VC维相关证明

毛润泽

September 18, 2016

这篇报告主要尝试证明或推导第四、五、六章中出现的一些数学方法，以期获得更深刻的理解。

首先是Hoeffding不等式。虽然没有学习过这个式子，但按照我的理解，它给出了随机变量与其期望的定量关系，即一个随机变量以不低于$1-\delta$的概率接近其期望值。

$$\text{Hoeffding Inequality:}\quad \text{设}x_1, x_2, \ldots, x_N\text{是}N\text{个随机变量，}\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

$$\text{且}x_i \in [a, b], i = 1, 2, \ldots, N.$$

$$\text{则}\forall\epsilon \in (0, 1):\ P(\bar{x} - \mathbb{E}(\bar{x}) > \epsilon) \le exp(-2\frac{N\epsilon^2}{(b-a)^2}). \tag{1}$$

接下来证明(1)式（如果想跳过这一部分证明，可以直接从第四页的(4)式处开始阅读）：

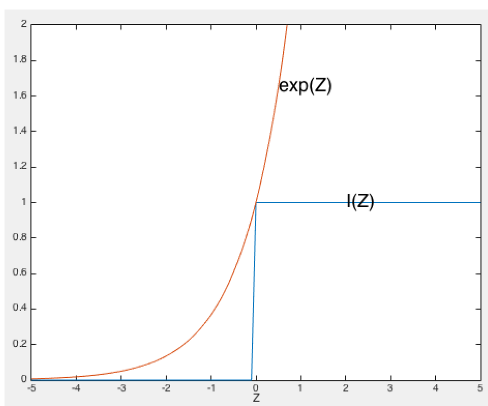

Figure 1: 指示函数与指数函数

首先根据$P(Z) \equiv \mathbb{E}[\mathbb{I}(Z)]$，利用指示函数(indicator function)将$P(\bar{x} - \mathbb{E}(\bar{x}) > \epsilon)$转化为$\mathbb{E}[\mathbb{I}(\bar{x} - \mathbb{E}(\bar{x}) - \epsilon > 0)]$. 其次， 如Figure 1所示，指示函数总是被指数函数bound住，所以有:

$$\mathbb{I}(Z > 0) \le e^{\eta Z} \qquad (\eta > 0)$$

$$
\begin{aligned}
\therefore P(\bar{x}-\mathbb{E}(\bar{x})>\epsilon) &= \mathbb{E}(\mathbb{I}(\bar{x}-\mathbb{E}(\bar{x})-\epsilon>0)) \\
&= \mathbb{E}(\mathbb{I}(N\bar{x}-N\mathbb{E}(\bar{x})-N\epsilon>0)) \\
&\leq \mathbb{E}(exp(\eta(N\bar{x}-N\mathbb{E}(\bar{x})-N\epsilon))) \\
&= e^{-\eta N\epsilon}\cdot\mathbb{E}(exp(\eta(N\bar{x}-N\mathbb{E}(\bar{x})))) \\
&= e^{-\eta N\epsilon}\cdot\mathbb{E}\Big(exp\Big(\eta\big(\sum_{i=1}^{N}x_i-\sum_{i=1}^{N}\mathbb{E}(x_i)\big)\Big)\Big) \\
&= e^{-\eta N\epsilon}\cdot\mathbb{E}\Big(\prod_{i=1}^{N}exp\big(\eta x_i-\eta\mathbb{E}(x_i)\big)\Big) \\
&= e^{-\eta N\epsilon}\cdot\prod_{i=1}^{N}\mathbb{E}\Big(exp\big(\eta x_i-\eta\mathbb{E}(x_i)\big)\Big) \qquad (2)
\end{aligned}
$$

