复杂度

前言

由PAC学习理论可以看出，PAC可学习性与假设空间 $\mathcal{H}$ 的复杂程度密切相关。假设空间 $\mathcal{H}$ 越复杂，从中寻找到目标概念的难度越大。

对于有限的假设空间，可以用其中包含假设的数目 $|\mathcal{H}|$ 来刻画假设空间的复杂度。在可分情形下，可以通过层层筛选的方式从有限的假设空间中寻找到目标概念。

然而，对于大多数学习问题来说，算法 $\mathfrak{L}$ 考虑的假设空间并非是有限的，直接使用 $|\mathcal{H}|$ 来刻画假设空间的复杂度不再有意义。

为此，本章将介绍刻画无限假设空间复杂度的方法，包括与数据分布 $\mathcal{D}$ 无关的 VC 维及其扩展 Natarajan 维，以及与数据分布相关的 Rademacher 复杂度。

前者计算简单，适用于许多学习问题，但其未考虑具体学习问题的数据特点，对假设空间复杂度的刻画较为粗糙；后者考虑了具体学习问题的数据特点，对假设空间复杂度的刻画较为准确，但计算复杂度较高，有时甚至是 NP-难问题。

数据分布无关

现实学习任务面对的通常是无限假设空间，例如实数域中的区间，$\mathbb{R}^d$ 空间中的所有线性超平面。

为了对这些无限假设空间进行研究，通常考虑其 VC 维（Vapnik-Chervonenkis Dimension）[Vapnik and Chervonenkis, 1971]。

在介绍 VC 维之前，先引入几个概念。

限制（restriction）

令 $\mathcal{H}$ 表示假设空间，其中的假设是 $\mathcal{X}$ 到 $\mathcal{Y}=\{-1,+1\}$ 的映射，对于数据集 $D=\{x_1,\ldots,x_m\}\subset\mathcal{X}$, $\mathcal{H}$ 在数据集 D 上的 限制 是从 D 到 $\{-1,+1\}^m$ 的一族映射：

$$\mathcal{H}_{|D}=\{(h(\boldsymbol{x}_1),\ldots,h(\boldsymbol{x}_m)):h\in\mathcal{H}\}. \qquad (3.1)$$

其中，h 在 D 上的限制是一个m维向量。

- 假设空间 $\mathcal{H}$ 中不同的假设对于 D 中示例赋予标记的结果可能相同，也可能不同。
- 尽管 $\mathcal{H}$ 可能包含无穷多个假设，但 $\mathcal{H}_{|D}$ 却是有限的，即 $\mathcal{H}$ 对 D 中示例赋予标记的可能结果数是有限的。
- 例如，对二分类问题，对于 m个示例最多有 2^m个可能的结果。

增长函数 (growth function)

增长函数 (growth function)

对于 $m \in \mathbb{N}$，假设空间 $\mathcal{H}$ 的增长函数 $\Pi_{\mathcal{H}}(m)$ 表示为

$$\Pi_{\mathcal{H}}(m) = \max_{\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\}\subset\mathcal{X}} \left|\{h(\boldsymbol{x}_1),\ldots,h(\boldsymbol{x}_m)|h\in\mathcal{H}\}\right|. \tag{3.2}$$

特别地，对二分类问题，有

$$\Pi_{\mathcal{H}}(m) = \max_{|D|=m} |\mathcal{H}_{|D}|. \tag{3.3}$$

增长函数表示假设空间对个示例所能赋予标记的最大可能的结果数。

增长函数在一定程度上描述了假设空间的表达能力，反映了假设空间的复杂程度！

对分（dichotomy）、打散（shattering）

对分（dichotomy）

- 对于二分类问题，假设空间 $\mathcal{H}$ 中的假设对 D 中的示例赋予标记的每种可能结果称为对 D 的一种对分。
- 该假设把 D 中的示例分成了正、负两类。

打散（shattering）

- 如果假设空间 $\mathcal{H}$ 能实现示例集 D 上的所有对分，即 $|\mathcal{H}_{|D}| = 2^m$，则称示例集 D 能被假设空间 $\mathcal{H}$ 打散，此时 $\Pi_{\mathcal{H}}(m) = 2^m$。

对分（dichotomy）、打散（shattering）

例如：

令 $\mathcal{H}$ 表示 $\mathbb{R}$ 上的阈值函数构成的集合，其中阈值函数表示为 $h_a(x) = \text{sign}(\mathbb{I}_{\{x<a\}} - \frac{1}{2})$，则 $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$。

令 $D = \{x_1\}$，如果取 $a = x_1 + 1$，则 $h_a(x_1) = +1$；
如果取 $a = x_1 - 1$，则 $h_a(x_1) = -1$。
因此 $\mathcal{H}$ 能打散 $D = \{x_1\}$。

令 $D' = \{x_1, x_2\}$，不妨假设 $x_1 < x_2$，则易知同时将 x_1 分类为 -1 但把 x_2 分类为 $+1$ 的结果不能被 $\mathcal{H}$ 中的任何阈值函数实现。
这是因为如果 $h_a(x_1) = -1$，则必有 $h_a(x_2) = -1$。
所以 $\mathcal{H}$ 不能打散 D'。

VC维 [Vapnik and Chervonenkis, 1971]

定义 3.1 [Vapnik and Chervonenkis, 1971] VC 维: 假设空间 $\mathcal{H}$ 的 VC 维是能被 $\mathcal{H}$ 打散的最大示例集的大小, 即

$$VC(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}. \tag{3.4}$$

$VC(\mathcal{H}) = d$ 表明存在大小为 d 的示例集能被假设空间 $\mathcal{H}$ 打散。

注意:

- 这并不意味着所有大小为 d 的示例集都能被假设空间 $\mathcal{H}$ 打散。
- VC 维的定义与数据分布 $\mathcal{D}$ 无关！因此，在数据分布未知时仍能计算出假设空间 $\mathcal{H}$ 的 VC 维。

VC维 [Vapnik and Chervonenkis, 1971]

定义 3.1 [Vapnik and Chervonenkis, 1971] VC 维: 假设空间 $\mathcal{H}$ 的 VC 维是能被 $\mathcal{H}$ 打散的最大示例集的大小, 即

$$VC(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}. \tag{3.4}$$

要证明一个具体的假设空间 $\mathcal{H}$ 的 VC 维为 d，需要证明两点:

- 存在大小为 d 的示例集 D 能被 $\mathcal{H}$ 打散;
- 任意大小为 $d+1$ 的示例集 D' 都不能被 $\mathcal{H}$ 打散。

用VC 维来衡量有限假设空间的复杂度更为准确，且更具优势。

- 令假设空间 $\mathcal{H}$ 为有限集合。对于任意数据集 D，有 $|\mathcal{H}_{|D}| \leq |\mathcal{H}|$ 。还可知当 $|\mathcal{H}| < 2^{|D|}$ 时，$\mathcal{H}$ 无法打散 D。因此，可得 $VC(\mathcal{H}) \leq \log_2 |\mathcal{H}|$ 。事实上，有限假设空间 $\mathcal{H}$ 的 $VC(\mathcal{H})$ 通常可以远小于 $\log_2 |\mathcal{H}|$。

两种假设空间的VC维

两种假设空间的 VC 维：

- 阈值函数的VC维：

令 $\mathcal{H}$ 表示所有定义在 $\mathbb{R}$ 上的阈值函数组成的集合，由上述讨论可知存在大小为 1 的示例集 D 能被 $\mathcal{H}$ 打散，但任意大小为 2 的示例集 $|D'|$ 都不能被 $\mathcal{H}$ 打散，于是根据定义可知 $VC(\mathcal{H}) = 1$ 。

- 区间函数的VC维：

令 $\mathcal{H}$ 表示所有定义在 $\mathbb{R}$ 上的区间组成的集合 $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ ，其中 $h_{a,b}(x) = (\mathbb{I}_{\{x \in (a,b)\}} - 1/2)$ 。令 $D = \{1, 2\}$ ，易知 $\mathcal{H}$ 能打散 D，因此 $VC(\mathcal{H}) \geq 2$ 。对于任意大小为 3 的示例集 $D' = \{x_1, x_2, x_3\}$，不妨设 $x_1 < x_2 < x_3$，则分类结果 $(+1, -1, +1)$ 不能被 $\mathcal{H}$ 中的任何区间函数实现，因为当 $h_{a,b}(x_1) = +1$ 且 $h_{a,b}(x_3) = +1$ 时，必有 $h_{a,b}(x_2) = +1$ 。所以 $\mathcal{H}$ 无法打散任何大小为 3 的示例集，即得出结论 $VC(\mathcal{H}) = 2$ 。

VC维与增长函数 [Sauer, 1972]

由增长函数的定义可知，VC维与增长函数关系密切，引理 3.1 [Sauer, 1972] 给出了二者之间的定量关系：

引理 3.1 若假设空间 $\mathcal{H}$ 的 VC 维为 d, 则对任意 $m \in \mathbb{N}$ 有

$$\Pi_{\mathcal{H}}(m) \leqslant \sum_{i=0}^{d} \binom{m}{i}. \tag{3.5}$$

注：Sauer 引理由 [Sauer, 1972] 而命名，但 Vapnik and Chervonenkis [1971] 和 Shelah [1972] 也分别独立地给出了该结果。

VC维与增长函数 [Sauer, 1972]

证明：利用数学归纳法证明。

当 $m=1,\ d=0$ 或 $d=1$ 时，引理成立。

假设引理对 $(m-1,d-1)$ 和 $(m-1,d)$ 成立。

令 $D=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\}$，$D'=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{m-1}\}$，

$$\mathcal{H}_{|D}=\left\{\left(h(\boldsymbol{x}_1),\ldots,h(\boldsymbol{x}_m)\right)\middle|h\in\mathcal{H}\right\},\tag{3.6}$$

$$\mathcal{H}_{|D'}=\left\{\left(h(\boldsymbol{x}_1),\ldots,h(\boldsymbol{x}_{m-1})\right)\middle|h\in\mathcal{H}\right\},\tag{3.7}$$

分别为假设空间在 D 和 D' 上的限制。任何假设 $h\in\mathcal{H}$ 对 $\boldsymbol{x}_m$ 的分类结果为 $+1$ 或 -1，因此任何出现在 $\mathcal{H}_{|D'}$ 的串都会在 $\mathcal{H}_{|D}$ 出现一次或者两次。令 $\mathcal{H}_{D'|D}$ 表示 $\mathcal{H}_{|D}$ 中出现两次的 $\mathcal{H}_{|D'}$ 中串组成的集合，即

$$\mathcal{H}_{D'|D}=\Big\{(y_1,\ldots,y_{m-1})\in\mathcal{H}_{|D'}\Big|\exists\, h,h'\in\mathcal{H},$$
$$(h(\boldsymbol{x}_i)=h'(\boldsymbol{x}_i)=y_i)\wedge(h(\boldsymbol{x}_m)\neq h'(\boldsymbol{x}_m)),1\leqslant i\leqslant m-1\Big\}.\tag{3.8}$$

VC维与增长函数 [Sauer, 1972]

考虑到 $\mathcal{H}_{D'|D}$ 中的串在 $\mathcal{H}_{|D}$ 中出现了两次，但在 $\mathcal{H}_{|D'}$ 中仅出现了一次，有

$$\left|\mathcal{H}_{|D}\right| = \left|\mathcal{H}_{|D'}\right| + \left|\mathcal{H}_{D'|D}\right|. \tag{3.9}$$

D' 的大小为 $m-1$，根据归纳的前提假设可得

$$\left|\mathcal{H}_{|D'}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d} \binom{m-1}{i}. \tag{3.10}$$

令 Q 表示能被 $\mathcal{H}_{D'|D}$ 打散的集合，由 Q 的定义可知 $Q \cup \{x_m\}$ 必能被 $\mathcal{H}_{|D}$ 打散。由于 $\mathcal{H}$ 的 VC 维为 d，因此 $\mathcal{H}_{D'|D}$ 的 VC 维最大为 $d-1$，于是有

$$\left|\mathcal{H}_{D'|D}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d-1} \binom{m-1}{i}. \tag{3.11}$$

VC维与增长函数 [Sauer, 1972]

综合式 (3.9) - (3.11) 可得

$$\begin{aligned}\left|\mathcal{H}_{|D}\right| &\leqslant \sum_{i=0}^{d}\binom{m-1}{i}+\sum_{i=0}^{d-1}\binom{m-1}{i} \\ &= \sum_{i=0}^{d}\left(\binom{m-1}{i}+\binom{m-1}{i-1}\right) \\ &= \sum_{i=0}^{d}\binom{m}{i}. \end{aligned} \tag{3.12}$$

