Notes for Fundations of Data Science

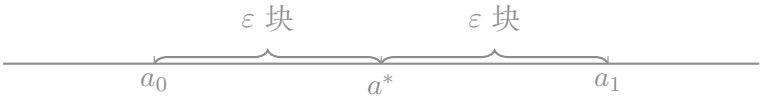
//** 笔记基本根据上课的顺序完成，但是对一些问题做了注释和补充 **//

1 VC 维与半空间

令 $\mathcal{H}$ 是实线上阈值函数构成的集合，即 $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$，其中，$h_a : \mathbb{R} \longrightarrow \{0,1\}$ 是一个函数，使得 $h_a(x) = 1_{[x<a]}$。即若 $x < a$ 则后者取 1，否则取 0。显然 $\mathcal{H}$ 是无限大小的。尽管如此，下面的引理表明 $\mathcal{H}$ 在 PAC 模型下采用 ERM 算法仍是可学习的。

lemma 令 $\mathcal{H}$ 为如之前定义的阈值函数类，那么，$\mathcal{H}$ 在采用 ERM 规则时是 PAC 可学习的，其样本复杂度 $m_{\mathcal{H}}(\varepsilon, \delta) \leq \lceil log(2/\delta)\varepsilon \rceil$。

证明： 令 a^* 为阈值，则相应的假设 $h^*(x) = \mathbf{1}_{[x<a^*]}$ 可以使得 $L_{\mathcal{D}}(h^*) = 0$。令 $\mathcal{D}_x$ 为域 $\mathcal{X}$ 上的边缘分布，令 $a_0 < a^* < a_1$ 使得：$\mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a_0, a^*)] = \mathbb{P}_{x\sim\mathcal{D}_x}[x \in (a^*, a_1)] = \varepsilon$



（如果 $\mathcal{D}_x(-\infty, a^*) \leq \varepsilon$ 则令 $a_0 = -\infty$，对于 a_1 采用类似的处理。）给定一个训练集 S，令 $b_0 = max\{x : (x, 1) \in S\}$，$b_1 = min\{x : (x, 0) \in S\}$（若在 S 中无正样本，令 $b_0 = -\infty$，同样，如果在 S 中无负样本，令 $b_1 = \infty$）。令 b_S 为与 ERM 假设 h_S 相关的阈值，即 $b_S \in (b_0, b_1)$。因此，$L_{\mathcal{D}}(h_S) \leq \varepsilon$ 成立的充分条件是 $b_0 \geq a_0$ 与 $b_1 \leq a_1$ 同时成立，换言之

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \varepsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0 \vee b_1 > a_1]$$

采用联合界，上式变为：

$$\mathbb{P}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S) > \varepsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] + \mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1] \tag{1}$$

当且仅当 S 中所有样本均分布在不同区间 (a_0, a^*) 中时，会出现 $b_0 < a_0$ 的情况，将这种情况出现的概率定义为 ε，即：

$$\mathbb{P}_{S\sim\mathcal{D}^m}[b_0 < a_0] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall(x, y) \in S, x \notin (a_0, a^*)] = (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

由于我们假定了 $m > log(2/\delta)/\varepsilon$，则上式最多为 $\delta/2$，同样，可以容易得到 $\mathbb{P}_{S\sim\mathcal{D}^m}[b_1 > a_1] \leq \delta/2$，联立 (1) 式引理得证！

（限制 $\mathcal{H}$ 在 C 上） 令 $\mathcal{H}$ 是从 $\mathcal{X}$ 到 $\{0,1\}$ 上的一个函数类，并且令 $C = \{c_1, \cdots, c_m\} \subset \mathcal{X}$。限制 $\mathcal{H}$ 在 C 上就是由来自 $\mathcal{H}$ 从 $\mathcal{X}$ 到 $\{0,1\}$ 的函数构成的集合。即：

$$\mathcal{H}_C = \{(h(c_1), \cdots h(c_m)) : h \in \mathcal{H}\}$$

其中，我们将每个从 C 到 $\{0,1\}$ 的函数表示为形如 $\{0,1\}^{|c|}$ 的向量。

如果限制 $\mathcal{H}$ 在 C 上是从 C 到 $\{0,1\}$ 的所有函数的集合，则假设类 $\mathcal{H}$ **打散**了有限集 $C \subset \mathcal{X}$，此时 $|\mathcal{H}_C| = c^{|\mathcal{H}|}$

令 $\mathcal{H}$ 是从 $\mathcal{X}$ 到 $\{0,1\}$ 的函数构成的假设类，令 m 是训练集的大小，假定存在大小为 2m 的集合 $C \subset \mathcal{X}$ 能被 $\mathcal{H}$ 打散，那么，对于任何学习算法 A，在 $\mathcal{X} \times \{0,1\}$ 上必定存在一个分布 $\mathcal{D}$ 和预测器 $h \in \mathcal{H}$ 使得 $L_{\mathcal{D}}(h) = 0$，但是对于所选样本集 $S \sim \mathcal{D}^m$ 至少以 $\frac{1}{7}$ 的概率有 $L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}$。

上述推论说明了如果假设列打散了大小为 $2m$ 的集合 C，那么我们将无法通过 m 个样本来学习 $\mathcal{H}$。从哲学上而言，如果有人可以解释每个现象，他的解释本身是毫无意义的。

