

淡蓝小点技术系列：统计学习理论的本质精炼介绍

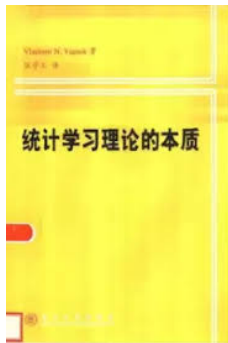
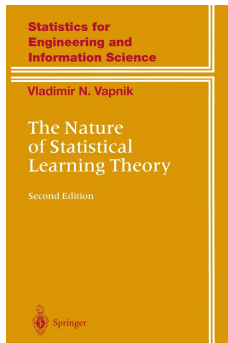
淡蓝小点Bluedotdot

微信: bluedotdot.cn

2024 年 4 月 30 日



本系列主要参考了Vapnik的《The Nature of Statistical Learning Theory》（第二版）及其译本（第一版）





- 人类的智慧：从实例中学习，通过对已知事物的总结分析得出规律，并对未发生的事物做出预测。这种能力被称为推广能力或泛化能力（**generalization**）。
- 人工智能：希望能用机器（计算机）来模拟这种学习能力，就是所谓机器学习。其目的是设计某种算法，它能够通过对已有数据的学习找到数据内在的依赖关系，从而对未知数据做预测。



1960s之前，概率统计学的两大核心基础包括：[大数定律](#)和[中心极限定理](#)。而这两大基础讨论都是样本数量趋于无穷时，试验中所呈现的规律。

- 大数定律：当试验次数趋于无穷时，样本出现的频率将收敛于事件发生的概率。不同的收敛强度对应不同的大数定律。
- 中心极限定理：当试验次数趋于无穷时，无论随机变量服从何种分布，随机变量之和将收敛于正态分布。

但对于现实中的学习问题，我们不可能得到无穷多样本。因此，1960至1980年间，统计学领域出现了一场革命：强调小样本统计学问题。



对于一种未知的函数依赖关系，如果我们想要基于已有观测（样本）去估计这种依赖关系，我们需要提前知道关于这一依赖关系的哪些信息？

- 基于Fisher的传统理论体系中，我们需要知道除了该依赖关系（概率分布函数）的有限个参数值以外的几乎所有一切信息。
- 新的学习理论下，只需要知道该依赖关系（分布函数/密度函数）所属于函数集的某些一般性质即可。



学习理论的主要内容包括：

- ① 统计推断一致性的充分必要条件及其相关概念
- ② 学习机器学习过程中收敛速度的界
- ③ 小样本下对泛化能力的控制（针对小样本的归纳推理原则）
- ④ 对各类问题的归纳推理（如模式识别、函数估计等）

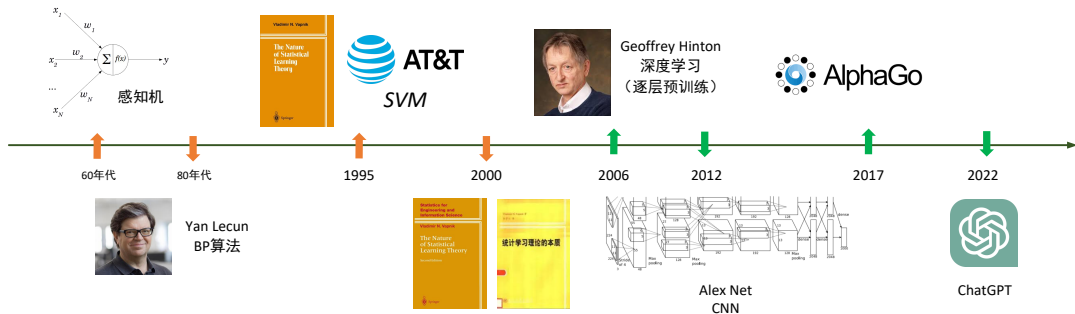


名词解释

- 学习理论: learning theory
- 依赖关系: dependency
- 函数集: set of functions
- 泛化能力: generalization
- 收敛: converge
- 一致性: consistency
- 归纳推理: inductive inference
- 原则: principle
- 非平凡: non-trivial
- 渐近性: asymptotic