在(2)式中，$\mathbb{E}\Big(exp\big(\eta x_i-\eta\mathbb{E}(x_i)\big)\Big)$可被bound住。因为形如$f(x)=e^x$的函数是下凸函数，所以有:

$$\forall x\in[a,b]:\quad f(x)\leq\frac{b-x}{b-a}f(a)+\frac{x-a}{b-a}f(b).$$

这一点根据向量的三点共线定理不难证明。

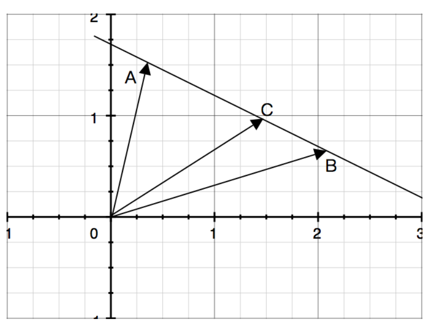

Figure 2: 向量三点共线

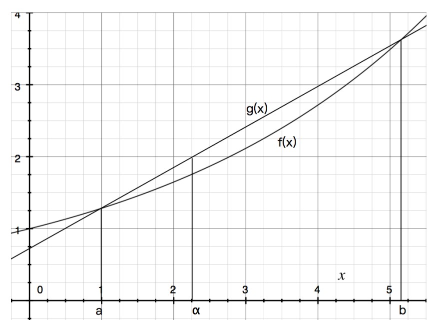

Figure 3: 下凸函数性质

如Figure 2所示，若A,B,C三点共线，且$\frac{\overrightarrow{AC}}{\overrightarrow{CB}}=\frac{\lambda}{1-\lambda}$，则$\overrightarrow{OC}=(1-\lambda)\overrightarrow{OA}+\lambda\overrightarrow{OB}$. 推广到任意维度均适用。如Figure 3所示，有:

$$
\begin{aligned}
\forall\alpha\in[a,b],\ f(\alpha)\leq g(\alpha) &= (1-\lambda)\cdot g(a)+\lambda\cdot g(b) \\
&= \frac{b-\alpha}{b-a}f(a)+\frac{\alpha-a}{b-a}f(b)
\end{aligned}
$$

$$
\begin{aligned}
\therefore exp(\eta x_i - \eta \mathbb{E}(x_i)) &= e^{-\eta \mathbb{E}(x_i)} \cdot e^{\eta x_i} \\
&\le e^{-\eta \mathbb{E}(x_i)} \cdot \left[\frac{b - x_i}{b-a} e^{\eta a} + \frac{x_i - a}{b-a} e^{\eta b} \right]
\end{aligned}
$$

对上面的不等式两边同时取期望，即得到

$$
\begin{aligned}
\mathbb{E}(epx(\eta x_i - \eta \mathbb{E}(x_i))) &\le e^{-\eta \mathbb{E}(x_i)} \cdot \left[\frac{b - \mathbb{E}(x_i)}{b-a} e^{\eta a} + \frac{\mathbb{E}(x_i) - a}{b-a} e^{\eta b} \right] \\
&= e^{-\eta \left(\mathbb{E}(x_i) - a \right)} \cdot \left[\frac{b - a - \left(\mathbb{E}(x_i) - a \right)}{b-a} + \frac{\mathbb{E}(x_i) - a}{b-a} e^{\eta(b-a)} \right] \\
&= e^{-\eta \left(\mathbb{E}(x_i) - a \right)} \cdot \left[1 - \frac{\mathbb{E}(x_i) - a}{b-a} + \frac{\mathbb{E}(x_i) - a}{b-a} e^{\eta(b-a)} \right] \qquad (3)
\end{aligned}
$$

令$\begin{cases} p = b - a \\ q = \frac{\mathbb{E}(x_i) - a}{b-a} \end{cases}$，则 $p \cdot q = \mathbb{E}(x_i) - a$. 明显$p, q$都是常数，于是有:

$$
\begin{aligned}
(3)\text{式} &= e^{-pq\eta} \cdot (1 - q + q e^{p\eta}) \\
&= exp\Big(-pq\eta + ln\big(1 - q + q e^{p\eta}\big) \Big)
\end{aligned}
$$