由 D 的任意性，式 (3.5) 成立。引理得证。

VC维与增长函数 [Sauer, 1972]

由引理 3.1 可以计算出增长函数的上界 [Sauer, 1972] :

定理 3.1 若假设空间 $\mathcal{H}$ 的 VC 维为 d, 则对任意整数 $m \geqslant d$ 有

$$\Pi_{\mathcal{H}}(m) \leqslant \left(\frac{e \cdot m}{d}\right)^{d} . \tag{3.13}$$

- 当假设空间 $\mathcal{H}$ 的 VC 维为无穷大时，任意大小的示例集 D 都能被 $\mathcal{H}$ 打散，此时有 $\Pi_{\mathcal{H}(m)} = 2^m$，增长函数随着数据集的大小指数级增长；
- 当 VC 维有限为 d 且 $m \geq d$ 时，由定理 3.1 可知增长函数随数据集的大小多项式级增长。

VC维与增长函数 [Sauer, 1972]

证明：

$$
\begin{aligned}
\Pi_{\mathcal{H}}(m) &\leqslant \sum_{i=0}^{d} \binom{m}{i} \\
&\leqslant \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \\
&= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^{i} \\
&\leqslant \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^{i} \\
&= \left(\frac{m}{d}\right)^{d} \left(1+\frac{d}{m}\right)^{m} \\
&\leqslant \left(\frac{e \cdot m}{d}\right)^{d} .
\end{aligned}
$$

多分类问题与Natarajan维

- VC维是针对二分类问题定义的。
- 对于多分类问题，需定义Natarajan维。

多分类问题：

假设空间$\mathcal{H}$包含的假设是 $\mathcal{X}$ 到 $\mathcal{Y}=\{0,\ldots,K-1\}$的映射，其中$K$为类别数。

打散(多分类问题)：

对于给定的一个集合 $D\subset\mathcal{X}$，若假设空间 $\mathcal{H}$ 中存在两个映射 $f_0,f_1:D\to\mathcal{Y}$ 满足以下两个条件：

- 对于任意的 $\boldsymbol{x}\in D$, $f_0(\boldsymbol{x})\neq f_1(\boldsymbol{x})$;
- 对于任意的集合 $B\subset D$ 存在一个映射 $h\in\mathcal{H}$ 使得

$$\forall\boldsymbol{x}\in B,h(\boldsymbol{x})=f_0(\boldsymbol{x})\text{ 且 }\forall\boldsymbol{x}\in D\backslash B,h(\boldsymbol{x})=f_1(\boldsymbol{x})\ , \tag{3.14}$$

则称集合 D 能被假设空间 $\mathcal{H}$ **打散**(多分类问题)。

Natarajan维 [Natarajan, 1989]

定义 3.2 Natarajan 维: 对于多分类问题的假设空间 $\mathcal{H}$, Natarajan 维是能被 $\mathcal{H}$ 打散的最大示例集的大小, 记作 $Natarajan(\mathcal{H})$.

显然，Natarajan维是VC维从二分类问题到多分类问题的扩展。

下面的定理表明，二分类问题的VC维与Natarajan维相同。

定理 3.2 类别数 $K = 2$ 时, $VC(\mathcal{H}) = Natarajan(\mathcal{H})$.

Natarajan维 [Natarajan, 1989]

证明： 首先证明 $VC(\mathcal{H}) \leq Natarajan(\mathcal{H})$。

令 D 表示大小为 $VC(\mathcal{H})$ 且能被 $\mathcal{H}$ 打散的集合。取多分类问题打散定义中的 f_0, f_1 为常值函数，即 $f_0 = 0, f_1 = 1$。由于 D 能被 $\mathcal{H}$ 打散，则对于任意集合 $B \subset D$，存在 h_B 使得 $\boldsymbol{x} \in B$ 时 $h_B(\boldsymbol{x}) = 0$, $\boldsymbol{x} \in D \backslash B$ 时 $h_B(\boldsymbol{x}) = 1$，即 $\mathcal{H}$ 能打散大小为 $VC(\mathcal{H})$ 的 D，于是有 $VC(\mathcal{H}) \leq Natarajan(\mathcal{H})$。

再证明 $VC(\mathcal{H}) \geq Natarajan(\mathcal{H})$。

令 D 表示大小为 $Natarajan(\mathcal{H})$ 且在多分类问题中能被 $\mathcal{H}$ 打散的集合。当 $K = 2$ 时，$f_0, f_1 : D \to \mathcal{Y} = \{0, 1\}$。取 $D_i^v = \{x \in D | f_i(x) = v\}$。对于 D 上的任意一种划分 $c : D \to \mathcal{Y}$，记 $D^+ = \{\boldsymbol{x} \in D | c(\boldsymbol{x}) = 1\}$, $D^- = \{\boldsymbol{x} \in D | c(\boldsymbol{x}) = 0\}$。令多分类问题打散定义中 $B = (D^+ \cap D_0^1) \cup (D^- \cap D_0^0)$，可知存在 $h \in \mathcal{H}$ 使得 $\forall \boldsymbol{x} \in D$ 有 $h(\boldsymbol{x}) = c(\boldsymbol{x})$，即 $\mathcal{H}$ 能打散大小为 $Natarajan(\mathcal{H})$ 的 D，于是有 $VC(\mathcal{H}) \geq Natarajan(\mathcal{H})$。

定理得证。

Natarajan维与增长函数 [Natarajan, 1989]

对于多分类问题，通过Natarajan维控制增长函数的增长速度，可得到下面的定理 [Natarajan, 1989]：

定理 3.3 若多分类问题假设空间 $\mathcal{H}$ 的 Natarajan 维为 d, 类别数为 K, 则对任意 $m \in \mathbb{N}$ 有

$$\Pi_{\mathcal{H}}(m) \leqslant m^d K^{2d}. \tag{3.15}$$

Natarajan维与增长函数 [Natarajan, 1989]

证明：利用数学归纳法证明。

当 $m=1,\ d=0$ 或 $d=1$ 时，定理成立。

假设定理对 $(m-1,d-1)$ 和 $(m-1,d)$ 成立。

令 $D=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\}$，对于 $\mathcal{Y}=\{0,\ldots,K-1\}$，$\mathcal{H}_{|D}$ 可以推广到多分类问题。