VC 维 假设类 $\mathcal{H}$ 的 VC 维，记作 $VCdim(\mathcal{H})$，是 $\mathcal{H}$ 可以打散的最大集合 $C \subset \mathcal{H}$ 的大小。如果 $\mathcal{H}$ 可以打散任意大小的集合，我们说 $\mathcal{H}$ 的 VC 维是无穷的。

定理： 设 $\mathcal{H}$ 是无穷 VC 维的假设类，那么 $\mathcal{H}$ 不是 PAC 可学习的。这是因为如果 $\mathcal{H}$ 有无穷的 VC 维，故对于任意 m 大小的训练集，总存在一个大小为 2m 且被打散的集合，结合前面的推论可证。

> 对于某个实例，希望证明 $VCdim(\mathcal{H}) = d$ 首先需要指出存在一个大小为 d 的集合可以被打散，同时对于任何大小为 d+1 的集合都不能被打散。

下面是一些常见的假设类例子：

–**轴平行矩形** 尽管对于指向全体轴平行矩形的假设类的 VC 维是四——这是满足了最大性，因为对于恰好分布在矩形四个角且被边联系的两点标签相同的情况，没有分隔器能够打散这四个点。同时对于 5 个点分布的任何情况，都存在某个 $\{0,1\}^m$ 不在 $\mathcal{H}$ 中，故 VC 维小于 5。
–**阈值函数** 显然 VC 维为 1。
–**实数区间** 实数线上的区间可以打散任意两个点的集合，但无法打散三个点的集合，因为无法孤立出第一个和最后一个点的子集。因此，区间的 VC-维度为 2。
–**实数成对区间** 考虑一对区间的族，其中一对区间被视为至少属于其中一个区间的点集。存在一个大小为四的集合可以被打散，但没有大小为五的集合可以被打散，因为无法孤立出第一个、第三个和最后一个点的子集。因此，成对区间的 VC-维度为 4。
–**有限实数集合** 有限实数集合系统可以打散任何有限实数集合，因此有限集合的 VC-维度是无限的。
–**凸多边形** 对于任意正整数 n，在单位圆上放置 n 个点。任何点的子集都是凸多边形的顶点。显然，该多边形不包含不在子集中的任何点。这表明凸多边形可以打散任意大的集合，所以其 VC-维度是无限的。

–**d 维半空间**

半空间 半空间类也是一个常见的假设类，它为二分类设计。即 $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{-1, +1\}$。半空间的定义如下：

$$HS_d = sign \circ L_d = \{x \mapsto sign(h_{w,b}(x)) : h_{w,b} \in L_d\}$$

换言之，每个 HS_d 半空间假设均被 $w \in \mathbb{R}^d$ 和 $b \in \mathbb{R}$ 参数化，当输入一个向量 x 时，假设返回一个标签 $sign(< w, x > +b)$。

这里引入两个半空间 ERM 方法的准则。一个是之前已经涉及的感知器算法，另一个是半空间类的线性规划。后者主要是说这样的分隔器的存在性。

不妨这样的问题是齐次情形下。令 $S = \{(x_i, y_i)\}_{i=1}^m$ 为 m 维训练集，假定样本是可分的，训练集上的 ERM 预测是 0 误差，即，我们可以寻找向量 $w \in \mathbb{R}^d$ 满足：

$$sign(< w, x_i >) = y_i \qquad \forall i = 1, \cdots, m$$

等价于：

$$y_i \langle w, x_i \rangle > 0, \qquad \forall i = 1, \cdots, m$$

令 w^* 满足该条件（因为我们可以假定可分，因此一定存在），定义 $\gamma = min_i(y_i \langle w^*, x_i \rangle)$ 并令 $\overline{w} = \frac{w^*}{\gamma}$。因此对于所有的 i，我们有：

$$y_i \langle \overline{w}, x_i \rangle = \frac{1}{\gamma} y_i \langle w^*, x_i \rangle \geq 1$$

存在性证毕。

半空间的 VC 维 1. 齐次半空间 $\mathbb{R}^d$ 的 VC 维是 d。

证明： 考虑向量集合 $e_1, e_2, \cdots, e_d$，其中 e_i 的第 i 个元素为 1，其余元素为 0；这个集合被半空间类打散。显然，对于 $y_1, \cdots, y_d$ 中的每一个标签，给定 $w = (y_1, \cdots, y_d)$ 有 $\langle w, e_i \rangle = y_i (\forall i)$。

而后，令 $x_1, \cdots, x_{d+1}$ 是 $\mathbb{R}^d$ 中的 d+1 个向量的集合，那么，一定有非全部为零的实数 $a_1, \cdots, a_{d+1}$ 满足 $\sum_{i=1}^{d+1} a_i x_i = 0$。令 $I = \{i : a_i > 0\}$ 且 $J = \{j : a_j < 0\}$，I,J 并非全空。首先假设他们均非空，即：

$$\sum_{i \in I} a_i x_i = \sum_{j \in J} |a_j| x_j$$

现在假定 $x_1, \cdots, x_{d+1}$ 被齐次类打散，那么必有向量 w 对于所有的 $i \in I$ 满足 $\langle w, x_i \rangle > 0$，对 $j \in J$ 有 $\langle w, x_j \rangle < 0$，从而：

$$0 < \sum_{i \in I} a_i \langle x_i, w \rangle = \langle \sum_{i \in I} a_i x_i, w \rangle = \sum_{j \in J} |a_j| \langle x_j, w \rangle < 0$$

