

信息论

——笔记整理

BarryMafu

2025 年 10 月 24 日

前言

2025 秋季，王立威教授的机器学习课程。

仍在施工中，请带好安全头盔！



BarryMafu

2025 年 10 月 24 日

目录

第一章 集中不等式	1
1.1 前置数学知识	1
1.2 Chernoff 界诱导的集中不等式	3
1.3 一则应用与推广	6
第二章 关于泛化的 VC 理论	7
2.1 机器学习的数学描述	7
2.2 有限假设空间下的结果	8
2.3 无穷假设空间下的放缩	9
2.4 VC 维度	11
附录 A 作业题目及解答	13
附录 B 学期课题	14
附录 C 往年题选及解答	15

第一章 集中不等式

1.1 前置数学知识

在介绍不等式前，先回顾一些概念. 以后会使用如下记号：

定义 1.1.1. 对于命题（或随机变量） u ，定义示性函数 (*indicator function*)

$$\mathbb{I}[u] = \begin{cases} 1 & , u \text{ 成立} \\ 0 & , \text{否则} \end{cases}$$

接下来，回顾一下概率论中的一些结论

定理 1.1.2. (*Markov* 不等式) 随机变量 X 满足 $X \geq 0$ ，且 $\mathbb{E}[X] < \infty$ ，则对于任意 $k \geq 0$ ，都有

$$\Pr[X \geq k] \leq \frac{\mathbb{E}[X]}{k}$$

定理 1.1.3. (*Chebyshev* 不等式) 随机变量 X 满足 $\mathbb{E}[X] < \infty$, $\text{Var}(X) = \sigma^2$ ，则对于任意 $k \geq 0$ ，都有

$$\Pr[|X - \mathbb{E}[X]| \geq k] \leq \frac{\sigma^2}{k^2}$$

布置了练习

练习 1.1.4. 已知随机变量 $X \sim \mathcal{N}(0, 1)$ ，定义

$$\Phi(t) := \Pr[X \geq t] = \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-\tau^2/2} d\tau$$

可以证明 Φ 并不是初等函数，现求一个初等函数 $f \sim \Phi$ ，即二者渐进等价。

分析 先分析一下这个问题：显然 $\Phi(t) \rightarrow 0$ ，所以要 $f \rightarrow 0$ 。现在 Φ 的形式非初等不便于分析，不妨考虑 $\Phi'(t) = \varphi(t)$ ，为此可以运用 L'Hospital 法则：

$$\lim_{t \rightarrow +\infty} \frac{\Phi(t)}{f(t)} = \lim_{t \rightarrow +\infty} \frac{\Phi'(t)}{f'(t)} = \lim_{t \rightarrow +\infty} \frac{-Ce^{-t^2/2}}{f'(t)}$$

其中 C 是某常数。我们希望 $f \sim \Phi$ ，也就是上述极限为 1，所以为了化简，我们希望 $f'(t)$ 中也出现 $e^{-t^2/2}$ 的形式。回忆到

$$v(x)e^{u(x)} = \left(v'(x) + u'(x)v(x) \right) \cdot e^{u(x)}$$

因此我们不妨设 $f(t)$ 形如 $g(t)e^{-t^2/2}$ ，此时

$$\lim_{t \rightarrow +\infty} \frac{-Ce^{-t^2/2}}{f'(t)} = \lim_{t \rightarrow +\infty} \frac{-Ce^{-t^2/2}}{[g'(t) - tg(t)] \cdot e^{-t^2/2}} = \lim_{t \rightarrow +\infty} \frac{-C}{g'(t) - tg(t)}$$

欲使上式为 1，就要

$$\lim_{t \rightarrow +\infty} tg(t) - g'(t) = \frac{1}{C}$$

简便起见只考虑 $C = 1$ ，之后再给 g 乘上系数。注意！这里千万不要把其当作 $-g'(t) + tg(t) = 1$ 这样的一阶线性常微分方程求解，因为其解不保证初等。如果你尝试求解 ODE 会发现解得 $f = \Phi$ 确实不初等。在这里，我们只需要考虑到 $t \rightarrow +\infty$ ，所以我们令 $g(t) = t^{-1}$ 即合意。

