七、The VC Dimension VC维

7.1 Definition of VC Dimension VC维的定义

先对上一章的内容作简单总结：如果一个假设空间存在突破点。则一定存在成长函数$m_H(N)$被某个上限函数$B(N,k)$所约束，可求出上限函数等于一个组合的求和形式$\sum_{i=0}^{k-1} C_N^i$，易知该形式的最高次项是N^{k-1}。图7-1a)和b）分别是以上限函数为成长函数上限的情况和以为成长函数上限的情况。

	$B(N,k)$	k=1	k=2	k=3	k=4	k=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	k=2	k=3	k=4	k=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7-1 a) 以上限函数为上限 b) 以N^{k-1}为上限

从图中可以看出在$N \geq 2$且$K \geq 3$的情况下，满足$B(N,k) \leq NK-1$，得到公式7-1。

$$m_H(N) = B(N,k) = \sum_{i=0}^{k-1} C_N^i \leq N^{k-1}$$

通过公式7-1和上一章的结论可以得出公式7-2。

$$\begin{aligned} P_D[|E_{in}(g) - E_{out}(g)| > \varepsilon] &\leq P_D[\exists h \in s.t. |E_{in} - E_{out}| > \varepsilon] \\ &\leq 4m_H(2N)\exp(-\frac{1}{8}\varepsilon^2 N) \\ &\leq^{\text{if K exists}} 4N^{k-1}\exp(-\frac{1}{8}\varepsilon^2 N) \end{aligned}$$

该公式的意义是在输入样本N很大时，VC限制一定成立，同时等式的左边也一定会在$k \geq 3$的情况下被以多项式形式（N^{k-1}）所约束（注意这里$N \geq 2$的条件没有了，原因很简单，VC限制是样本N很大的情况下产生的，因此一定满足$N \geq 2$的条件），而在k<3的情况下有其他的限制可以满足（比如前几章提到的如正射线之类的分类不需要多项式形式的限制也可以约束住成长函数）。

至此得知，满足以下几个条件，机器便可以学习：

1. 假设空间的成长函数有一个突破点k（有好的假设空间H）；
2. 输入数据样本N足够的大（有好的输入样本集D）；

1和2通过VC限制共同推出了E_{in}和E_{out}有很大的可能很接近。

1. 一个算法A能够找出一个使 E_{in} 足够小的g（好的算法A）；

再结合1和2得出的结论就可以进行学习了（当然这里需要一点好的运气）。

接下来介绍一下这一节的正题，VC维度或者VC维（VC dimension）是什么意思。

它的定义和突破点（break point）有很大关系，是最大的一个不是突破点的数。

VC维是假设空间的一个性质，数据样本可以被完全二分的最大值。用d_{VC}作为VC维的数学符号，假如突破点存在的话，即最小的突破点减去1，如公式7-3所示；如果不存在突破点的话，则VC维为无限大。 $d_{VC} = ^{'}$ 最小的$k' - 1$

如果输入数据量N小于VC维d_{VC}，则有可能输入数据D会被完全的二分类，这里不是一定，只能保证存在。

如果输入数据量N（或者用k表示）大于VC维d_{VC}，则有k一定是假设空间H的突破点。

使用VC维d_{VC}对公式7-1进行重写，在$N \geq 2$ 且$d_{VC}ge2$时，如公式7-4所示。

$$m_H(N) \leq N^{d_{VC}}$$

对第五章中提到的几种分类，使用VC维取代突破点，表示VC维与成长函数的关系，如表7-1所示。

表7-1 VC维与成长函数的关系

正射线	$d_{VC}=1$	$m_H(N)=N+1$
一维空间的感知器	$d_{VC}=2$	$m_H(N)=2N$
间隔为正的分类	$d_{VC}=2$	$m_H(N)=\frac{1}{2}N^2+\frac{1}{2}N+1$
凸图形分类	$d_{VC}=\infty$	$m_H(N)=2^N$
二维平面的感知器	$d_{VC}=3$	$m_H(N)<N^3$ 在$N\geq 3$ 时