上式是一个矛盾式，最后如果 J（I）非空，那么上式的右侧（左侧）的不等号也会矛盾。

2. 非齐次半空间 $\mathbb{R}^d$ 的 VC 维是 d+1。

证明： 首先，就像证明上一个定理一样，容易知道向量集合 $0, e_1, \cdots, e_d$ 被非齐次半空间类打散，然后，假设向量 $x_1, x_2, \cdots, x_{d+2}$ 被非齐次半空间打散，但是，这相当于 $\mathbb{R}^{d+1}$ 空间中能被齐次向量打散的向量有 d+2 个，这和前面的定理矛盾。（矛盾的产生相当于分类后的两个子集的 convex hull（凸包）相交）

Radon Theorem 对于任意 $S \subseteq \mathbb{R}^d$ 且 $|S| \geq d + 2$，S 可以被分为两个不相交的子集 A 和 B，使得 $\text{conv}(A) \cap \text{conv}(B) \neq \emptyset$，这里 $\text{conv}(X)$ 表示点集 X 的凸包。

证明： 不妨设 $|S| = d+2$，对其中的每一个点，考虑由其作为列向量构成的 $d \times (d+2)$ 维矩阵 A，添加一个由全 1 构成的行向量到 A，记修改后的矩阵为 B。显然这个矩阵的

秩至多为 $d+1$ 并且它的每一列几乎相互独立。设非零向量 $x=(x_1,x_2,\cdots,x_{d+2})$ 使得 $Bx=0$, 对 x 重排使得 $x_1,x_2,\cdots,x_S\geq 0$ 并且 $x_{S+1},x_{S+2},\cdots,x_{d+2}<0$。归一化 x 使得 $\sum_{i=1}^{S}|x_i|=1$。设 $b_i(a_1)$ 分别是 $B(A)$ 的第 i 列，那么 $\sum_{i=1}^{s}|x_i|b_i=\sum_{i=s+1}^{d+2}|x_i|b_i$, 从而 $\sum_{i=1}^{s}|x_i|a_i=\sum_{i=s+1}^{d+2}|x_i|a_i$ 且 $\sum_{i=1}^{s}|x_i|=\sum_{i=s+1}^{d+2}|x_i|$。由于 $\sum_{i=1}^{s}|x_i|=1$ 且 $\sum_{i=s+1}^{d+2}|x_i|=1$, 每一边 $\sum_{i=1}^{s}|x_i|a_i=\sum_{i=s+1}^{d+2}|x_i|a_i$ 都是 A 的列的凸组合，这证明了定理。因此，S 可以被分为两个集合，第一个集合由重新排列后的前 S 个点构成，第二个集合由点 $s+1$ 到 $d+2$ 组成。他们的凸包相交如所需，证毕。

d 维球体

d 维球体是一组形式为 $\{x|\parallel x-x_0\parallel\leq r\}$ 的点。球体的 VC 维为 d+1，与半空间相同。首先我们指出，由 d 个单位坐标向量和原点组成的 d+1 个点的集合可以被球体打散。假设 A 是这 d+1 个点的任意子集，设 a 是 A 中的单位向量的数量，我们的球的中心 $a+0$ 将是 A 中的向量的和，对于每个 A 中的单位向量，他们到中心的距离都是 $\sqrt{a-1}$，而对于 A 以外的每个单位向量，他们到此中心的距离都是 $\sqrt{a+1}$，原点到中心的距离是 $\sqrt{a}$，因此适当选择半径即可使得 A 中的点恰好都处于超球内。接下来需要说明任何大小为 $d+2$ 的点集都不能被保证打散。否则，假设可以被打散的这样的点集是 S，对于 S 的任意划分 A_1,A_2, 都存在球体 B_1,B_2 使得：

$$B_1\cap S=A_1,\qquad B_2\cap S=A_2$$

虽然后 B_1,B_2 可能相交，但是他们的交集中没有 S 点，很容易看出，存在一个垂直于连接两个球心的直线的超平面，它将所有的 A_1 点放在一侧，而将所有的 A_2 点放在另一侧，这说明半空间可以打散 S，这是一个矛盾。因此，没有 d+2 个点的点集可以被球体打散。

PAC 学习的基本定理

统计学习的基本定理

令 $\mathcal{H}$ 是一个由 $\mathcal{X}$ 到 $\{0,1\}$ 的映射函数构成的假设类，且令损失函数为 0-1 损失，那么，下面的叙述等价：

- 1. $\mathcal{H}$ 有一致收敛性；
- 2. 任何 ERM 规则都是对于 $\mathcal{H}$ 成功的不可知 PAC 学习器；
- 3. $\mathcal{H}$ 是不可知 PAC 可学习的；
- 4. $\mathcal{H}$ 是 PAC 可学习的；
- 5. 任何 ERM 规则都是对于 $\mathcal{H}$ 成功的 PAC 学习器；
- 6. $\mathcal{H}$ 的 VC 维有限。