解答. 构造初等函数

$$f(t) = \frac{1}{\sqrt{2\pi}} \cdot \frac{e^{-t^2/2}}{t}$$

根据 L'Hospital 法则，可以验证

$$\lim_{t \rightarrow +\infty} \frac{\Phi(t)}{f(t)} = \lim_{t \rightarrow +\infty} \frac{\Phi'(t)}{f'(t)} = \lim_{t \rightarrow +\infty} \frac{-e^{-t^2/2}}{\left(-\frac{1}{t^2} - t \cdot \frac{1}{t}\right) e^{-t^2/2}} = \lim_{t \rightarrow \infty} \frac{1+t^2}{t^2} = 1$$

因此 f 和 Φ 渐进等价。 ■

事实上，我们上面给出的是 Mills 渐近展开的首项，完整的是：

$$\Phi(t) \sim \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} \cdot \left(\frac{1}{t} - \frac{1}{t^3} + \frac{1 \cdot 3}{x^5} - \frac{1 \cdot 3 \cdot 5}{x^7} + \cdots \right)$$

推论 1.1.5. (*Markov 不等式推论*) 随机变量 X , 其矩 $\mathbb{E}[X], \mathbb{E}[X^2], \dots, \mathbb{E}[X^r]$ 均存在, 则对于任意 $k \geq 0$, 都有

$$\Pr[X \geq k] \leq \min_{t \in [r]} \frac{\mathbb{E}[X^t]}{k^t}$$

定义 1.1.6. (*矩母函数*) 对于随机变量 X , 定义矩母函数 (*MGT, Moment Generating Function*) 如下

$$M_X(t) := \mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}[X^k]$$

定理 1.1.7. (*Chernoff 界*) 对于随机变量 X , 其矩母函数存在, 则对于任意 $k \geq 0$

$$\Pr[X \geq k] \leq \inf_{t>0} \frac{M_X(t)}{e^{tk}}$$

1.2 Chernoff 界诱导的集中不等式

现在考虑随机变量 X, X_1, X_2, \dots i.i.d., 服从 Bernoulli 分布 $B(1, p)$. 对于 $\delta > 0$, 一方面使用 Chebyshev 不等式; 另一方面使用中心极限定理 (CLT) 并结合 [练习 1.1.4](#) 的结果, 不难推出

$$\Pr \left[\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| \geq \delta \right] \leq \begin{cases} \frac{p(1-p)}{n\delta^2} = \mathcal{O}\left(\frac{1}{n}\right) & (\text{Chebyshev}) \\ e^{-\mathcal{O}(n)} & (\text{CLT}) \end{cases}$$

Chebyshev 仅使用了二阶矩的信息, 得到的结果太松弛了. 而 CLT 利用了完整的分布信息, 但其得到的结果在数学上并不严谨, 因为 CLT 需要 $n \rightarrow \infty$. 接下来, 我们借鉴 CLT 的方法使用 Chernoff 界给出一个紧致且严格的证明.

定理 1.2.1. 随机变量 X, X_1, X_2, \dots i.i.d. 服从 Bernoulli 分布 $B(1, p)$, 则对于任意 $\delta > 0$

$$\Pr \left[\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| \geq \delta \right] \leq e^{-\mathcal{O}(n)}$$

证明. 根据 Chernoff 界, 有

$$\Pr \left[\frac{1}{n} \sum_{i=1}^n X_i - p \geq \delta \right] = \Pr \left[\sum_{i=1}^n X_i \geq n(p + \delta) \right] \leq \inf_{t>0} e^{-nt(p+\delta)} \mathbb{E} [e^{t \sum X_i}]$$

而计算可得

$$\mathbb{E} [e^{t \sum X_i}] = \prod_{i=1}^n \mathbb{E} [e^{t X_i}] = (\mathbb{E} [e^{t X}])^n = (pe^t + (1-p))^n$$

令 $A = e^{p+\delta}$, 则有

$$\inf_{t>0} e^{-nt(p+\delta)} \mathbb{E} [e^{t \sum X_i}] = \left(\inf_{t>0} \frac{pe^t + 1 - p}{A^t} \right)^n = e^{-\mathcal{O}(n)}$$

至此证毕. □

下面介绍一些信息论相关的记号 (熵):