令

$$\mathcal{H}_k=\left\{h\in\mathcal{H}_{|D}\middle|h(\boldsymbol{x}_1)=k\right\},k=0,\ldots,K-1, \tag{3.16}$$

基于 $\mathcal{H}_k$ 可以定义如下集合：

$$\forall i\neq j,\mathcal{H}_{ij}=\left\{h\in\mathcal{H}_i\middle|\exists h'\in\mathcal{H}_j,h(\boldsymbol{x}_l)=h'(\boldsymbol{x}_l),l=2,\ldots,m\right\}, \tag{3.17}$$

$$\bar{\mathcal{H}}=\mathcal{H}_{|D}-\cup_{i\neq j}\mathcal{H}_{ij}. \tag{3.18}$$

根据联合界 (Union Bound) 不等式 (1.14) 可知

$$\left|\mathcal{H}_{|D}\right|\leqslant\left|\bar{\mathcal{H}}\right|+\left|\cup_{i\neq j}\mathcal{H}_{ij}\right|\leqslant\left|\bar{\mathcal{H}}\right|+\sum_{i\neq j}\left|\mathcal{H}_{ij}\right|. \tag{3.19}$$

Natarajan维与增长函数 [Natarajan, 1989]

根据定义，$\bar{\mathcal{H}}$ 在 $D-\{\boldsymbol{x}_1\}$ 上无相同假设，且 $Natarajan(\bar{\mathcal{H}}) \leq d$ 。根据归纳的前提假设，可知

$$|\bar{\mathcal{H}}| \leqslant \Pi_{\bar{\mathcal{H}}}(m-1) \leqslant (m-1)^d K^{2d}. \tag{3.20}$$

同时，对于任意 $\mathcal{H}_{ij}$，其 Natarajan 维最多为 $d-1$，否则 $\mathcal{H}$ 的 Natarajan 维将超过 d。同样根据归纳的前提假设，有

$$\forall i \neq j, |\mathcal{H}_{ij}| \leqslant \Pi_{\mathcal{H}_{ij}}(m) \leqslant m^{d-1} K^{2(d-1)}. \tag{3.21}$$

综上可得

$$\begin{aligned} \left|\mathcal{H}_{|D}\right| \leqslant \left|\bar{\mathcal{H}}\right| + \sum_{i \neq j} |\mathcal{H}_{ij}| &\leqslant \Pi_{\bar{\mathcal{H}}}(m-1) + \sum_{i \neq j} \Pi_{\mathcal{H}_{ij}}(m) \\ &\leqslant (m-1)^d K^{2d} + K^2 m^{d-1} K^{2(d-1)} \\ &\leqslant m^d K^{2d}. \end{aligned} \tag{3.22}$$

由 D 的任意性，(3.15) 成立。定理得证。

数据分布相关

VC维

VC维的定义与数据分布无关。因此基于VC维的分析结果是分布无关、数据独立的，也就是说对任意数据分布都成立。

- 一方面，基于VC维的分析结果具有一定 的“普适性”；
- 另一方面，由于没有考虑数据自身，基于VC维的分析结果通常比较“松”，对那些与学习问题的典型情况相差甚远的较“坏”分布来说尤其如此。

Rademacher 复杂度

- 是另一种刻画假设空间复杂度的工具；
- 与VC维不同的是，它在一定程度上考虑了数据分布。

最小经验误差

给定数据集 $D = \{(x_1, y_1), \ldots, (x_m, y_m)\}$，假设 h 的经验误差为

$$\begin{aligned}\hat{E}(h) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(h(\boldsymbol{x}_i) \neq y_i) \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{1 - y_i h(\boldsymbol{x}_i)}{2} \\ &= \frac{1}{2} - \frac{1}{2m}\sum_{i=1}^{m} y_i h(\boldsymbol{x}_i). \end{aligned} \tag{3.23}$$

其中 $\frac{1}{m}\sum_{i=1}^{m} y_i h(\boldsymbol{x}_i)$ 体现了预测值 $h(\boldsymbol{x}_i)$与样例真实标记 y_i 之间的一致性，若 $h(\boldsymbol{x}_i) = y_i,\ 1 \leq i \leq m$，则 $\frac{1}{m}\sum_{i=1}^{m} y_i h(\boldsymbol{x}_i)$ 取得最大值1，也就是说具有最小经验误差的假设是

$$\arg\max_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} y_i h(\boldsymbol{x}_i). \tag{3.24}$$

Rademacher随机变量

现实任务中样例的标记有时会受到噪声的影响，即对某些样例 $(\boldsymbol{x}_i, y_i)$来说，其 y_i 或许已经受到随机因素的影响，不再是 $\boldsymbol{x}_i$ 的真实标记。在此情形下，选择假设空间 $\mathcal{H}$ 中在训练集上表现最好的假设，有时还不如选择 $\mathcal{H}$ 中事先已考虑了随机噪声的假设。

Rademacher随机变量

考虑随机变量 σ_i，它以 0.5 的概率取值 -1，以 0.5 的概率取值 $+1$，称其为 Rademacher 随机变量。基于 σ_i 可将 (3.24) 改写为

$$\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(\boldsymbol{x}_i). \tag{3.25}$$

对所有的 σ_i 求期望可得

$$\mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(\boldsymbol{x}_i) \right] . \tag{3.26}$$

Rademacher随机变量

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)\right] . \tag{3.26}$$

式 (3.26) 和增长函数有着相似的作用，体现了假设空间在数据集 D上的表示能力，取值范围为 $[0,1]$。当式 (3.26) 取值为 1 时，意味着对任意 $\boldsymbol{\sigma}=\{\sigma_1,\ldots,\sigma_m\}$ $\sigma_i\in\{-1,+1\}$，有

$$\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)=1 . \tag{3.27}$$

在有限假设空间的情况下，即 $\exists\, h\in\mathcal{H},\ \text{s.t.}\ h(\boldsymbol{x}_i)=\sigma_i,\ 1\le i\le m$，类似于 $|\mathcal{H}_{|D}|=2^m$，也就有 $\Pi_{\mathcal{H}}(m)=2^m$，即 $\mathcal{H}$ 能打散 D 。

总的来说，式 (3.26) 的值越接近 1，假设空间的表示能力越强。

经验Rademacher复杂度 [Koltchinskii, 2001]