VC 维不仅能够用于描述 PAC 可学习性，还可以决定样本复杂度。

统计学习定理的定量形式

令 $\mathcal{H}$ 是一个由 $\mathcal{X}$ 到 $\{0,1\}$ 的映射函数构成的假设类，且令损失函数为 0-1 损失。假定 $VCdim(\mathcal{H}) = d < \infty$，那么，存在绝对常数 C_1, C_2 使得：

- $\mathcal{H}$ 有一致收敛性，若其样本复杂度满足：

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}.$$

- $\mathcal{H}$ 是不可知 PAC 可学习的，若其样本复杂度满足：

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}.$$

- $\mathcal{H}$ 是 PAC 可学习的，若其样本复杂度满足：

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}.$$

我们所述的基本定理是针对二分类问题的。对于其他学习问题，如采用绝对值损失或者平方损失的回归问题

对于上述定理，$1 \to 2$ 是很显然的，同时 $2 \to 3, 3 \to 4, 2 \to 5$ 都是显然的。结合没有免费的午餐定理，$4 \to 6, 5 \to 6$ 也是易见的，定理的难点在于 $6 \to 1$，证明过程主要基于：

- 如果 $VCdin(\mathcal{H}) = d$，即使 $\mathcal{H}$ 是无限的，当他限制在一个有限集合 $C \subset \mathcal{X}$ 时，其有效规模 $|\mathcal{H}_C|$ 只有 $O(|C|^d)$，说明这是随 $|C|$ 呈多项式增长的，这就是后面的 Sauer 引理；
- 假设类有一个小的有效规模的时候（$|\mathcal{H}_C|$ 随着 $|C|$ 按照多项式的方式增长。）

生长函数/打散函数

令 $\mathcal{H}$ 是假设类，$\mathcal{H}$ 的生长函数，记作 $\tau_{\mathcal{H}}(m) : \mathbb{N} \to \mathbb{N}$，定义为：

$$\tau_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X}, |C| = m} |\mathcal{H}_C|$$

即，$\tau_{\mathcal{H}}(m)$ 就是从大小为 m 的集合 C 到 $\{0,1\}$ 不同函数的个数，其可由限制 $\mathcal{H}$ 在 C 上获得。显然如果 $VCdim(\mathcal{H}) = d$，则对于所有的 $m \leq d$ 都有 $\tau_{\mathcal{H}}(m) = 2^m$。

(Sauer-Shelah-Perles lemma) 令 $\mathcal{H}$ 是一个假设类，且 $VCdim(\mathcal{H}) \leq d < \infty$，则对于所有的 m，$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$。特别地，如果 $m > d + 1$，那么 $\tau_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$

证明： 对于 $\forall C = \{c_1, \cdots, c_m\}$，有：

$$\forall \mathcal{H}, |\mathcal{H}_c| \leq |\{B \subseteq C : \mathcal{H}\text{打散}B\}| \qquad (1)$$

上式对于证明引理是充分的，因为如果 $VCdim(\mathcal{H}) \leq d$ 那么将不存在规模大于 d 且被 $\mathcal{H}$ 打散的集合，因此：

$$|\{B \subseteq C : \mathcal{H}\text{打散}B\}| \leq \sum_{i=0}^{d} \binom{m}{i}$$

利用二项式展开不难证明上式右端至多为 $(en/d)^d$。现在只需证明 (1) 式成立。采用归纳法，对于 $m=1$ 的情况，无论 $\mathcal{H}$ 是何种形式，(1) 式的两边或均等于 1 或均等于 2（规定空集总是可以被打散)。下面假定对于集合规模 $k<m$ 者均成立，对于集合规模为 m 的情况：固定 $\mathcal{H}$ 以及 $C=\{c_1,\cdots,c_m\}$。另外，记 $C'=\{c_2,\cdots,c_m\}$，并定义如下两个集合：

$$Y_0=\{(y_2,\cdots,y_m):(0,y_2,\cdots,y_m)\in\mathcal{H}_C\vee(1,y_2,\cdots,y_m)\in\mathcal{H}_C\}$$

$$Y_1=\{(y_2,\cdots,y_m):(0,y_2,\cdots,y_m)\in\mathcal{H}_C\wedge(1,y_2,\cdots,y_m)\in\mathcal{H}_C\}$$

易见 $|\mathcal{H}_C|=|Y_0|+|Y_1|$。另外，由于 $Y_0=\mathcal{H}_C$，考虑 $\mathcal{H}$ 在 C' 上的归纳假设，有：

$$|Y_0|=|\mathcal{H}_C|\le|\{B\subseteq C':\mathcal{H}\text{打散}B\}|=|\{B\subseteq C:c_1\notin B\wedge\mathcal{H}\text{打散}B\}|$$

接下来，定义 $\mathcal{H}'\subset\mathcal{H}$ 为：

$$\begin{aligned}\mathcal{H}'&=\{h\in\mathcal{H}:\exists h'\in\mathcal{H}\ s.t.\ (1-h'(c_1),h'(c_2),\cdots,h'(c_m))&(1)\\&=\{h(c_1),h(c_2),\cdots,h(c_m)\}&(2)\end{aligned}$$