定义 1.2.2. (熵) 对于随机变量 X , 设其服从分布列 $p = (p_1, \dots, p_n)$, 则称其熵 (entropy) 为

$$H(X) := \sum_{i=1}^n p_i \log_2 p_i \text{ (bits)} = \sum_{i=1}^n p_i \ln p_i \text{ (nats)}$$

定义 1.2.3. (相对熵, KL 散度) 对于两个分布列 $P = (p_1, \dots, p_n)$ 和 $Q = (q_1, \dots, q_n)$, 定义其相对熵 (relative entropy) 为

$$D(P \| Q) := \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$$

定义 1.2.4. (Bernoulli 相对熵) 对于两个 Bernoulli 分布的分布列 $P = (p, 1-p), Q = (q, 1-q)$, 定义其 Bernoulli 相对熵为

$$D_B(p \| q) := D(P \| Q)$$

有了这个定义, 我们可以将**定理1.2.1**定量地写成

定理 1.2.5. (*Chernoff*) 随机变量 X, X_1, X_2, \dots *i.i.d.* 服从 *Bernoulli* 分布 $B(1, p)$, 则对于任意 $\delta > 0$

$$\Pr \left[\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| \geq \delta \right] \leq e^{-n \cdot D_B(p+\delta||p)}$$

注意到如果 $\mathbb{E}[X] = p$ 且 $X \in [0, 1]$, 那么根据 Jensen 不等式有

$$\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t(X \cdot 1 + (1-X) \cdot 0)}] \leq \mathbb{E}[Xe^{t \cdot 1}] + \mathbb{E}[(1-X)e^{t \cdot 0}] = pe^t + 1 - p$$

所以套用之前的方法能得到更普适的结果:

推论 1.2.6. (*Chernoff*) 随机变量 X, X_1, X_2, \dots *i.i.d.* 满足 $X \in [0, 1]$ 且 $\mathbb{E}[X] = p$, 则对于任意 $\delta > 0$

$$\Pr \left[\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| \geq \delta \right] \leq e^{-n \cdot D_B(p+\delta||p)}$$

事实上, 继续使用 Jensen 不等式能否给出更宽泛的结果:

推论 1.2.7. 随机变量 X_1, X_2, \dots 两两独立, 满足 $X_i \in [0, 1]$. 记 $p_i = \mathbb{E}[X_i]$, $p = \frac{1}{n} \sum_i p_i$, 则对于任意 $\delta > 0$

$$\Pr \left[\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| \geq \delta \right] \leq e^{-n \cdot D_B(p+\delta||p)}$$

究其本质而言, 这是由于中心极限定理保证了均值的分布趋于正态分布. 而正态分布的“拖尾”是指数级下降的, 因而保证了整体指数级集聚.

另外, 在实际应用当中, 我们常常做放缩 $D_B(p + \delta||p) \geq 2\delta^2$. 这个放缩在 $p \approx \frac{1}{2}$ 时较为接近, 而在 $p \approx 0$ 或 1 时较为松弛.

Chernoff 界还有一个著名的推广:

定理 1.2.8. (*Hoeffding* 不等式) 设随机变量 X_1, X_2, \dots 两两独立, 满足 $X_i \in [a_i, b_i]$ (其中 $-\infty < a_i < b_i < \infty$). 记 $p_i = \mathbb{E}[X_i]$, $p = \frac{1}{n} \sum_i p_i$, 则对于任意 $\delta > 0$

$$\Pr \left[\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| \geq \delta \right] \leq \exp \left\{ \frac{2n^2 \delta^2}{\sum_{i=1}^n (b_i - a_i)^2} \right\}$$

值得强调的是, **独立**是集中不等式成立的重要条件, 如果没有该条件 (例如取 $X_1 = X_2 = \dots$) 那么平均分布就是原分布, 并不会出现“集中”性的表现.

1.3 一则应用与推广

考虑这样一个场景：有 N 个比特（可考虑成球）， a_1, a_2, \dots, a_N ($a_i \in \{0, 1\}$). 现在希望随机抽取 n 次，我们有两种抽取方式：有放回 (draw with replacement) 和无放回 (draw without replacement). 我们记有放回的结果为 X_1, \dots, X_n ；无放回的结果为 Y_1, \dots, Y_n .