考虑实值函数空间 $\mathcal{F}: \mathcal{Z} \to \mathbb{R}$，令 $Z = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$，其中 $\boldsymbol{z}_i \in \mathcal{Z}$，将式(3.27)中的 $\mathcal{X}$ 和 $\mathcal{H}$ 替换为 $\mathcal{Z}$ 和 $\mathcal{F}$ 可得 [Koltchinskii, 2001]

定义 3.3 函数空间 $\mathcal{F}$ 关于 Z 的**经验 Rademacher 复杂度**为

$$\hat{R}_Z(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\boldsymbol{z}_i)\right]. \tag{3.28}$$

这里的 Z 是一个给定的集合。经验Rademacher复杂度衡量了函数空间 $\mathcal{F}$ 与随机噪声在数据集 Z 中的相关性。

Rademacher复杂度

对从分布 $\mathcal{D}$ 独立同分布采样得到的大小为 m 的集合 Z 求期望可得

定义 3.4 函数空间 $\mathcal{F}$ 关于 $\mathcal{Z}$ 在分布 $\mathcal{D}$ 上的 Rademacher **复杂度**为

$$R_m(\mathcal{F}) = \mathbb{E}_{Z \subset \mathcal{Z}:|Z|=m} \left[\hat{R}_Z(\mathcal{F}) \right]. \tag{3.29}$$

- Rademacher复杂度依赖于具体学习问题及数据分布，是为具体学习问题量身定制的。
- 基于Rademacher复杂度可以比基于VC维推导出更紧的泛化误差界。
- 需要注意到，在Rademacher复杂度的定义中 σ 是 $\{-1, +1\}$ 上服从均匀分布的随机变量。如果将均匀分布改为其他分布，会得到其他一些复杂度的定义。

Rademacher复杂度

经验Gauss复杂度 [Bartlett and Mendelson, 2003]

定义 3.5 函数空间 $\mathcal{F}$ 关于 Z 的**经验 Gauss 复杂度**为

$$\hat{G}_Z(\mathcal{F}) = \mathbb{E}_{\boldsymbol{g}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} g_i f(\boldsymbol{z}_i) \right], \tag{3.30}$$

其中$\boldsymbol{g} = \{g_1, \ldots, g_m\}$服从高斯分布，即标准正态分布。

Gauss复杂度：对经验Gauss复杂度求期望可得

定义 3.6 函数空间 $\mathcal{F}$ 关于 $\mathcal{Z}$ 在分布 $\mathcal{D}$ 上的 **Gauss 复杂度**为

$$G_m(\mathcal{F}) = \mathbb{E}_{Z \subset \mathcal{Z}:|Z|=m} \left[\hat{G}_Z(\mathcal{F}) \right]. \tag{3.31}$$

Rademacher复杂度与增长函数

Rademacher复杂度与前面介绍VC维用到的增长函数之间也具有一定的关系，首先引入定理 [Massart, 2000]：

定理 3.4 令 $A \subset \mathbb{R}^m$ 为有限集合且 $r = \max_{\boldsymbol{x} \in A} \|\boldsymbol{x}\|_2$，有

$$\mathbb{E}_{\boldsymbol{\sigma}} \left[\frac{1}{m} \sup_{\boldsymbol{x} \in A} \sum_{i=1}^{m} \sigma_i x_i \right] \leqslant \frac{r\sqrt{2 \ln |A|}}{m}, \tag{3.32}$$

其中 σ_i 为独立且符合 $\{-1, +1\}$ 上均匀分布的随机变量，x_i 为向量 $\boldsymbol{x}$ 的分量。

Rademacher复杂度与增长函数

证明： 对任意 $t > 0$ 使用 Jensen不等式 (1.8) 可得

$$\begin{aligned}
\exp\left(t\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) &\leqslant \mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(t\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\exp\left(t\sum_{i=1}^{m}\sigma_i x_i\right)\right] \\
&\leqslant \sum_{\boldsymbol{x}\in A}\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(t\sum_{i=1}^{m}\sigma_i x_i\right)\right]. \qquad (3.33)
\end{aligned}$$

Rademacher复杂度与增长函数

接着使用 σ 的独立性及Hoeffding不等式 (1.26) 可得

$$\begin{aligned}
\exp\left(t\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) &\leqslant \sum_{\boldsymbol{x}\in A}\prod_{i=1}^{m}\mathbb{E}_{\sigma_i}\left[\exp(t\sigma_i x_i)\right] \\
&\leqslant \sum_{\boldsymbol{x}\in A}\prod_{i=1}^{m}\exp\left(\frac{t^2(2x_i)^2}{8}\right) \\
&= \sum_{\boldsymbol{x}\in A}\exp\left(\frac{t^2}{2}\sum_{i=1}^{m}x_i^2\right) \\
&\leqslant \sum_{\boldsymbol{x}\in A}\exp\left(\frac{t^2r^2}{2}\right) \\
&= |A|\exp\left(\frac{t^2r^2}{2}\right). \qquad (3.34)
\end{aligned}$$

Rademacher复杂度与增长函数

对不等式两边取对数，可得

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leqslant \frac{\ln|A|}{t}+\frac{tr^2}{2}. \tag{3.35}$$

当 $t=\frac{\sqrt{2\ln|A|}}{r}$ 时右侧取最小值，即

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leqslant r\sqrt{2\ln|A|}. \tag{3.36}$$

两边除以 m 定理得证。

Rademacher复杂度与增长函数

由定理 3.4 可得关于Rademacher复杂度与增长函数之间关系的推论：

推论 3.1 假设空间 $\mathcal{H}$ 的 Rademacher 复杂度 $R_m(\mathcal{H})$ 与增长函数 $\Pi_{\mathcal{H}}(m)$ 满足

$$R_m(\mathcal{H}) \leqslant \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(m)}{m}}. \tag{3.37}$$

Rademacher复杂度与增长函数

证明： 对于$D=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\}$，$\mathcal{H}_{|D}$为假设空间$\mathcal{H}$在$D$上的限制。由于$h\in\mathcal{H}$的值域为$\{-1,+1\}$，可知$\mathcal{H}_{|D}$中的元素为模长$\sqrt{m}$的向量。因此，由定理 3.4 可得

$$\begin{aligned} R_m(\mathcal{H}) &= \mathbb{E}_D\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{u\in\mathcal{H}_{|D}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i u_i\right]\right] \\ &\leqslant \mathbb{E}_D\left[\frac{\sqrt{m}\sqrt{2\ln|\mathcal{H}_{|D}|}}{m}\right]. \end{aligned} \tag{3.38}$$