定义$f(\eta) = -pq\eta + ln(1 - q + q e^{p\eta})$, 于是有:

$$
\begin{aligned}
f'(\eta) &= -pq + \frac{pq e^{p\eta}}{1 - q + q e^{p\eta}} \\
&= -pq + \frac{pq}{(1-q) e^{-p\eta} + q} \\
\therefore f''(\eta) &= p^2 \cdot \frac{q(1-q) e^{-p\eta}}{\Big((1-q) e^{-p\eta} + q \Big)^2}
\end{aligned}
$$

根据均值不等式“调几算方”的定理，几何平均数不大于算术平均数，所以:

$$
\sqrt{q \cdot (1-q) e^{-p\eta}} \le \frac{q + (1-q) e^{-p\eta}}{2}, \quad \text{即} \ \frac{q(1-q) e^{-p\eta}}{\Big((1-q) e^{-p\eta} + q \Big)^2} \le \frac{1}{4}
$$

$$
\therefore f''(\eta) \le \frac{p^2}{4} = \frac{(b-a)^2}{4}
$$

将$f(\eta)$在$\eta = 0$处进行泰勒展开，得到:

$$
f(\eta) = f(0) + f'(0) \cdot \eta + \frac{f''(\xi)}{2} \cdot \eta^2
$$

根据$\begin{cases} f(0) &= 0 \\ f'(0) &= 0 \\ f''(\eta)_{max} &= \frac{(b-a)^2}{4} \end{cases}$, 得到$f(\eta) \leq \frac{\eta^2(b-a)^2}{8}$, 从而得到:

$$(3)\text{式} \leq e^{\eta^2(b-a)^2/8}$$

$$\begin{aligned} \therefore (2)\text{式} &\leq e^{-\eta N\epsilon} \cdot \prod_{i=1}^{N} e^{\eta^2(b-a)^2/8} \\ &= exp\Big(-\eta N\epsilon + \frac{N}{8}\eta^2\big(b-a\big)^2\Big) \qquad (\ \forall \eta > 0\ ) \end{aligned}$$

设$g(\eta) = \frac{N}{8}(b-a)^2\eta^2 - N\epsilon\eta$, 这是一个关于$\eta$的二次函数:

$$\text{当}\eta = \frac{4\epsilon}{(b-a)^2}\text{时，} g(\eta)_{min} = -\frac{2N\epsilon^2}{(b-a)^2}$$

$$\text{于是，} P(\bar{x} - \mathbb{E}(\bar{x}) > \epsilon) \leq exp\big(-2\frac{N\epsilon^2}{(b-a)^2}\big) \text{ 得证。}$$

对于假设空间$\mathcal{H}$中某个假设h, 和训练集$\mathcal{D}$中的数据(x_i, y_i), 记 $\xi_i = \mathbb{I}_{x_i\in\mathcal{D}}(y_i \neq h(x_i))$, 则$\bar{\xi} = \mathbb{E}(\mathbb{I}_{x\in\mathcal{D}}(y \neq h(x))) = P_{x\in\mathcal{D}}(y \neq h(x)) = E_{in}$, $\therefore \mathbb{E}(\bar{\xi}) = \mathbb{E}(E_{in}) = E_{out}$. 又因为$\xi_i \in \{0,1\}$, 所以$\begin{cases} a = 0 \\ b = 1 \end{cases}$. 代入到Hoeffding不等式中，得到:

$$P(E_{in} - E_{out} > \epsilon) = P(\bar{\xi} - \mathbb{E}(\bar{\xi}) > \epsilon) \leq exp(-2N\epsilon^2)$$

$$\begin{aligned} \therefore P(|E_{in} - E_{out}| > \epsilon) &= P(E_{in} - E_{out} > \epsilon) + P(-E_{in} + E_{out} > \epsilon) \\ &\leq 2exp(-2N\epsilon^2) \end{aligned} \tag{4}$$