对于有放回的情形 $\{X_i\}$ ，每次抽取都是独立同分布的 Bernoulli 采样，因此可以归约到 Chernoff 界. 但对于无放回的情形 $\{Y_i\}$ ，还有没有集中不等式呢？注意到这里我们打破了独立这一条件， $\{Y_i\}$ 之间应该是负相关的（直觉上也是容易想见的）.

不妨来比较 $\{X_i\}$ 和 $\{Y_i\}$ 的收敛情况，用一步 Chernoff 不等式，我们实际上希望比较

$$\mathbb{E}[e^{t(X_1 + \dots + X_n)}] \quad \text{和} \quad \mathbb{E}[e^{t(Y_1 + \dots + Y_n)}]$$

可以证明右式是小于等于左式的，意味着 $\{Y_i\}$ 有着更强的集中（收敛）性. 具体细节太繁琐了，以下是证明的大致思路：

$$\mathbb{E}[e^{t(X_1 + \dots + X_n)}] = 1 + t \sum_i \mathbb{E}[X_i] + \frac{t^2}{2} \sum_{i,j} \mathbb{E}[X_i X_j] + \frac{t^3}{6} \sum_{i,j,k} \mathbb{E}[X_i X_j X_k] + \dots$$

$$\mathbb{E}[e^{t(Y_1 + \dots + Y_n)}] = 1 + t \sum_i \mathbb{E}[Y_i] + \frac{t^2}{2} \sum_{i,j} \mathbb{E}[Y_i Y_j] + \frac{t^3}{6} \sum_{i,j,k} \mathbb{E}[Y_i Y_j Y_k] + \dots$$

零阶和一阶项都是一样的，来关注二阶项. 其中形如 $\mathbb{E}[X_i^2]$ 和 $\mathbb{E}[Y_i^2]$ 的项是一样的，故只需比较

$$\sum_{i < j} \mathbb{E}[X_i X_j] \quad \text{和} \quad \sum_{i < j} \mathbb{E}[Y_i Y_j]$$

而对于任意的 $i < j$ 都有

$$\begin{aligned} \mathbb{E}[Y_i Y_j] &= \Pr[Y_i = 1, Y_j = 1] = \Pr[Y_i = 1] \Pr[Y_j = 1 | Y_i = 1] \\ &\leq \Pr[Y_i = 1] \Pr[Y_j = 1] = \Pr[X_i = 1] \Pr[X_j = 1] = \mathbb{E}[X_i X_j] \end{aligned}$$

于是便可以证明原不等式，进而给出了 $\{Y_i\}$ 的集中不等式.

第二章 关于泛化的 VC 理论

泛化，指的是学习到的模型对于未知数据的预测能力。半世纪前，Vapnik-Chervonenkis 理论（简称 VC 理论）被提出，尝试从数学角度定量地刻画了所谓泛化能力。值得一提，现如今 VC 理论被指出并不能完整地刻画“泛化”，即仍有该理论未囊括的额外因素，但其思想是重要且值得介绍的。

2.1 机器学习的数学描述

首先介绍一个基本概念：

定义 2.1.1.（模型，函数类，假设空间）给定输入空间 \mathcal{X} 和输出空间 \mathcal{Y} ，那么由其确定的模型 (*Model*)，函数类 (*Function Class*) 或假设空间 (*Hypothesis Space*)（这三者是同义词）为

$$\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathcal{Y}\}$$

例如：全体二次函数、线性函数或 CNN 都可以叫做模型。

现在考虑一个简单的有监督学习的模型，有数据集 $(x_1, y_1), \dots, (x_n, y_n)$ ，其中 $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$ ，所有数据 i.i.d. 且服从 $D_{X,Y}$ 。我们称每个 x_i 为**实例** (instance)，每个 y_i 为**标签** (label)。

有了数据，便要进行训练。这里便可以看出假设空间的重要性了，我们总是在假设空间中进行学习，而非天马行空毫无约束。记假设空间为 \mathcal{F} ，现在我们要选择 $\hat{f} \in \mathcal{F}$ 使得其在数据集 $\{(x_i, y_i)\}_{i=1}^n$ 上有一个较小的损失 (loss) 或误差 (error)。

那么如何评判 \hat{f} 的好坏？一个直观上的评估就是在 $D_{X,Y}$ 上的错误率.