又因为$|\mathcal{H}_{|D}|\leq\Pi_{\mathcal{H}}(m)$，有

$$R_m(\mathcal{H}) \leqslant \mathbb{E}_D\left[\frac{\sqrt{m}\sqrt{2\ln\Pi_{\mathcal{H}}(m)}}{m}\right] = \sqrt{\frac{2\ln\Pi_{\mathcal{H}}(m)}{m}}, \tag{3.39}$$

定理得证。

分析实例之一（以线性超平面为例）

线性超平面

对于二分类问题，线性超平面假设空间 $\mathcal{H}$ 可以表示为

$$\left\{h_{\boldsymbol{w},b} : h_{\boldsymbol{w},b}(\boldsymbol{x}) = \operatorname{sign}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right) = \operatorname{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) + b\right)\right\}, \qquad (3.40)$$

$b = 0$ 时为齐次线性超平面。

典型的线性超平面是将 $\boldsymbol{w}, b$ 放缩后，满足 $\min_{(\boldsymbol{x},y)\in D} |\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b| = 1$ 的 $(\boldsymbol{w}, b)$ 所构成的超平面。

对线性超平面进行VC维和Rademacher复杂度分析，有

定理 3.5 $\mathbb{R}^d$ 空间中，由齐次线性超平面构成的假设空间的 VC 维为 d.

线性超平面的VC维

证明： 令 $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$ 表示 $\mathbb{R}^d$ 中的 d 个单位向量，$D = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$。对于任意$d$个标记 $y_1, \ldots, y_d$，取 $\boldsymbol{w} = (y_1, \cdots, y_d)$，则有 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{e}_i = y_i$，所以 D 能被齐次线性超平面打散。

令 $D' = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1}\}$ 为 $\mathbb{R}^d$ 中任意 $d+1$个向量，则必存在不全为 0 的实数 $a_1, \ldots, a_{d+1}$使得 $\sum_{i=1}^{d+1} a_i \boldsymbol{x}_i = 0$。令 $I = \{i : a_i > 0\}$, $J = \{j : a_j < 0\}$，则 I, J 至少一个不为空集。首先假设两者都不为空集，则有

$$\sum_{i \in I} a_i \boldsymbol{x}_i = \sum_{j \in J} |a_j| \boldsymbol{x}_j. \tag{3.41}$$

下面采用反证法。假设 D' 能被 $\mathcal{H}$ 打散，则存在向量 $\boldsymbol{w}$ 使得 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i > 0, i \in I$, 且 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j < 0, j \in J$。由此可得

$$0 < \sum_{i \in I} a_i (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w}) = \left(\sum_{i \in I} a_i \boldsymbol{x}_i \right)^{\mathrm{T}} \boldsymbol{w} = \left(\sum_{j \in J} |a_j| \boldsymbol{x}_j \right)^{\mathrm{T}} \boldsymbol{w} = \sum_{j \in J} |a_j| (\boldsymbol{x}_j^{\mathrm{T}} \boldsymbol{w}) < 0. \tag{3.42}$$

此式矛盾，即 D' 能被 $\mathcal{H}$ 打散不成立。当 I, J 只有一个不为空集时同理可证。

综上可知，$VC(\mathcal{H}) = d$，定理得证。

线性超平面的VC维

定理 3.6 $\mathbb{R}^d$ 空间中，由非齐次线性超平面构成的假设空间的 VC 维为 $d+1$.

证明： 由定理 3.5 的证明可知 $D=\{0,\boldsymbol{e}_1,\ldots,\boldsymbol{e}_d\}$ 能被非齐次线性超平面 $\mathcal{H}$打散。下面将非齐次线性超平面转化为齐次线性超平面：

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=\boldsymbol{w}'^{\mathrm{T}}\boldsymbol{x},\quad \boldsymbol{w}\in\mathbb{R}^d,\boldsymbol{x}\in\mathbb{R}^d,\boldsymbol{w}'\in\mathbb{R}^{d+1},\boldsymbol{x}'\in\mathbb{R}^{d+1}, \tag{3.43}$$

其中 $\boldsymbol{w}'=(\boldsymbol{w};b)$, $\boldsymbol{x}'=(\boldsymbol{x};1)$ 。如果 $D'=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{d+2}\}$ 能被 $\mathbb{R}^d$ 中非齐次线性超平面打散，则 $D''=\{\boldsymbol{x}'_1,\ldots,\boldsymbol{x}'_{d+2}\}$ 能被 $\mathbb{R}^{d+1}$ 中齐次线性超平面打散，这与定理3.5矛盾。

因此，非齐次线性超平面的VC维为 $d+1$ 。

线性超平面的Rademacher复杂度

线性超平面的假设空间复杂度不仅可基于VC维进行刻画，还可基于Rademacher复杂度刻画。但Rademacher复杂度与数据相关，因此在计算Rademacher复杂度时需要将分布 $\mathcal{D}$ 限制在某一范围内。

定理 3.7 令 $D \subset \{\boldsymbol{x} : \|\boldsymbol{x}\| \leqslant r\}$ 是大小为 m 的数据集, 则典型超平面族 $\mathcal{H} = \{\boldsymbol{x} \mapsto \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} : \|\boldsymbol{w}\| \leqslant \Lambda\}$ 的经验 Rademacher 复杂度满足

$$\widehat{R}_D(\mathcal{H}) \leqslant \sqrt{\frac{r^2\Lambda^2}{m}}. \tag{3.44}$$

不难发现，定理 3.7 只给出了Rademacher复杂度的上界。这是由于Rademacher复杂度依赖于数据分布，使得计算Rademacher复杂度的具体数值相当困难。

线性超平面的Rademacher复杂度

证明：

$$
\begin{aligned}
\widehat{R}_D(\mathcal{H}) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup \sum_{i=1}^{m} \sigma_i \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i\right] \\
&= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup \boldsymbol{w}^{\mathrm{T}} \sum_{i=1}^{m} \sigma_i \boldsymbol{x}_i\right] \\
&\leqslant \frac{\Lambda}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{m} \sigma_i \boldsymbol{x}_i\right\|\right] \\
&\leqslant \frac{\Lambda}{m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{m} \sigma_i \boldsymbol{x}_i\right\|^2\right]\right]^{1/2} \\
&= \frac{\Lambda}{m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{m} \sigma_i \sigma_j (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j)\right]\right]^{1/2} \\
&\leqslant \frac{\Lambda}{m}\left[\sum_{i=1}^{m} \|\boldsymbol{x}_i\|^2\right]^{1/2} \\
&\leqslant \sqrt{\frac{r^2\Lambda^2}{m}}, \qquad (3.45)
\end{aligned}
$$