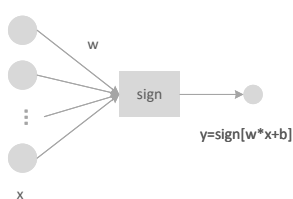


学习理论诞生的历史阶段

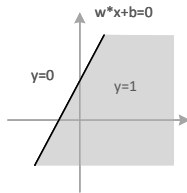




感知机的诞生（60s）→ 学习理论基础（60~70s）→ 神经网络（80s）→ 神经网络替代方法（90s） 感知机
由Rosenblatt基于McCulloch-Pitts模型提出



Perceptron 感知机



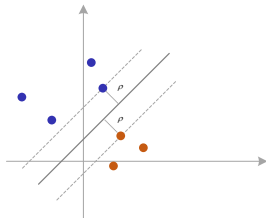
- 感机器被看作是第一个可学习的机器
- 尝试将多个感知机组合起来
- 不知如何同时选择参数值



感知机的诞生（60s）→ 学习理论基础（60~70s）→ 神经网络（80s）→ 神经网络替代方法（90s）

学习过程分析始于Novikoff关于感知机的一个定理，这也是学习理论的开始。假设训练数据可分并且多层感知机中，最后一层的输入为 \mathbf{z} （前几层将 \mathbf{x} 变换为 \mathbf{z} ），那么若

- ① 如果 \mathbf{z} 的范数是有界的，如 $|\mathbf{z}| \leq R$
- ② 训练数据能被间隔 ρ 分开： $\sup_w \min_i y_i (\mathbf{z}_i \cdot \mathbf{w}) > \rho$
- ③ 感知机经过充分的训练



那么最多经过 $N \leq \lceil \frac{R^2}{\rho^2} \rceil$ 次对 \mathbf{w} 的修正后，超平面就能将训练数据分开



感知机的诞生（60s）→ 学习理论基础（60~70s）→ 神经网络（80s）→ 神经网络替代方法（90s）

Novikoff证明了如果数据可分，那么感知机就一定能把数据分开。即使是对于无限数据，只要数据是可分的，感知机经有限次修正后就能将数据分开。

更进一步的，如果我们设定规则：若感知机经过 $k(k = 1, 2, \dots)$ 次修正后，在接下来的 m_k 个数据内都没有再对超平面做修正就停止训练，其中

$$m_k = \frac{1 + 2 \ln k - \ln \eta}{-\ln(1 - \epsilon)}$$

① 感知机的训练会在 l 步以内停止（在 l 以内将所有数据分开），其中

$$l \leq \frac{1 + 4 \ln \frac{R}{\rho} - \ln \eta}{-\ln(1 - \epsilon)} \left[\frac{R^2}{\rho^2} \right]$$

② 停止时学到的超平面，在测试数据集上出错率小于 ϵ 的概率为 $1 - \eta$



感知机的诞生（60s）→ 学习理论基础（60~70s）→ 神经网络（80s）→ 神经网络替代方法（90s）

基于以上结论，当时的很多学者认为使学习机具有泛化性（有较小的测试误差）的唯一因素就是使它在训练集上的误差最小。至此，对学习过程的研究就演化为两个分支：

- 对学习过程的应用分析：主要目标就是寻找同时构造所有神经元系数的方法，使所得到的超平面在训练数据上的错误率最小（To get a good generalization it is sufficient to choose the coefficients of the neuron that provide the minimal number of training errors. The principle of minimizing the number of training errors is a self-evident inductive principle, and number of training errors is a self-evident inductive principle, and from the practical point of view does not need justification.）。
- 对学习过程的理论分析：寻找能够达到最好的推广性能的归纳原则，并构造算法来实现这一原则（The principle of minimizing the number of training errors is not self-evident and needs to be justified. It is possible that there exists another inductive principle that provides a better level generalization ability.）。