即，$\mathcal{H}'$ 包含了那些在 C' 上适用但在 c_1 上不适用的假设。在这样的定义下，显然地，如果 $\mathcal{H}'$ 打散了集合 $B\subseteq C'$，那么它将同时打散集合 $B\cup\{c_1\}$，反之亦成立。将 $Y_1=\mathcal{H}'_{C'}$ 与上述事实联立，并考虑 $\mathcal{H}'$ 在 C' 上的归纳假设，可得：

$$\begin{aligned}|Y_1|=|\mathcal{H}'_{C'}|&\le|\{B\subseteq C':\mathcal{H}'\text{打散}B\}|&(3)\\&=|\{B\subseteq C':\mathcal{H}'\text{打散}B\cup\{c_1\}\}|&(4)\\&=|\{B\subseteq C:c_1\in B\wedge\mathcal{H}'\text{打散}B\}|&(5)\\&\le|\{B\subseteq C:c_1\in B\wedge\mathcal{H}\text{打散}B\}|&(6)\end{aligned}$$

综上有：

$$\begin{aligned}|\mathcal{H}_C|&=|Y_0|+|Y_1|&(7)\\&\le|\{B\subseteq C:c_1\notin B\wedge\mathcal{H}\text{打散}B\}|+|\{B\subseteq C:c_1\in B\wedge\mathcal{H}\text{打散}B\}|&(8)\\&=|\{B\subseteq C:\mathcal{H}\text{打散}B\}|&(9)\end{aligned}$$

theorem-小规模的类的一致收敛性 令 $\mathcal{H}$ 是一个类，令 $\tau_{\mathcal{H}}$ 是其生长函数，那么，对于每个 $\mathcal{D}$ 以及每个 $\delta\in(0,1)$，对于任意 $S\sim\mathcal{D}^m$，都以至少 $1-\delta$ 的概率有下式成立：

$$|L_{\mathcal{D}}(h)-L_s(h)|\le\frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}$$

证明：

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h)-L_S(h)|\right]\le\frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}\tag{3}$$

由于随机变量 $\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h)-L_S(h)|$ 是非负的，因此该定理可直接由马尔可夫不等式推出。

为了给出式 (3) 左半部分的界，我们首先注意到对于每个 $h \in \mathcal{H}$，我们可以重写 $L_{\mathcal{D}}(h) = \mathbb{E}_{S' \sim \mathcal{D}^m}[L_S(h)]$，其中 $S' = z'_1, \cdots, z'_m$ 为新增的独立同分布样本。因此

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right] = \mathbb{E}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} |\mathbb{E}_{S' \sim \mathcal{D}^m} L_{S'}(h) - L_S(h)|\right]$$

利用三角不等式的一般形式可得

$$|\mathbb{E}_{S' \sim \mathcal{D}^m}[L_{S'}(h) - L_S(h)]| \leq \mathbb{E}_{S' \sim \mathcal{D}^m}|L_{S'}(h) - L_S(h)|$$

考虑到期望的上界小于上界的期望，故而：

$$\sup_{h \in \mathcal{H}} \mathbb{E}_{S \sim \mathcal{D}^m}|L_{S'}(h) - L_S(h)| \leq \mathbb{E}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} |L_{S'}(h) - L_S(h)|\right]$$

之前的两个不等式也可由 Jensen 不等式得到，联立上述各式，有

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|\right] \leq \mathbb{E}_{S, S' \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} |L_{S'}(h) - L_S(h)|\right] \tag{10}$$

$$= \mathbb{E}_{S, S' \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}(\ell(h, z'_i) - \ell(h, z_i))\right|\right] \tag{4}$$

等式右边的期望与两个独立同分布的样本 $S = z_1, \cdots, z_m$ 和 $S' = z'_1, \cdots, z'_m$ 有关。由于所有 $2m$ 个向量都是独立同分布的，因此我们将随机变量 z_i 换名为 z'_i 不会产生任何变化，这样之后，式 (6.5) 中的项 $(\ell(h, z'_i) - \ell(h, z_i))$ 将变为项 $-(\ell(h, z'_i) - \ell(h, z_i))$。因此，对于每个 $\boldsymbol{\sigma} \in \{\pm 1\}^m$，式 (6.5) 等价于：

$$\mathbb{E}_{S, S' \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(\ell(h, z'_i) - \ell(h, z_i))\right|\right]$$

由于该式对于每个 $\boldsymbol{\sigma} \in \{\pm 1\}^m$ 成立，如果我们随机地对 $\boldsymbol{\sigma}$ 的每个分量按照在 $\{\pm 1\}$ 上的均匀分布来采样，记作 $U_{\pm}$，该式也是成立的。因此，式 (4) 也等价于

$$\mathbb{E}_{\boldsymbol{\sigma} \sim U_{\pm}^m}\mathbb{E}_{S, S' \sim \mathcal{D}^m}\left[\sup_{h \in \mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(\ell(h, z'_i) - \ell(h, z_i))\right|\right]$$

由于期望是线性的，该式亦等价于

$$\mathbb{E}_{S, S' \sim \mathcal{D}^m}\mathbb{E}_{\boldsymbol{\sigma} \sim U_{\pm}^m}\left[\sup_{h \in \mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(\ell(h, z'_i) - \ell(h, z_i))\right|\right]$$