定义$|E_{in} - E_{out}| > \epsilon$为$\mathcal{BAD}$, 则对于某个确定的假设$h$, 有:

$$P(h\ is\ \mathcal{BAD}) \leq 2exp(-2N\epsilon^2)$$

$$\begin{aligned} \therefore P(\mathcal{H}\ is\ \mathcal{BAD}) &= P(\exists h \in \mathcal{H}:\ h\ is\ \mathcal{BAD}) \\ &= P\Big(\bigcup_{h_i\in\mathcal{H}} \big(h_i\ is\ \mathcal{BAD}\big)\Big) \\ &\leq \sum_{h_i\in\mathcal{H}} P(h_i\ is\ \mathcal{BAD}) \end{aligned} \tag{5}$$

为方便起见，以下所有讨论均以二分类问题为例。定义$\mathcal{H}$中的假设对$\mathcal{D}$中示例赋予标记的每种可能结果为对$\mathcal{D}$的一种“对分”(dichotomy). 比如$\mathcal{D}$有三个示例x_1, x_2, x_3, 则$\{(1, -1, -1)\}$是一种对分, $\{(1, -1, 1)\}$也是一种对分。

显然，若两个不同的假设h_i, h_j, 在整个$\mathcal{X}$上实现了相同的对分, 则它们具有相同的E_{in}和E_{out}, 那么$(h_i\ is\ \mathcal{BAD})$与$(h_j\ is\ \mathcal{BAD})$就是完全对等的事件。将这样实现相同对分的h视为同一个，(5)式就有可能是收敛的。

但事实上即使这样，由于样本空间$\mathcal{X}$是无限大的，在$\mathcal{X}$上能实现的对分也是无穷多种可能的，相应的E_{in}和E_{out}也有无穷多种，所以(5)式仍是趋向于无穷的。因此需要用另一个有限的数据集代替无限大的$\mathcal{X}$, 以得到有限数量的对分，使得(5)式收敛。

从$(\mathcal{X}-\mathcal{D})$中再随机抽取N个独立同分布的点，组成$\mathcal{D}'$。这N个点是隐形的、未知的，只是存在于理论分析中，实际操作时并不真正需要它们，所以又称ghost sample。这样的话，一个假设h除了在$\mathcal{D}$上有经验误差E_{in}，还会在$\mathcal{D}'$上有另一个误差E'_{in}，用E'_{in}来代替原来的E_{out}。由于$|\mathcal{D}+\mathcal{D}'|=2N$，所以最多可能有$2^{2N}$种对分。将具有相同对分的$h$视为同一个，也就最多有$2^{2N}$个$h$，那么(5)式就可以收敛了。

看到这里，也许你要说，这是在自欺欺人。机器学习的目标是找到一个假设h，它的E_{in}和E_{out}相差不会很大。但现在用E'_{in}来代替E_{out}，我们找到的h，只能满足E_{in}和E'_{in}相差不大。但别忘了，$\mathcal{D}'$是从$\mathcal{X}$中独立同分布地采样出来的，所以它在一定程度上可以代表$\mathcal{X}$。E_{in}和E'_{in}相差不大，说不定就能保证，E_{in}和E_{out}相差也不会太大。事实上，我们有如下结论:

$$P\Big[\exists h\in\mathcal{H}\ s.t.\ |E_{in}(h)-E_{out}(h)|>\epsilon\Big]\le 2P\Big[\exists h\in\mathcal{H}\ s.t.\ |E_{in}(h)-E'_{in}(h)|>\frac{\epsilon}{2}\Big] \tag{6}$$