定义 2.1.2. (训练误差) 有监督学习的情况下, 对于数据集 $\{(x_i, y_i)\}_{i=1}^n$ 和假设 \hat{f} , 训练误差 (*Training Error*) 为

$$\frac{1}{n} \sum_{i=1}^n \mathbb{I}[y_i \neq \hat{f}(x_i)]$$

定义 2.1.3. (泛化误差) 有监督学习的情况下, 设数据 (X, Y) 服从分布 $D_{X,Y}$, 对于假设 \hat{f} , 泛化误差 (*Generalization Error*) 为

$$\Pr_{(X,Y) \sim D_{X,Y}} [Y \neq \hat{f}(X)]$$

自然地, 我们可以得到泛化差距

定义 2.1.4. 承之前所有记号, \hat{f} 的泛化差距 (*Generalization Gap*) 为

$$\Pr_{(X,Y) \sim D_{X,Y}} [Y \neq \hat{f}(X)] - \frac{1}{n} \sum_{i=1}^n \mathbb{I}[y_i \neq \hat{f}(x_i)]$$

如果我们记 $Z_i := \mathbb{I}[y_i \neq \hat{f}(x_i)]$ 以及 $Z := \mathbb{I}[Y \neq \hat{f}(X)]$. 那么泛化差距就可以写成 $\mathbb{E}[Z] - \frac{1}{n} \sum_i Z_i$. 由于每个 Z_i 和 Z 都服从某个 Bernoulli 分布, 这看起来似乎就像是我们在前一章集中不等式当中描述的样子. 那么泛化差距应该随着数据集的大小 n 的增长指数级收敛至 0. 也就意味着泛化永远成立? 但事实并非如此, 这个 \hat{f} 是由 $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$ 确定的, 因此 \hat{f} 依赖于 \mathcal{D} . 此时 Z_1, \dots, Z_n 根本不独立, 甚至某些程度上正相关, 因此不能应用 Chernoff 界. (另外, 回忆一下 1.3 节中负相关才能放缩, 正相关时不一定成立)

接下来探讨何时泛化差距会比较小.

2.2 有限假设空间下的结果

先来考虑假设空间 \mathcal{F} 是有限集的情况 ($|\mathcal{F}| < \infty$). 请注意该情况是过于简化的, 因为线性模型都是无穷集. 记 $\mathcal{F} = \{f_1, \dots, f_{|\mathcal{F}|}\}$, 考虑算法最差的情况下对于任意 $\varepsilon > 0$

$$\begin{aligned}
& \Pr_{\text{worst case}} \left[\Pr_{(X,Y) \sim D_{X,Y}} [Y \neq \hat{f}(X)] - \frac{1}{n} \sum_{i=1}^n \mathbb{I}[y_i \neq \hat{f}(x_i)] \geq \varepsilon \right] \\
& \leq \sum_{j=1}^{|\mathcal{F}|} \Pr \left[\Pr_{(X,Y) \sim D_{X,Y}} [Y \neq f_j(X)] - \frac{1}{n} \sum_{i=1}^n \mathbb{I}[y_i \neq f_j(x_i)] \geq \varepsilon \right] \\
& \leq |\mathcal{F}| \cdot e^{-2n\varepsilon^2}
\end{aligned}$$

可以看到 \mathcal{F} 有限时泛化误差几乎必然随着 n 增大而收敛到一个较小的值. 但这样的分析不足以支撑 $|\mathcal{F}|$ 是无穷的情况, 这就需要使用 VC 理论了.

