degrees of freedom

因此，VC dimension可以看作是假说集合复杂程度的另一种表征：VC dimension越大的假说集合越复杂，能够shatter掉的样本数目越多。

5. Interpreting VC Dimension

回忆VC bound的叙述: For any $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$ and 'statistical' large $\mathcal{D}$, for $N \geq 2$, $d_{VC} \geq 2$:

$$P_{\mathcal{D}}\Big[|E_{in}(g) - E_{out}(g)| > \epsilon\Big] \leq 4 \cdot (2N)^{d_{VC}} \cdot exp(-\frac{1}{8}\epsilon^2 N)$$

上面的这种叙述形式实质在告诉我们：“坏事”(演算法回传的hypothesis g的经验误差与泛化误差相差很多)发生的概率不是很高(被一个上界bound住)。“坏事”发生的概率不高，等价于“好事”会以较高的概率发生。因此我们可以将VC bound重写成另一种形式：with probability $\geq 1 - \delta$，we have $|E_{in}(g) - E_{out}| \leq \epsilon$。只需令VC bound右侧等于$\delta$即可解出$\epsilon$的表达式：

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

因此，我们有至少$1 - \delta$的概率：

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

然而我们一般对泛化误差的下界没有兴趣，只保留上界。因此我们有：至少$1 - \delta$的概率：

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

我们将上式右侧的根号项称为penalty for model complexity，即模型复杂度，记作$\Omega(N, \mathcal{H}, \delta)$。

我们以VC dimension为横轴，Error为纵轴建立坐标系。当使用具有更高VC dimension的假说集合时(+合理的演算法)，我们期望能够得到具有更小经验误差的hypothesis，即in-sample error会下降；但是由于模型复杂度提升，Ω项会增大，对应于图中上升的红线。两者相加即为泛化误差的上界，因此泛化误差先增大后减小，呈山谷型。这张图给我们的最大启示是：使用复杂度高的模型并不总能得到好的结果，甚至会得到非常“危险”的结果(特别是在没有很多训练数据的时候)。

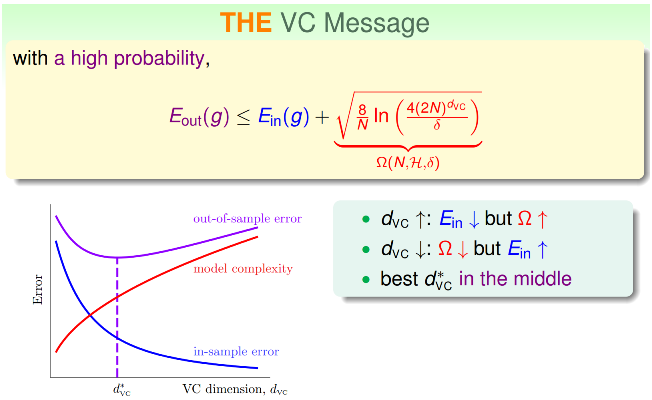

图 5: VC Message

VC bound的另一个用法是计算样本复杂度(sample complexity)，也就是在设定容忍误差ϵ与精度要求δ的基础上，计算最少所需样本数。例如，假说集合的$d_{VC}=3$，$\epsilon=0.1$，$\delta=0.1$，那么通过解：

$$4 \cdot (2N)^{d_{VC}} \cdot exp(-\frac{1}{8}\epsilon^2 N) \le \delta$$

得到$N \ge 29300$，因此VC bound告诉我们至少需要近三万个样本。然而，实践中我们只需要大概$N = 10d_{VC}$个样本就足够了，对于上例也就是30个。为何理论结果与实践结果相差这么远？或者换一种问法，为何VC bound的上界这么松(loose)？这是因为：

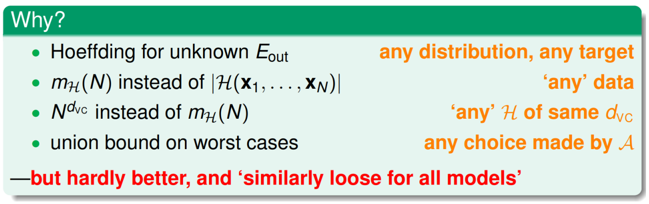

图 6: Looseness of VC Bound

总而言之，因为我们为了简化计算、扩大适用范围，在推导VC Bound时放大了多次上界，所以导致VC Bound会高估泛化误差。