对上述条件1中好的假设空间重新做一个定义，即有限的VC维d_{VC}。

一个有限的VC维总是能够保证寻找到的近似假设g满足$E_{in}(g) \approx E_{out}(g)$，这一结论与下述部分没有关系：

1. 使用的算法A，即使很大$E_{in}(g)$，也依然能满足上述的性质；
2. 输入数据的分布P；
3. 未知的目标函数f。

即VC维可应对任意的假设空间，任意的数据分布情况，任意的目标函数。

满足这一性质可以得到如图7-2所示的流程图，其中灰色的部分表示上述几个不会影响$E_{in}(g) \approx E_{out}(g)$这一结果的部分。

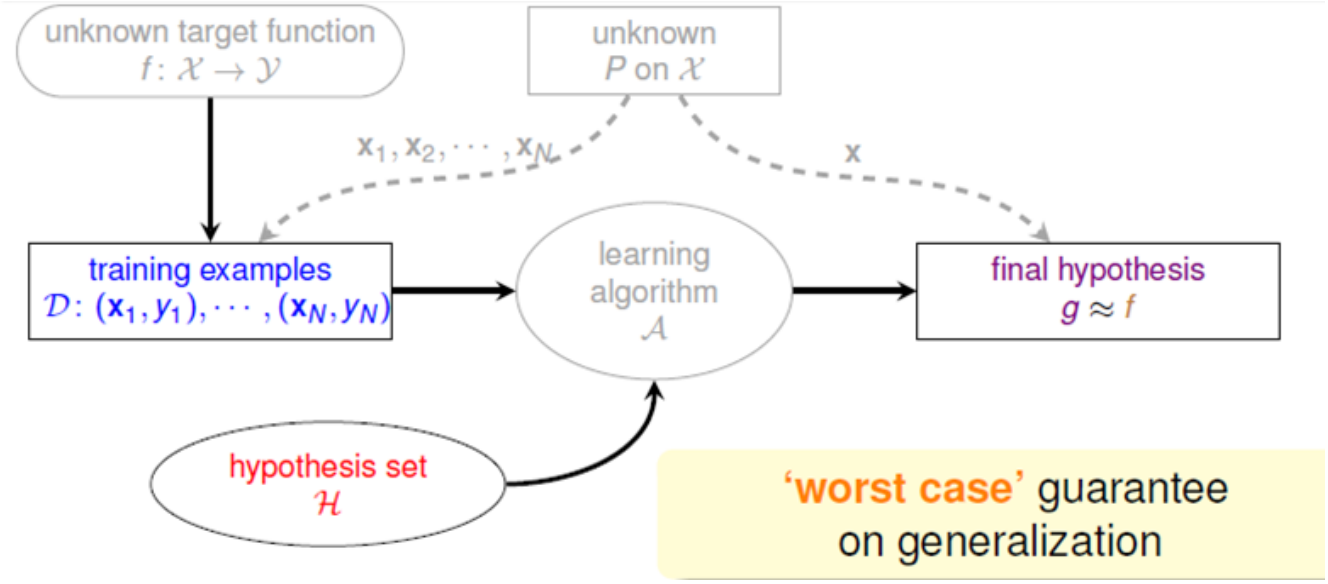

图7-2 由VC维保证机器可以学习的流程图

7.2 VC Dimension of Perceptrons 感知器的VC纬

以下两个条件保证了2维线性可分的数据是可以学习的。

1. 线性可分的数据通过PLA算法运行足够长的时间（T步骤足够大），则会找出一条可以正确分类的直线，使得样本中没有产生分错类的情况，即$E_{in}(g) = 0$;
2. 在训练样本和整个数据集都服从同一分布P的前提下，有VC限制保证了，在且训练样本N足够大时，$E_{in}(g) \approx E_{out}(g)$。

以上两个条件共同得出$E_out(g) \approx 0$的结论。

这一节讨论的是PLA能否处理维数大于二维的数据。

从上一节的内容得知：只要求出d_{VC}是一个有限数，则可以使用VC限制来保证$E_{in}(g) \approx E_{out}(g)$。于是问题变成了在维数大于二维时，感知器的VC维$d_{VC}$如何表示（能否表示成一个有限数）。

两种已知感知器的VC维表示。1维感知器的VC维：$d_{VC} = 2$；2维感知器的VC维：$d_{VC} = 3$。

能否以此类推得出d维感知器的VC维：$d_{VC} = d + 1$呢？

上述只是一种猜想，接下来是对此猜想进行证明，证明的思路也很传统，证明等于号的成立分为两步：证明大于等于d_{VC}以及小于等于$d_{VC} \leq d + 1$。

证明大于等于的思路：证明存在d+1数量的某一数据集可以完全二分；证明小于等于的思路：证明任何d+2数量的数据集都不可以完全二分。

首先证明大于等于。因为只需要证明存在，不妨构造一个输入样本集，假设样本为一个行向量，其中第一个样本为0向量，第二个样本是其第一个分量为1其他分量为0的向量，第三个样本是其第二个分量为1其他分量为0的向量，以此类推，第d+1个样本是其第d个分量为1其他分量为0的向量，如：，，，...，，在感知器中样本X如公式7-5所示，其中每一个样本加上默认的第0个分量，其值为1（从阈值b变成w_0所乘的样本分量）。