接下来，固定 S 与 S'，令 C 为在 S 与 S' 中同时出现的实例集。那么，我们可以取只在 $h \in \mathcal{H}_C$ 的上确界，因此：

$$\mathbb{E}_{\boldsymbol{\sigma} \sim U_{\pm}^m}\left[\sup_{h \in \mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(\ell(h, z'_i) - \ell(h, z_i))\right|\right]$$

$$= \mathbb{E}_{\boldsymbol{\sigma} \sim U_{\pm}^m}\left[\max_{h \in \mathcal{H}_C} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i(\ell(h, z'_i) - \ell(h, z_i))\right|\right]$$

固定某个 $h \in \mathcal{H}_C$，并记 $\theta_h = \frac{1}{m}\sum_{i=1}^{m}\sigma_i(\ell(h,z_i') - \ell(h,z_i))$。由于 $\mathbb{E}[\theta_h] = 0$ 且 θ_h 是在 $[-1,1]$ 取值的独立变量的平均值，因此由 Hoeffding 不等式，对于每个 $\rho > 0$

$$\mathbb{P}\left[|\theta_h| > \rho\right] \leq 2\exp\left(-2m\rho^2\right)$$

利用在 $h \in \mathcal{H}_C$ 上的联合界，可以得到，对于任意 $\rho > 0$

$$\mathbb{P}\left[\max_{h\in\mathcal{H}_C}|\theta_h| > \rho\right] \leq 2|\mathcal{H}_C|\exp\left(-2m\rho^2\right)$$

最后，由书中引理可知，上式表明

$$\mathbb{E}\left[\max_{h\in\mathcal{H}_C}|\theta_h|\right] \leq 4 + \sqrt{\frac{\log(|\mathcal{H}_C|)}{2m}}$$

联立上述各式与 $\tau_{\mathcal{H}}$ 的定义，可得

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|\right] \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}$$

回到统计学习的基本定理，欲证有限 VC 维的假设类有一致收敛性，须证：

$$m_{\mathcal{H}}^{UC}(\varepsilon,\delta) \leq 4\frac{16d}{(\delta\varepsilon)^2}log\left(\frac{16d}{(\delta\varepsilon)^2}\right) + \frac{16dlog(2e/d)}{(\delta\varepsilon)^2} \qquad (2)$$

利用 Sauer 定理，对于 $m > d$，有 $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$。该式与小规模的类的一致收敛性定理联立可得下式以至少 $1-\delta$ 的概率成立：

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{4 + \sqrt{dlog(2em/d)}}{\delta\sqrt{2m}}$$

不妨 $\sqrt{dlog(2em/d)} \geq 4$，从而：

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \frac{1}{\delta}\sqrt{\frac{2dlog(2em/d)}{m}}$$

为保证上式至多为 ε，需要：

$$m \geq \frac{2dlog(m)}{(\delta\varepsilon)^2} + \frac{2dlog(2e/d)}{(\delta\varepsilon)^2}$$

这等价于 (2) 式。

集合系统

称一个集合 X 以及被限制在 X上的假设类$\mathcal{H}$ 的一个 pair 为一个集合系统，记作 $(X,\mathcal{H})$。

设 $(X,\mathcal{H}_1)$ 以及 $(X,\mathcal{H}_2)$ 为同一个同一个基础集合 X 上的两个集合系统。定义另一个集合系统，称为交集系统 $(X,\mathcal{H}_1\cap\mathcal{H}_2)$，其中 $\mathcal{H}_1\cap\mathcal{H}_2 = \{h_1\cap h_2 | h_1\in\mathcal{H}_1, h_2\in\mathcal{H}_2\}$。换句话说，取 $\mathcal{H}_1$ 中的每个集合与 $\mathcal{H}_2$ 中的每个集合的交集。例如，如果 $U = \mathbb{R}^d$ 且 $\mathcal{H}_1$ 和 $\mathcal{H}_2$ 都是所有半空间的集合，那么 $\mathcal{H}_1\cap\mathcal{H}_2$ 由两个半空间的交集定义的所有集合组成。这相当于取两个阈值门输出的布尔 AND，是除了单个门之外最基本的神经网络之一。如果重复该过程并取 k 个半空间的交，下面的引理限制了过程中的生长函数的界：

假设 $(X, \mathcal{H}_1)$ 以及 $(X, \mathcal{H}_2)$ 为同一个基础集合 X 上的两个集合系统，则有：

$$m_{\mathcal{H}_1 \cap \mathcal{H}_2}(n) \leq m_{\mathcal{H}_1}(n) \cdot m_{\mathcal{H}_2}(n)$$

假设 S 是把假设类的交限制在 A 上的结果，注意到这相当于把假设类分别限制在 A 上的结果的交，利用排列组合规则不难得到上述定理。
上述定理的推广：
对于给定的 k 个 concepts：$h_1, h_2, \cdots, h_k$ 和一个 bool 函数 f，定义：

$$comb_f(h_1, \cdots, h_k) = \{x \in X | f(h_1(x), \cdots, h_k(x)) = 1\}$$

这里使用 $h_i(x)$ 表示 x 是否属于 h_i 的指示器。例如，如果 f 是 AND 函数，那么 $comb_f(h_1, \cdots, h_k)$ 就是 h_i 的交集；如果 f 是多数投票函数，那么 $comb_f(h_1, \cdots, h_k)$ 是由超过一半的集合 h_i 包含的点组成的集合，$comb_f(h_1, \cdots, h_k)$ 也可以被视为深度为两层的神经网络，等等。我们继续定义 $COMB_{f,k} = \{comb_f(h_1, \cdots, h_k) | h_i \in \mathcal{H}\}$，可得以下引理：

$$m_{COMB_{f,k}(\mathcal{H})}(n) \leq m_{\mathcal{H}}(n)^k$$

VC-维数的组合类 如果假设类 $\mathcal{H}$ 的 VC-维数为 V，那么对于任意布尔函数 f 和整数 k，组合类 $\text{COMB}_{f,k}(\mathcal{H})$ 的 VC 维为 $O(kV \log(kV))$。