接下来证明(6)式。

要满足$\exists h\in\mathcal{H}\ s.t.\ |E_{in}(h)-E_{out}(h)|>\epsilon$，只要让$|E_{in}(h)-E_{out}(h)|$的上确界大于$\epsilon$即可。所以可以记$P\Big[\exists h\in\mathcal{H}\ s.t.\ |E_{in}(h)-E_{out}(h)|>\epsilon\Big]$为$P\Big[\sup\limits_{h\in\mathcal{H}}(|E_{in}-E_{out}|)>\epsilon\Big]$。同样的，记$P\Big[\exists h\in\mathcal{H}\ s.t.\ |E_{in}(h)-E'_{in}(h)|>\frac{\epsilon}{2}\Big]$为$P\Big[\sup\limits_{h\in\mathcal{H}}(|E_{in}-E'_{in}|)>\frac{\epsilon}{2}\Big]$。

$$\begin{aligned}P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E'_{in}|)>\frac{\epsilon}{2}\Big]&\ge P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E'_{in}|)>\frac{\epsilon}{2}\bigcap\sup_{h\in\mathcal{H}}(|E_{in}-E_{out}|)>\epsilon\Big]\\&=P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E_{out}|)>\epsilon\Big]\times\\&\quad P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E'_{in}|)>\frac{\epsilon}{2}\ \Big|\ \sup_{h\in\mathcal{H}}(|E_{in}-E_{out}|)>\epsilon\Big]\end{aligned} \tag{7}$$

假设$|E_{in}-E_{out}|$的上确界在$h=h^*$处取到，且满足$|E_{in}(h^*)-E_{out}(h^*)|>\epsilon$. 则:

$$\begin{aligned}&P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E'_{in}|)>\frac{\epsilon}{2}\ \Big|\ \sup_{h\in\mathcal{H}}(|E_{in}-E_{out}|)>\epsilon\Big]\\\ge&P\Big(|E_{in}(h^*)-E'_{in}(h^*)|>\frac{\epsilon}{2}\ \Big|\ |E_{in}(h^*)-E_{out}(h^*)|>\epsilon\Big)\end{aligned} \tag{8}$$

分别记$|E_{in}(h^*)-E_{out}(h^*)|>\epsilon$为事件A，$|E'_{in}(h^*)-E_{out}(h^*)|<\frac{\epsilon}{2}$为事件B，$|E_{in}(h^*)-E'_{in}(h^*)|>\frac{\epsilon}{2}$为事件C。那么已知A、B的情况下，可以得出C:

$$\begin{aligned}|E_{in}(h^*)-E'_{in}(h^*)|&=\Big|\big[E_{in}(h^*)-E_{out}(h^*)\big]-\big[E'_{in}(h^*)-E_{out}(h^*)\big]\Big|\\&\ge\Big|E_{in}(h^*)-E_{out}(h^*)\Big|-\Big|E'_{in}(h^*)-E_{out}(h^*)\Big|\\&\ge\epsilon-\frac{\epsilon}{2}\\&=\frac{\epsilon}{2}\end{aligned}$$

$$\begin{aligned}
&\therefore A\cap B\subseteq C\\
\text{又}\ &\because A\cap B\subseteq A\\
&\therefore A\cap B\subseteq A\cap C\\
&\therefore P(A\cap B)\le P(A\cap C)
\end{aligned}$$

上面的不等式两边同除以$P(A)$，根据贝叶斯公式得到:

$$P(B|A)\le P(C|A)$$

$$\therefore (8)\text{式}\ge P\Big(|E'_{in}(h^*)-E_{out}(h^*)|<\frac{\epsilon}{2}\ \Big|\ |E_{in}(h^*)-E_{out}(h^*)|>\epsilon\Big)$$

h^*的选择仅与$\mathcal{D}$有关，而与$\mathcal{D}'$无关。所以$|E'_{in}(h^*)-E_{out}(h^*)|<\dfrac{\epsilon}{2}$和$|E_{in}(h^*)-E_{out}(h^*)|>\epsilon$是互相独立的事件。

$$\begin{aligned}
\therefore (8)\text{式}&\ge P\Big(|E'_{in}(h^*)-E_{out}(h^*)|<\frac{\epsilon}{2}\ \Big|\ |E_{in}(h^*)-E_{out}(h^*)|>\epsilon\Big)\\
&=P\Big(|E'_{in}(h^*)-E_{out}(h^*)|<\frac{\epsilon}{2}\Big)\\
&=1-P\Big(|E'_{in}(h^*)-E_{out}(h^*)|\ge\frac{\epsilon}{2}\Big)\\
&\ge 1-2exp\Big(-2N\cdot(\frac{\epsilon}{2})^2\Big)\\
&=1-2e^{-N\epsilon^2/2}
\end{aligned}$$