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & X_{d+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & & \ddots & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

很容易证明该矩阵式可逆：除了第一行之外的每一行都减去第一行得到一个对角矩阵，因此为满秩，可逆。

需要证明的是可以完全二分，关注的重点为输出标记向量 $y_T = []y_1, y_2, y_3, \cdots, y_{d+1}]$。只要找出各种权值向量$W$ 能将上述输入样本集X映射到全部的二分情况下y的上就可以了。

需要证明的是可以完全二分，关注的重点为输出标记向量 。只要找出各种权值向量
W能将上述输入样本集X映射到全部的二分情况下的y$上就可以了。

已知感知器可以使用$sign(Xw) = y$ 表示。而只要权值向量使得$Xw = y$成立，就一定满足$sign(Xw) = y$ 的需求。假设其中输入矩阵如公式7-5所示，即X是可逆矩阵，输出向量y的任意一种二分类情况都可以被一个假设函数w划分出，原因是权值向量 w 满足 $w = X^{-1}y$ ，即任何一种二分类情况都会有一个权向量w与之对应，因此满足$d_{vc} \geq d + 1$成立。

证明小于等于的思路：证明小于等于是不能如上，举一个特殊输入数据集，因为其要证明在所有的情况下，证明过程稍微复杂一些，先以一个2维空间的例子为切入点。

假设一个2维空间，则需要观察在2+2个输入数据量，不妨假设这四个输入样本分别是$x_1 = [0, 0]$，$x_2 = [1, 0]$，$x_3 = [0, 1]$，$x_4 = [1, 1]$。输入数据集X如公式7-6所示。

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

可以想象在标记 y_1为-1，y_2与y_3为+1时，不可以为-1，如图7-3所示。

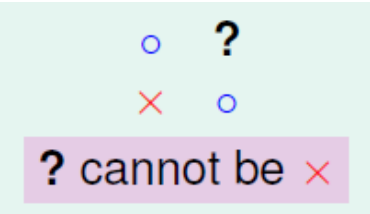

图7-3 2维数据样本不可二分的情况

用数学的形式如何表达？首先根据这四个样本可以确保公式7-7成立。 $x_4 = x_3 + x_2 - x_1$

该公式等号两边同时左乘权值向量w依旧成立，但在满足 $y_1 = -1$，y_2与y_3为+1这一条件时，公式左边一定大于0，如公式7-8所示。 $wx_4 = wx_3 + wx_2 - wx_1 > 0$ 其中，$wx_3 = 1, wx_2 = -1, wx_1 = -1$

这种样本间线性依赖（linear dependence）的关系导致了无法二分。

那在高维空间结果如何呢？

假设在d维空间中d+2个样本，其输入样本集如公式7-9所示。 $X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \vdots & \vdots \\ 1 & X_{d+2} \end{bmatrix}$

因为样本间的线性依赖关系，一定可以得到公式$7-10$，其中$$a_i(i = 1, 2, \cdots, d+2)$$表示系数，该系数可正可负，也可以等于0，但是不可以全为0.

$$x_4 = a_1X_1 + a_2X_2 + \cdots + a_{d+1}X_{d+1}$$

此处使用反证法：设存在一个这样的二分情况，$$y^T = [sign(a_1), sign(a_2), \cdots, sign(a_{d+1}), -1]$$，在公式$7-10$的等号两边左乘权值向量$$W$$得公式$7-11$。

w^TX*{d+2}* = *a*1w^Tx*1* + *a*2w^Tx*2* + \cdots + *a*{d+1}w^Tx_{d+1} > 0

P_D[|E*{in}(g)* - *E*{out}(g) > \varepsilon |] \le 4(2N)^{d_{VC}}exp(- \frac{1}{8} \varepsilon^2 N)

不等式的右边用符号δ表示，则好事情$|E_{in}(g) - E_{out}(g) \leq \varepsilon|$发生的几率一定大于等于$1 - \delta$，$E_{in}(g)$与$E_{out}(g)$接近程度可以使用含有$delta$的公式表示，如公式$7 - 13$所示。

\delta = 4(2N)^{d_{VC}}exp(-frac{1}{8} \varepsilon^2 N) \frac{\delta}{4(2N)^{d_{VC}}} = exp(-\frac{1}{8} \varepsilon^2N)