证明. 设 $\text{COMB}_{f,k}(\mathcal{H})$ 的 VC 维为 n。根据定义，存在一个由 n 个点组成的集合 S，使得 S 被 $\text{COMB}_{f,k}(\mathcal{H})$ 打散。

根据 Sauer 引理，我们知道用 $\mathcal{H}$ 中的集合对 S 中的点进行划分的方式至多有 $\sum_{i=0}^{V} \binom{n}{i} \leq n^V$ 种。由于 $\text{COMB}_{f,k}(\mathcal{H})$ 中的每个集合是由 $\mathcal{H}$ 中的 k 个集合确定的，因此不同的 k-元组至多有 $(n^V)^k = n^{kV}$ 种。这意味着用 $\text{COMB}_{f,k}(\mathcal{H})$ 对 S 中的点进行划分的方式至多有 n^{kV} 种。

因为 S 被打散，所以必须有 $2^n \leq n^{kV}$，即等价于 $n \leq kV \log_2(n)$。我们可以通过以下步骤求解这个不等式：

首先，假设 $n \geq 16$，则有 $\log_2(n) \leq \sqrt{n}$，因此 $kV \log_2(n) \leq kV\sqrt{n}$，这进一步意味着 $n \leq (kV)^2$。

为了得到更紧的界，我们将 $n \leq (kV)^2$ 代入原不等式，得到：$\log_2(n) \leq 2\log_2(kV)$。进一步有：

$$n \leq kV \cdot 2\log_2(kV) = 2kV \log_2(kV).$$

因此，$\text{COMB}_{f,k}(\mathcal{H})$ 的 VC-维数为 $O(kV \log(kV))$（上界是比较紧的，符合大 O 法则）。 □

概率分布下的集合系统 设 $(X, \mathcal{H})$ 是一个集合系统，D 是 X 上的概率分布，且 n 是满足以下条件的整数：

$$n \geq \frac{8}{\varepsilon} \quad \text{和} \quad n \geq \frac{2}{\varepsilon}\left(\log_2(2m_{\mathcal{H}}(2n)) + \log_2\frac{1}{\delta}\right).$$

令 S_1 是从 D 中抽取的 n 个点组成的集合。那么，以至少 $1-\delta$ 的概率，$\mathcal{H}$ 中每个概率质量大于 ϵ 的集合都会与 S_1 相交。

注意，n 同时出现在上述不等式的两边。如果 $\mathcal{H}$ 的 VC-维数有限为 d，这不会导致循环依赖，因为根据 Sauer 引理，有 $\log(\pi_{\mathcal{H}}(2n)) = O(d \log n)$。因此，形式为 $n \geq a \log n$（其中 $a \geq 4$ 是正整数）的不等式可以由 $n \geq ca \ln a$ 推导出来，从而消除了右边的 n。

证明. 令 A 表示存在一个集合 $h \in \mathcal{H}$，其概率质量大于或等于 ϵ 且与 S_1 不相交的事件。从 D 中再抽取一个包含 n 个点的集合 S_2。令 B 表示存在一个集合 $h \in \mathcal{H}$，该集合与 S_1 不相交但包含 S_2 中至少 $\frac{\epsilon}{2}n$ 个点的事件。即：

$$B : \exists h \in \mathcal{H},\ |S_1 \cap h| = 0 \quad \text{但} \quad |S_2 \cap h| \geq \frac{\epsilon}{2}n.$$

根据切比雪夫不等式，如果 $n \geq 8/\epsilon$，则 $\mathbb{P}(B \mid A) \geq \frac{1}{2}$。具体来说，如果 h 与 S_1 不相交且其概率质量大于或等于 ϵ，则 h 至少有 $\frac{1}{2}$ 的概率包含新随机集合 S_2 中至少 $\frac{\epsilon}{2}n$ 个点。这意味着：

$$\mathbb{P}(B) \geq \mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\mathbb{P}(A) \geq \frac{1}{2}\mathbb{P}(A).$$

因此，要证明 $\mathbb{P}(A) \leq \delta$，希望证明 $\mathbb{P}(B) \leq \frac{\delta}{2}$。为此，我们考虑另一种方式来选择 S_1 和 S_2。从 D 中随机抽取一个包含 $2n$ 个点的集合 S_3，然后将 S_3 随机分成两个相等的部分；令 S_1 为第一部分，S_2 为第二部分。显然，这种方法得到的 S_1 和 S_2 的概率分布与独立选择是相同的。