2.3 无穷假设空间下的放缩

现在考虑 $|\mathcal{F}| = \infty$ 的情况, 依旧有

$$\Pr_{\text{worst case}} \left[\Pr_{(X,Y) \sim D_{X,Y}} [Y \neq \hat{f}(X)] - \frac{1}{n} \sum_{i=1}^n \mathbb{I}[y_i \neq \hat{f}(x_i)] \geq \varepsilon \right] \quad (2.1)$$

$$\leq \Pr \left[\sup_{f \in \mathcal{F}} \Pr_{(X,Y) \sim D_{X,Y}} [Y \neq f(X)] - \frac{1}{n} \sum_{i=1}^n \mathbb{I}[y_i \neq f(x_i)] \geq \varepsilon \right] \quad (2.2)$$

我们想区分假设空间的大小关系 (需要一个比基数更好的度量方式), 例如五次函数类应该比线性函数类要大. 逐步来讨论, 先证明一个引理

引理 2.3.1. 设 $X, X_1, \dots, X_n, X_{n+1}, \dots, X_{2n}$ 是 *i.i.d* 的随机变量且服从 *Bernoulli* 分布. 记 $\nu_1 = \frac{1}{n} \sum_{i=1}^n X_i, \nu_2 = \frac{1}{n} \sum_{i=n+1}^{2n} X_i$, 并设 $\mathbb{E}[X] = p$. 若 $n > \varepsilon^{-2} \ln 2$, 则

$$\frac{1}{2} \Pr \left[|\nu_1 - p| \geq 2\varepsilon \right] \leq \Pr \left[|\nu_1 - \nu_2| \geq \varepsilon \right] \leq 2 \Pr \left[|\nu_1 - p| \geq \frac{\varepsilon}{2} \right]$$

证明. 我们分两部分证明.

- ① 右半式较为简单, 若 $|\nu_1 - \nu_2| \geq \varepsilon$, 那么必然有 $|\nu_1 - p| \geq \frac{\varepsilon}{2}$ 或 $|\nu_2 - p| \geq \frac{\varepsilon}{2}$, 而后面二者概率相等, 故

$$\Pr \left[|\nu_1 - \nu_2| \geq \varepsilon \right] \leq 2 \Pr \left[|\nu_1 - p| \geq \frac{\varepsilon}{2} \right]$$

② 再来看左半式, 由于 ν_1 和 ν_2 是两个独立的随机变量, 所以

$$\begin{aligned}\Pr[|\nu_1 - \nu_2| \geq \varepsilon] &\geq \Pr[|\nu_1 - p| \geq 2\varepsilon, |\nu_2 - p| \leq \varepsilon] \\ &= \Pr[|\nu_1 - p| \geq 2\varepsilon] \cdot \Pr[|\nu_2 - p| \leq \varepsilon]\end{aligned}$$

而根据集中不等式**定理1.2.5** (加上放缩) 可知

$$\Pr[|\nu_2 - p| \leq \varepsilon] \geq 1 - e^{-2n\varepsilon^2} \geq \frac{1}{2}$$

综上所述.

□

以上引理称作双重样本技巧 (Double Sample Trick), 使用该技巧可以将**公式2.2**进一步放缩为

$$\leq 2 \Pr_{(x_i, y_i)_{i=1}^{2n}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \mathbb{I}[y_i \neq f(x_i)] - \frac{1}{n} \sum_{i=n+1}^{2n} \mathbb{I}[y_i \neq f(x_i)] \geq \frac{\varepsilon}{2} \right]$$

现在不等式的形式看起来简单许多, 记 $z_i := (x_i, y_i)$, 及 $\phi_f(z_i) := \mathbb{I}[y_i \neq f(x_i)]$. 这个形式是极度对称的, 想象我们施加一个随机的置换 $\sigma \in S_{2n}$, 得到 $z_{\sigma(1)}, z_{\sigma(2)}, \dots, z_{\sigma(2n)}$, 这边的前 n 项和后 n 项还是和原来一样 i.i.d, 所以我们可以将上式展开成期望:

$$\begin{aligned}&\Pr_{z_1, \dots, z_{2n}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \phi_f(z_i) - \frac{1}{n} \sum_{i=n+1}^{2n} \phi_f(z_i) \geq \varepsilon' \right] \\ &= \mathbb{E}_{z_1, \dots, z_{2n}} \left[\Pr_{\sigma \in S_{2n}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \phi_f(z_{\sigma(i)}) - \frac{1}{n} \sum_{i=n+1}^{2n} \phi_f(z_{\sigma(i)}) \geq \varepsilon' \right] \right]\end{aligned}$$

抛开期望来看, 考虑中间的概率, 可以看作取定了数据 $\{z_i\}$, 再随机打乱 (shuffle). 而随机打乱也可以看成逐个不放回地取出, 即**小节1.3**的内容, 应用该结论可知如果没有 \sup 号的话, 有 $\Pr \leq e^{-\mathcal{O}(n\varepsilon^2)}$.