\frac{4(2N)^{d_{VC}}}{\delta} = exp(\frac{1}{8} \varepsilon^2N)
$$

$$\ln(\frac{4(2N)^{d_{VC}}}{\delta}) = \frac{1}{8}\varepsilon^2 N$$

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

其中与接近程度，即被称为泛化误差（$generalizationerror$），通过上式证明该误差小于等于

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

因此$E_{out}(g)$的范围公式可以用$7 - 14$表示

E{in}(g) - \sqrt{\frac{8}{N} \ln(\frac {4(2N)^{d{VC}}}{\delta})} \le E{out(g)} \le E{in}(g) + \sqrt{\frac{8}{N} \ln(\frac {4(2N)^{d_{VC}}}{\delta})} $$

其中最左边的公式一般不去关心，重点是右边的限制，表示错误的上限。

$\sqrt{\frac{8}{N}\ln(\frac{4(2N)^{d_{VC}}}{\delta})}$又被写成函数$\Omega(N, H, \delta)$称为模型复杂度（model complexity）。

通过一个图表来观察VC维到底给出了哪些重要信息，如图7-5所示。

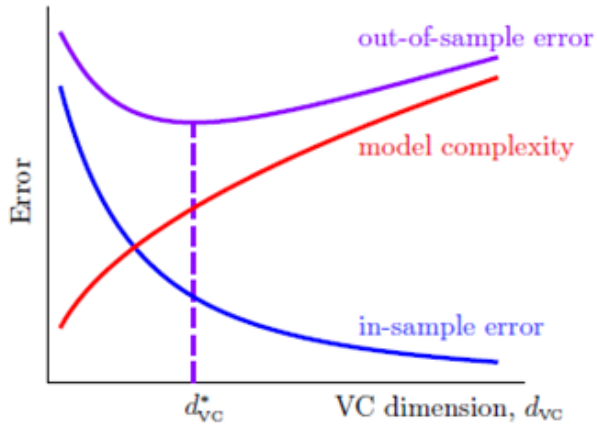

图 7-5 错误率与VC维的关系

其中蓝色线表示E_{in}随VC维d_{VC}的变化；红色部分为模型复杂度$\Omega(N, H, \delta)$随d_{VC}的变化；紫色部分为E_{out}随VC维d_{VC}的变化，其中E_{out}可以表示用E_{in}与$\Omega(N, H, \delta)$表示成的$E_{out} < E_{in} + \sqrt{\frac{8}{N}\ln(\frac{4(2N)^{d_{VC}}}{\delta})}$形式。

其中E_{in}随着d_{VC}的增加而减小，不难理解，因为越大可以选择的假设空间就越大，就有可能选择到更小的E_{in}；模型复杂度$\Omega(N, H, \delta)$随d_{VC}的增加而增加也不难理解，从$\Omega(N, H, \delta) = \sqrt{\frac{8}{N}\ln(\frac{4(2N)^{d_{VC}}}{\delta})}$就能得出关系；而$E_{out}$因为这前两者的求和，因此出现了一个先降低后增加的过程，使其最小的取值为d_{VC}^*。使得E_{out}最小才是学习的最终目的，因此寻找d_{VC}^*很重要。

VC维除了表示模型复杂度之外还可以表示样本的复杂度（sample complexity）。

假设给定需求$\varepsilon = 0.1$，$\delta = 0.1$，$d_{VC} = 3$，求N为多少时，即输入样本多大时才可以满足这些条件。我们将N的各个数量级代入公式$4(2N)^{d_{VC}} \exp(-\frac{1}{8}\varepsilon^2 N)$，与$delta = 0.1$作比较，得到图7-6。

N	bound
100	2.82×10^{7}
1,000	9.17×10^{9}
10,000	1.19×10^{8}
100,000	1.65×10^{-38}
29,300	9.99×10^{-2}

图7-6 N的取值与的关系

从图中可以得出在N差不多2万9千时，才可以训练出满足条件的模型，这是一个很大的数字，即数据的复杂度和VC维存在一个理论上的关系，$N \approx 10000 d_{VC}$ 。但在实际应用中，倍数远比这小的多，大概是$N \approx 10 d_{VC}$。造成这一现象的原因使自VC限制是个非常宽松的约束。宽松的原因主要来自以下四个：

1. 霍夫丁不需要知道未知的E_{out}，VC限制可以用于各种分布，各种目标函数；
2. 成长函数$m_H(N)$取代真正的二分类个数本身就是一个宽松的上界，VC限制可以用于各种数据样本；
3. 使用二项式$N^{d_{VC}}$作为成长函数的上界使得约束更加宽松，VC限制可以用于任意具有相同VC维d_{VC}的假设空间；
4. 联合限制（union bound）并不是一定会选择出现不好事情的假设函数，VC限制可以用于任意算法。

其实很难找出在这四个方面都可以任意选择且比VC限制约束更紧的限制了，了解VC限制的重点其实也并不是它的约束宽松与否，而是它给予我们的哲学启示。