分析实例之二（以SVM为例）

SVM

由于原样本空间往往是线性不可分的，SVM通常需要将原样本空间映射到可分的高维空间，并在高维空间中训练线性超平面进行分类。由定理 3.6 可知，若高维空间的维度为 d，则在高维空间中SVM的VC维为 $d+1$ 。

在实际运用中，映射后的高维空间维数通常很大甚至接近无穷，使得依赖空间维度的VC维失去了实际意义。这时就需要一种与空间维数无关的VC维进行刻画。虽然这种刻画方法与空间维数无关，但仍需要对超平面加以限制。

对于限制后的超平面可以得到下面的定理 [Vapnik, 1998]：

定理 3.8 令 $D \subset \{\boldsymbol{x} : \|\boldsymbol{x}\| \leqslant r\}$，超平面族 $\{\boldsymbol{x} \mapsto \mathrm{sign}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) : \min_{\boldsymbol{x}\in D} |\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}| = 1 \wedge \|\boldsymbol{w}\| \leqslant \Lambda\}$ 的 VC 维 d 满足下面的不等式

$$d \leqslant r^2\Lambda^2. \tag{3.46}$$

SVM的VC维

证明： 令 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}$ 为能被超平面族打散的集合，则对于任意 $y = (y_1, \cdots, y_d) \in \{-1, +1\}^d$ 存在 $\boldsymbol{w}$ 使得

$$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i) \geqslant 1, 1 \leqslant i \leqslant d. \tag{3.47}$$

对这些不等式求和可得

$$d \leqslant \boldsymbol{w}^{\mathrm{T}} \sum_{i=1}^{d} y_i \boldsymbol{x}_i \leqslant \|\boldsymbol{w}\| \left\| \sum_{i=1}^{d} y_i \boldsymbol{x}_i \right\| \leqslant \Lambda \left\| \sum_{i=1}^{d} y_i \boldsymbol{x}_i \right\|. \tag{3.48}$$

由于式 (3.48) 对任意 $y \in \{-1, +1\}^d$ 都成立，对其两边按 $y_1, \ldots, y_d$ 服从 $\{-1, +1\}$ 独立且均匀的分布取期望可得

SVM的VC维

$$
\begin{aligned}
d &\leqslant \Lambda \mathbb{E}_y \left[\left\| \sum_{i=1}^{d} y_i \boldsymbol{x}_i \right\| \right] \\
&\leqslant \Lambda \left[\mathbb{E}_y \left[\left\| \sum_{i=1}^{d} y_i \boldsymbol{x}_i \right\|^2 \right] \right]^{1/2} \\
&= \Lambda \left[\sum_{i,j=1}^{d} \mathbb{E}_y [y_i y_j] (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j) \right]^{1/2} \\
&= \Lambda \left[\sum_{i=1}^{d} (\boldsymbol{x}_i \cdot \boldsymbol{x}_i) \right]^{1/2} \\
&\leqslant \Lambda \left[d r^2 \right]^{1/2} \\
&= \Lambda r \sqrt{d}, \qquad (3.49)
\end{aligned}
$$

从而可知，$\sqrt{d} \leq \Lambda r$，定理得证。

分析实例之三（以多层神经网络为例）

多层神经网络

在多分类情况下，当标记集合为 $\mathcal{Y}$ 时，函数族 $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ 的增长函数有如下性质：

$$\Pi_{\mathcal{F}}(m) = \max_{D \subset \mathcal{X}} \left| \mathcal{F}_{|D} \right|, \tag{3.50}$$

其中 D 是大小为 m 且独立同分布从 $\mathcal{X}$ 中采样得到的训练集，易知 $\Pi_{\mathcal{F}}(m) \leq |\mathcal{Y}|^m$。

引理 3.2 令 $\mathcal{F}^{(1)} \subset \mathcal{Y}_1^{\mathcal{X}}$, $\mathcal{F}^{(2)} \subset \mathcal{Y}_2^{\mathcal{X}}$ 为两个函数族, $\mathcal{F} = \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$ 为它们的笛卡尔积, 有

$$\Pi_{\mathcal{F}}(m) \leqslant \Pi_{\mathcal{F}^{(1)}}(m) \cdot \Pi_{\mathcal{F}^{(2)}}(m) \tag{3.51}$$

分析实例之三（以多层神经网络为例）

引理 3.2 令 $\mathcal{F}^{(1)} \subset \mathcal{Y}_1^{\mathcal{X}}$, $\mathcal{F}^{(2)} \subset \mathcal{Y}_2^{\mathcal{X}}$ 为两个函数族, $\mathcal{F} = \mathcal{F}^{(1)} \times \mathcal{F}^{(2)}$ 为它们的笛卡尔积, 有

$$\Pi_{\mathcal{F}}(m) \leqslant \Pi_{\mathcal{F}^{(1)}}(m) \cdot \Pi_{\mathcal{F}^{(2)}}(m) \tag{3.51}$$

证明： 对于大小为 m 且独立同分布从 $\mathcal{X}$ 中采样得到的训练集 $D \subset \mathcal{X}$，根据笛卡尔积的定义可知

$$\begin{aligned} \left|\mathcal{F}_{|D}\right| &= \left|\mathcal{F}_{|D}^{(1)}\right| \left|\mathcal{F}_{|D}^{(2)}\right| \\ &\leqslant \Pi_{\mathcal{F}^{(1)}}(m) \cdot \Pi_{\mathcal{F}^{(2)}}(m). \end{aligned} \tag{3.52}$$

由 D 的任意性，引理得证。

分析实例之三（以多层神经网络为例）

引理 3.3 令 $\mathcal{F}^{(1)} \subset \mathcal{Y}_1^{\mathcal{X}}$, $\mathcal{F}^{(2)} \subset \mathcal{Y}_2^{\mathcal{X}}$ 为两个函数族, $\mathcal{F} = \mathcal{F}^{(2)} \circ \mathcal{F}^{(1)}$ 为它们的复合函数族, 有

$$\Pi_{\mathcal{F}}(m) \leqslant \Pi_{\mathcal{F}^{(2)}}(m) \cdot \Pi_{\mathcal{F}^{(1)}}(m). \tag{3.53}$$