引入记号:

$$N^{\mathcal{F}}(z_1, \dots, z_m) := \left| \{(\phi_f(z_1), \dots, \phi_f(z_m)) : f \in \mathcal{F}\} \right|$$

值得注意, 由于集合中的每个元素可看作 0-1 比特串, 故该集合显然有限. 这样一来就解决了放缩 \sup 时问题, 于是原期望可以进一步放缩

$$\begin{aligned}
&\leq \mathbb{E}_{z_1, \dots, z_{2n}} \left[N^{\mathcal{F}}(z_1, \dots, z_{2n}) \cdot \Pr_{\sigma \in S_{2n}} \left[\frac{1}{n} \sum_{i=1}^n \phi_f(z_{\sigma(i)}) - \frac{1}{n} \sum_{i=n+1}^{2n} \phi_f(z_{\sigma(i)}) \geq \varepsilon' \right] \right] \\
&\leq \mathbb{E}_{z_1, \dots, z_{2n}} \left[N^{\mathcal{F}}(z_1, \dots, z_{2n}) \cdot e^{-\mathcal{O}(n\varepsilon'^2)} \right] \\
&= e^{-\mathcal{O}(n\varepsilon'^2)} \mathbb{E}_{z_1, \dots, z_{2n}} [N^{\mathcal{F}}(z_1, \dots, z_{2n})]
\end{aligned}$$

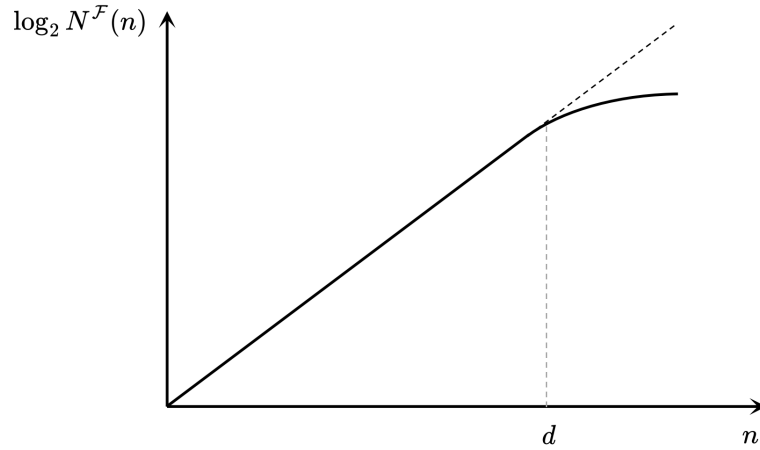
继续使用记号简化, 记

$$N^{\mathcal{F}}(m) := \max_{z_1, \dots, z_m} N^{\mathcal{F}}(z_1, \dots, z_m)$$

最终可以给出**原式2.2**的一个上界 $2N^{\mathcal{F}}(2n)e^{-\mathcal{O}(n\varepsilon^2)}$.

2.4 VC 维度

很显然 $N^{\mathcal{F}}(n) \leq 2^n$, 但可能从某一刻起可能就无法取到上界, 变成 $N^{\mathcal{F}}(n) < 2^n$ (如下图). 下面具体讨论这一现象.



我们设 d 是使得 $N^{\mathcal{F}}(n) = 2^n$ 成立的 n 的上确界. 现在想要对于 $n > d$ 给出 $N^{\mathcal{F}}(n)$ 的一个上界. 我们可以不妨考虑 0^{d+1} 不能被取到的情况, 直接得到以下命题 (至于为什么可以如此不妨假设, 等待之后详细补充, 现暂时留作思考)

命题 2.4.1. 对于 $n > d$, 有

$$N^{\mathcal{F}}(n) \leq \sum_{k=0}^d \binom{n}{k} = \mathcal{O}(n^d)$$

下面正式给出 VC 维度的定义

定义 2.4.2. (VC 维度)

附录 A 作业题目及解答

附录 B 学期课题

附录 C 往年题选及解答