实际中，一般都会要求$N\epsilon^2>4$。所以:

$$(8)\text{式}=1-2e^{-N\epsilon^2/2}>\frac{1}{2}$$

代入到(7)式，得到:

$$P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E_{out}|)>\epsilon\Big]\le 2P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E'_{in}|)>\frac{\epsilon}{2}\Big]$$

(6)式得证。这样，$\mathcal{H}$能实现的对分就是有限多种的。

为了应用Hoeffding不等式，要将E'_{in}换成E_{in}的期望——本来是E_{out}，现在则是$\frac{E_{in}+E'_{in}}{2}$。E_{in}和E'_{in}相差$\dfrac{\epsilon}{2}$的话，我们可以认为E_{in}和$\frac{E_{in}+E'_{in}}{2}$相差$\dfrac{\epsilon}{4}$。

$$\begin{aligned}
\therefore P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E_{out}|)>\epsilon\Big]&\le 2P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E'_{in}|)>\frac{\epsilon}{2}\Big]\\
&=2P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-\frac{E_{in}+E'_{in}}{2}|)>\frac{\epsilon}{4}\Big]\\
&\le 2\cdot 2|\mathcal{H}|exp\Big(-2N\cdot(\frac{\epsilon}{4})^2\Big)\\
&=4|\mathcal{H}|exp(-\frac{1}{8}N\epsilon^2)
\end{aligned}\tag{9}$$

建立$\mathcal{D}+\mathcal{D}'\to\{0,1\}$的映射，共有$2^{2N}$种映射，所以假设空间$\mathcal{H}$最多产生$2^{2N}$种对分。所以:

$$
\begin{aligned}
(9)\text{式} &\le 4\cdot 2^{2N}\cdot exp(-\frac{1}{8}N\epsilon^2)\\
&= exp\big[(2N+2)ln2-\frac{1}{8}N\epsilon^2\big]
\end{aligned}
$$

很明显，上式的指数部分$(2N+2)ln2-\dfrac{1}{8}N\epsilon^2$在$\epsilon=1$时可以取到最小值，仍然是大于0的。所以上式大于1，这个upper bound没有意义。究其原因，是因为将$|\mathcal{H}|$过高地估计为指数形式。我们希望能够找到一个关于N的多项式，来做一个更准确的上界评估。

方法如下。

定义关于假设空间$\mathcal{H}$和$m\in\mathbb{N}$的增长函数(Growth Function)$G_{\mathcal{H}}(m)$为:

$$
G_{\mathcal{H}}(m)=\max_{x_1,x_2,\ldots,x_m\in\mathcal{X}}\Big|\big\{\big(h(x_1),h(x_2),\ldots,h(x_m)\big)\mid h\in\mathcal{H}\big\}\Big|
$$

增长函数代表了假设空间$\mathcal{H}$对m个点，最多有多少种赋予标记的方式，即对分的方式。以2D-perceptron为例:

$$
\begin{aligned}
G_{\mathcal{H}}(1)&=2\\
G_{\mathcal{H}}(2)&=4\\
G_{\mathcal{H}}(3)&=8\\
&\cdots\cdots
\end{aligned}
$$

看似$G_{\mathcal{H}}(m)=2^m$，但事实上当$m=4$时，2D-perceptron无法实现$2^4=16$种对分。这可以通过画图来证明。对于平面上的4个点，无论它们怎么分布，都无法画直线来实现16种对分方式，最多只有14种。所以我们称$m=4$是2D-perceptron的break point，而$m=3$则称为它的VC维。