证明： 对于大小为 m 且独立同分布从 $\mathcal{X}$ 中采样得到的训练集 $D \subset \mathcal{X}$，根据笛卡尔积的定义可知

$$\begin{aligned}\mathcal{F}_{|D} &= \left\{ \left(f_2\left(f_1(\boldsymbol{x}_1)\right), \ldots, f_2\left(f_1(\boldsymbol{x}_m)\right) \right) \middle| f_1 \in \mathcal{F}^{(1)}, f_2 \in \mathcal{F}^{(2)} \right\} \\ &= \bigcup_{\boldsymbol{u}_i \in \mathcal{F}_{|D}^{(1)}} \left\{ \left(f_2(\boldsymbol{u}_1), \ldots, f_2(\boldsymbol{u}_m) \right) \middle| f_2 \in \mathcal{F}^{(2)} \right\}. \end{aligned} \tag{3.54}$$

分析实例之三（以多层神经网络为例）

因此有

$$
\begin{aligned}
|\mathcal{F}_{|D}| &\leqslant \sum_{\boldsymbol{u}_i \in \mathcal{F}^{(1)}_{|D}} \left| \left\{ \left(f_2(\boldsymbol{u}_1), \ldots, f_2(\boldsymbol{u}_m) \right) \middle| f_2 \in \mathcal{F}^{(2)} \right\} \right| \\
&\leqslant \sum_{\boldsymbol{u}_i \in \mathcal{F}^{(1)}_{|D}} \Pi_{\mathcal{F}^{(2)}}(m) \\
&= \left| \mathcal{F}^{(1)}_{|D} \right| \cdot \Pi_{\mathcal{F}^{(2)}}(m) \\
&\leqslant \Pi_{\mathcal{F}^{(2)}}(m) \cdot \Pi_{\mathcal{F}^{(1)}}(m). \qquad (3.55)
\end{aligned}
$$

由 D 的任意性，引理得证。

分析实例之三（以多层神经网络为例）

一般来说，神经网络中的每个结点 v 计算一个函数

$$\varphi\left((\boldsymbol{w}^{(v)})^{\mathrm{T}}\boldsymbol{x}-\theta^{(v)}\right), \tag{3.56}$$

其中 φ 被称为激活函数，$\boldsymbol{w}^{(v)}$ 是与结点 v 相关的权值参数，$\theta^{(v)}$ 是与结点 v 相关的阈值参数，φ 以 $\boldsymbol{x}$ 为输入，输出激活信号。本节主要考虑使用符号激活函数 $\varphi(t)=\mathrm{sign}(t)$ 的多层神经网络。假设输入空间 $\mathcal{X}=\mathbb{R}^{d_0}$，一个$l$层的多层网络可以简化为一系列映射的复合：

$$f_l\circ\cdots\circ f_2\circ f_1(\boldsymbol{x}), \tag{3.57}$$

其中

$$\begin{aligned} &f_i:\mathbb{R}^{d_{i-1}}\mapsto\{\pm 1\}^{d_i}, 1\leqslant i\leqslant l-1\\ &f_l:\mathbb{R}^{d_{l-1}}\mapsto\{\pm 1\}. \end{aligned} \tag{3.58}$$

分析实例之三（以多层神经网络为例）

f_i 是一个多维到多维的映射，可以将其分解为若干个二值多元函数，对于 f_i 的每个分量 $f_{i,j}: \mathbb{R}^{d_{i-1}} \mapsto \{\pm 1\}$ 表示为：

$$f_{i,j}(\boldsymbol{u}) = \mathrm{sign}\left((\boldsymbol{w}^{i,j})^{\mathrm{T}} \cdot \boldsymbol{u} - \theta^{i,j}\right), \tag{3.59}$$

其中 $\boldsymbol{w}^{i,j} \in \mathbb{R}^{d_i - 1}$, $\theta^{i,j} \in \mathbb{R}$ 分别为关于第 i 层第 j 个结点的权值参数与阈值参数，将多元函数 $f_{i,j}(\boldsymbol{u})$ 的函数族记为 $\mathcal{F}^{(i,j)}$，关于第 i 层的函数族可以表示为

$$\mathcal{F}^{(i)} = \mathcal{F}^{(i,1)} \times \cdots \times \mathcal{F}^{(i,d_i)}, \tag{3.60}$$

从而整个多层神经网络的函数族可以表示为

$$\mathcal{F} = \mathcal{F}^{(l)} \circ \cdots \circ \mathcal{F}^{(2)} \circ \mathcal{F}^{(1)}. \tag{3.61}$$

分析实例之三（以多层神经网络为例）

根据引理 3.1，引理 3.2 和定理 3.1 可得

$$\Pi_{\mathcal{F}}(m) \leqslant \prod_{i=1}^{l} \Pi_{\mathcal{F}^{(i)}}(m) \leqslant \prod_{i=1}^{l} \prod_{j=1}^{d_i} \Pi_{\mathcal{F}^{(i,j)}}(m)$$

$$\leqslant \prod_{i=1}^{l} \prod_{j=1}^{d_i} \left(\frac{e \cdot m}{d_{i-1}+1} \right)^{d_{i-1}+1}. \tag{3.62}$$

令

$$N = \sum_{i=1}^{l} \sum_{j=1}^{d_i} (d_{i-1}+1) \tag{3.63}$$

表示整个多层神经网络的参数数目，可以将式 (3.62) 化简为

$$\Pi_{\mathcal{F}}(m) \leqslant (e \cdot m)^N, \tag{3.64}$$

多层神经网络的VC维

进一步可以计算出 $\mathcal{F}$ 的 VC 维的界:

定理 3.9 令 $\mathcal{F}$ 表示对应多层神经网络的函数族, 其 VC 维 $VC(\mathcal{F}) = O(N\log_2(N))$.

证明: 假设能被 $\mathcal{F}$ 打散的最大样本集大小为 d, 则 $\Pi_{\mathcal{F}}(d) = 2^d$, 由式 (3.64) 可得

$$2^d \leqslant (de)^N, \tag{3.65}$$

化简得知 $d = O\big(N\log_2(N)\big)$, 定理得证。