现在，考虑在 S_3 已经被抽取但尚未随机划分成 S_1 和 S_2 时的情况。即使 $\mathcal{H}$ 可能包含无限多个集合，我们知道它与 S_3 的不同交集至多有 $m_{\mathcal{H}}(2n)$ 个。即：

$$|\{S_3 \cap h \mid h \in \mathcal{H}\}| \leq m_{\mathcal{H}}(2n).$$

因此，为了证明 $\mathbb{P}(B) \leq \frac{\delta}{2}$，只需证明对于任意给定的 S_3 的子集 h'，在随机划分 S_3 成 S_1 和 S_2 的情况下，满足 $|S_1 \cap h'| = 0$ 但 $|S_2 \cap h'| \geq \frac{\epsilon}{2}n$ 的概率至多为 $\frac{\delta}{2m_{\mathcal{H}}(2n)}$。

为此，首先注意到如果 h' 包含的点少于 $\frac{\epsilon}{2}n$，则不可能有 $|S_2 \cap h'| \geq \frac{\epsilon}{2}n$。对于 h' 包含超过 $\frac{\epsilon}{2}n$ 个点的情况，随机划分 S_3 使得 h' 中没有点落入 S_1 的概率至多为 $\left(\frac{1}{2}\right)^{\epsilon n/2}$。根据定理中的 n 的界，我们有：

$$2^{-\epsilon n/2} \leq 2^{-\log(2m_{\mathcal{H}}(2n)) + \log \delta} = \frac{\delta}{2m_{\mathcal{H}}(2n)},$$

这正是我们所需要的。因此，$\mathbb{P}(B) \leq \frac{\delta}{2}$，从而 $\mathbb{P}(A) \leq \delta$。

这种通过两种方式选择 S_1 和 S_2 的论证方法称为“双重抽样”或“幽灵样本”方法。关键思想是推迟某些随机选择直到问题被转化为有限规模的问题。双重抽样在其他上下文中也很有用。 □

我们现在将 VC-维的概念应用到机器学习中。在机器学习中，我们有一个目标概念 c^*，例如垃圾邮件，并且有一组假设 $\mathcal{H}$，这些假设是我们声称的垃圾邮件集合。设 $\mathcal{H}' = \{h \triangle c^* \mid h \in \mathcal{H}\}$ 是 $\mathcal{H}$ 中假设的错误区域集合。注意，$\mathcal{H}'$ 和 $\mathcal{H}$ 具有相同的 VC 维和生长函数。我们现在抽取一个训练样本 S（邮件集），并应用定理 5.14 到 $\mathcal{H}'$，以论证对于所有满足 $\mathbb{P}(h \triangle c^*) \geq \epsilon$ 的假设 h，几乎可以肯定地有 $|S \cap (h \triangle c^*)| > 0$。换句话说，几乎可以肯定的是，只有真实误差较低的假设才会完全符合训练样本。这在下面的定理陈述中得到了形式化。

样本界定理 对于任何类 $\mathcal{H}$ 和分布 D，如果从 D 中抽取的训练样本 S 的大小满足：

$$n \geq \frac{2}{\epsilon}\left(\log(2m_H(2n)) + \log\frac{1}{\delta}\right),$$

那么以概率大于或等于 $1-\delta$，对于所有 $h \in \mathcal{H}$，若其真实误差 $L_D(h) \geq \epsilon$，则其训练误差 $L_S(h) > 0$。等价地，对于所有 $h \in \mathcal{H}$，若其训练误差 $\mathrm{err}_S(h) = 0$，则其真实误差 $L_D(h) < \epsilon$。

证明： 该证明是通过将概率分布下的集合系统定理应用于 $\mathcal{H}' = \{h \triangle c^* \mid h \in \mathcal{H}\}$ 得出的。□

类比上面的定理我们也能得出前面的 $6 \to 1$ 的证明。

最后，我们应用 Sauer 引理将上述定理中的生长函数替换为 VC-维。我们对样本界定理进行这样的改写：

> 对于任何类 $\mathcal{H}$ 和分布 D，一个大小为
>
> $$O\left(\frac{1}{\epsilon}\left(\mathrm{VCdim}(\mathcal{H})\log\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right)$$
>
> 的训练样本 S 足以确保以概率大于或等于 $1-\delta$, 对于所有 $h \in \mathcal{H}$, 若其真实误差 $L_D(h) \geq \epsilon$, 则其训练误差 $L_S(h) > 0$。等价地，对于所有 $h \in \mathcal{H}$，若其训练误差 $L_S(h) = 0$，则其真实误差 $L_D(h) < \epsilon$。

VC-维和描述一个集合所需的比特数并不是唯一可以用来推导泛化保证的复杂度度量。已经有很多工作研究了各种各样的度量。其中一个度量称为 Rademacher 复杂度，它衡量给定假设类 $\mathcal{H}$ 对随机噪声的拟合程度。给定一组 n 个样本 $S = \{x_1, \ldots, x_n\}$，H 的经验 Rademacher 复杂度定义为

$$R_S(H) = \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\max_{h\in H}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(x_i)\right],$$

其中 $\sigma_i \in \{-1, 1\}$ 是独立的随机标签，且 $\mathbb{P}[\sigma_i = 1] = \frac{1}{2}$。例如，如果你对 S 中的点赋予随机的 ± 1 标签，而 $\mathcal{H}$ 中最好的分类器平均误差为 0.45，则 $R_S(H) = 0.55 - 0.45 = 0.1$。可以证明，以概率大于或等于 $1-\delta$，对于所有 $h \in \mathcal{H}$，其真实误差小于或等于训练误差加上 $R_S(H) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}$。