对于给定的N，如果存在某一种情形下的N个点，使得$\mathcal{H}$可以实现所有的对分，那么称$\mathcal{H}$ shatter了N。于是可以定义：某个假设空间$\mathcal{H}$的break point是指一个最小的N，使得任意的N个点都无法被该假设空间shatter；而VC维是一个最大的N，使得存在N个点，可以被该假设空间shatter。

对于假设空间$\mathcal{H}$,定义它的VC维为：$\mathcal{VC}(\mathcal{H})=\max\{m\mid G_{\mathcal{H}}(m)=2^m\}$，break point为：$k=\min\{m\mid G_{\mathcal{H}}(m)<2^m\}$，两者关系为：$k=\mathcal{VC}(\mathcal{H})+1$.

很明显，如果$\mathcal{H}$的VC维存在，则有：$\forall m>\mathcal{VC}(\mathcal{H})$, $G_{\mathcal{H}}(m)<2^m$。所以，当N足够大时，也许就可以得到：$G_{\mathcal{H}}(N)\ll 2^N$。

问题就在于，如何确定$G_{\mathcal{H}}(m)$的表达式，或者说，如何证明$G_{\mathcal{H}}(m)$是一个关于m的多项式。为了更具一般性地讨论，我们不研究特定的假设空间$\mathcal{H}$，而是以break point或VC维来代表某一类假设空间。定义:

$$
B(m,k)=\max\{G_{\mathcal{H}}(m)\mid \text{the break point of } \mathcal{H} \text{ is k}\}
$$

也就是说，$B(m,k)$表示了break point为k的假设空间，对m个点最多有多少种对分方式。根据以上对于break point的定义，我们可以得到:

$$
B(m,k)=\begin{cases}2^m & m<k\\ 2^m-1 & m=k\\ ? & m>k\end{cases}
$$

当$m > k$时，假设$B(m,k) = 2a + b$。其中a是指$B(m,k)$所有对分方式中成对出现的对分，所谓的成对，即两种对分，对点$x_1, x_2, \dots, x_{m-1}$赋予了相同标记，而对点x_m分别标记为1和-1。b则是剩下的无对称的对分。

首先证明$a + b \le B(m-1,k)$:

$$
\begin{aligned}
&\because m > k \\
&\therefore \text{任意m个点，不能shatter k个点} \\
&\therefore \text{取}a+b\text{的前}m-1\text{个点，也不能shatter k个点} \\
&\therefore a + b \le B(m-1,k)
\end{aligned}
$$

其次证明$a \le B(m-1,k-1)$:

$$
\begin{aligned}
&\quad \text{假设取}a\text{的前}m-1\text{个点，可以shatter其中k - 1个点} \\
&\because a \text{ 成对} \\
&\therefore \text{加上}x_m\text{，}a\text{可以shatter k个点} \\
&\quad \text{与不能shatter k个点矛盾} \\
&\therefore \text{假设不成立。取}a\text{的前}m-1\text{个点，不能shatter其中k -1个点} \\
&\therefore a \le B(m-1,k-1)
\end{aligned}
$$

综上，我们得到:

$$
\begin{aligned}
B(m,k) &= 2a + b \\
&= (a+b) + b \\
&\le B(m-1,k) + B(m-1,k-1)
\end{aligned}
$$

根据上述结论，可以用数学归纳法证明: $m \ge k$时，$B(m,k) \le \sum_{i=0}^{k-1} \binom{m}{i}$，方法如下。

首先，$n = k = 3$时，$B(3,3) = 7 = \sum_{i=0}^{2} \binom{3}{i}$；$n = 3$，$k = 2$时，$B(3,2) = 4 = \sum_{i=0}^{1} \binom{3}{i}$。假设$m{=}m_0$，$k = k_0$时不等式成立，且$m = m_0$，$k = k_0 - 1$时也成立。则对于$m = m_0 + 1$，$k = k_0$，有:

$$\begin{aligned}
B(m_0+1,k_0) &\le B(m_0,k_0)+B(m_0,k_0-1)\\
&\le \sum_{i=0}^{k_0-1}\binom{m_0}{i}+\sum_{i=0}^{k_0-2}\binom{m_0}{i}\\
&=\binom{m_0}{0}+\sum_{i=0}^{k_0-2}\left[\binom{m_0}{i}+\binom{m_0}{i+1}\right]\\
&=\binom{m_0+1}{0}+\sum_{i=0}^{k_0-2}\binom{m_0+1}{i+1}\\
&=\sum_{i=0}^{k_0-1}\binom{m_0+1}{i}
\end{aligned}$$

所以，对于$m=m_0+1,\ k=k_0$，不等式亦成立。同理可以推出$m=m_0+1,\ k=k_0-1$时，不等式也成立。由这两个条件又可以推导出$m=m_0+2$时的情形，以至于无穷。于是，不等式:

$$B(m,k)\le\sum_{i=0}^{k-1}\binom{m}{i},\quad (m\ge k) \tag{10}$$

得证。

可以证明(10)式右边部分是一个m的多项式。首先，$k-1$就是$\mathcal{H}$的VC维，将它设为d。不等式两边同乘以$(\frac{d}{m})^d$，得到:

$$\begin{aligned}
(\frac{d}{m})^d\cdot B(m,k) &\le (\frac{d}{m})^d\cdot\sum_{i=0}^{d}\binom{m}{i}\\
\therefore (\frac{d}{m})^d\cdot B(m,k) &\le \sum_{i=0}^{d}(\frac{d}{m})^i\binom{m}{i} \qquad (\because \frac{d}{m}<1)\\
&\le \sum_{i=0}^{m}(\frac{d}{m})^i\binom{m}{i}\\
&=(1+\frac{d}{m})^m
\end{aligned}$$

而$(1+\frac{d}{m})^{\frac{m}{d}}<e$，所以$(1+\frac{d}{m})^m<e^d$。故:

$$\begin{aligned}
(\frac{d}{m})^d\cdot B(m,k) &\le e^d\\
\therefore B(m,k) &\le e^d\cdot(\frac{d}{m})^{-d}\\
&=(\frac{e\cdot m}{d})^d
\end{aligned}$$

所以，只要假设空间$\mathcal{H}$的VC维存在，则$B(m,k)$可以被一个关于m的多项式bound住。相应的，$G_{\mathcal{H}}(m)$也被bound住了。将(10)式代入到(9)式中，得到:

$$P\Big[\sup_{h\in\mathcal{H}}(|E_{in}-E_{out}|)>\epsilon\Big]\le 4G_{\mathcal{H}}(2N)exp\big(-\frac{1}{8}N\epsilon^2\big) \tag{11}$$

其中:

$$G_{\mathcal{H}}(2N)\le\sum_{i=0}^{\mathcal{VC}(\mathcal{H})}\binom{2N}{i}$$

想让(11)式的右边部分尽量小，可以增大N，或者增大ϵ。也就是说，想让机器学习得到更好的结果，可以通过增大训练规模和放松约束两种方法来达到。一般来说，似乎增大训练规模是个更加靠谱的选择。N增大时，$G_{\mathcal{H}}(2N)$对于减小右边的式子有阻碍的作用，而$exp(-\frac{1}{8}N\epsilon^2)$则起到促进的作用。但前者是多项式增长的，后者是指数缩小的，所以右半边的式子会渐渐缩小。

事实上，上述定理告诉我们，只要$\mathcal{H}$的VC维存在，那么:

$$\forall\delta\in(0,1),\ \exists N_0\in\mathbb{N},\ s.t.\ \forall N>N_0:\quad P(\mathcal{H}\ is\ \mathcal{BAD})\le\delta$$

所以，机器学习在这种情形下，是可以实现的，或者说是，PAC可学习的。