感知机的诞生（60s）→ **学习理论基础（60~70s）** → 神经网络（80s）→ 神经网络替代方法（90s）

在感知机提出及BP算法被应用到神经网络上的这段时间内，学习理论的基础被创立并得到很多重要的研究结论。首先是关于经验风险最小化的：

- 1968年，指示函数集（用于判定问题）的VC熵及VC维被提出，基于这些概率发现了泛函空间的大数定律，并且得到了学习过程收敛速率的非渐近界的主要结论。这些结论在1971年被完全证明并发表。
- 1976~1981年间，上述概念被推广到了实函数集，推广后的主要结论包括：实函数集上的泛函空间的大数定律、完全有界函数集和无界函数上一致收敛速率的界以及结构风险最小化。
- 1989年，找到了经验风险最小化归纳原则和极大似然估计一致性的充要条件。
- 1990年，开始研究（合成构造）能控制泛化能力的学习机



感知机的诞生（60s）→ 学习理论基础（60~70s）→ 神经网络（80s）→ 神经网络替代方法（90s）

其次是关于不适定问题的研究（ill-posed problem），不适定问题的研究对于学习理论的发展起到了重要帮助。不适定问题是基于适定性问题（well-posed problem）定义的：设求解等式方程 $Af = F, f \in \mathcal{X}, F \in \mathcal{Y}$ ，其中 $A: \mathcal{X} \rightarrow \mathcal{Y}$ 是一个算子，如果该等式的解同时满足以下三个条件，则说这个问题是适定性的

- ① 对任意一个 $F \in \mathcal{Y}$ ，该方程都有解（即 A 是满射）
- ② 该方程的解总是唯一的（ A 是单射）
- ③ 若 F 发生微小的扰动，方程的解是稳定的（或者说：当 F 发生微小变换时 f 的变化也很微小；方程解随初始条件的变化是连续的； A^{-1} 是连续的）

如果上述条件不能同时满足，就说问题是不适定的。



感知机的诞生（60s）→ 学习理论基础（60~70s）→ 神经网络（80s）→ 神经网络替代方法（90s）

不适定问题的关键是第三点。设有 F_δ ， $\|F - F_\delta\| < \delta$ ，其中 δ 是一个任意小量。如 f 和 f_δ 分别是方程的解，不适定性是说此时 $\|f - f_\delta\|$ 可能很大。

不适定问题意味着：若最小化 $R(f)$ 得到 f_δ ，则不能保证 f_δ 是 f 的一个好的近似，即使 δ 趋近于0

$$R(f) = \|Af - F_\delta\|^2$$

即使因果关系是一一映射的，问题仍有可能是不适定的



感知机的诞生（60s）→ 学习理论基础（60~70s）→ 神经网络（80s）→ 神经网络替代方法（90s）

到1960s人们发现，最小化带正则项的泛函 $R(f)$ ，当 δ 趋近于0时，其结果能收敛到想要的结果 f 上。

$$R^*(f) = \|Af - F_\delta\|^2 + \gamma(\delta)\Omega(f)$$

这里的 $\Omega(f)$ 一种特殊类型的泛函， $\gamma(\delta)$ 一个恰当选择的常数，它的值与噪声水平 δ 有关。

最小化泛函 $R(f)$ 原本以为是不证自明（self-evident）的，结果它无法保证是正确的；最小化泛函 $R^*(f)$ 不是不证自明的，但是它反而是正确的。



感知机的诞生（60s）→ **学习理论基础（60~70s）** → 神经网络（80s）→ 神经网络替代方法（90s）

第三个方面是密度估计的非参数方法（**nonparametric methods**）的出现。传统学习理论的研究是从一个较小泛围的函数（也就是参数模型**parametric model**）中去找到一个合适的密度函数，采用的方法是传统意义上“不证自明”的方法如极大似然估计。当将候选函数从较小的泛围拓展到较宽泛的泛围后，原先的方法就不适用了，因此人们提出了新的方法，其中就利用正则化技术。

现在，我们已经可以从一个较宽泛的函数集中去估计目标函数了。



感知机的诞生（60s）→ 学习理论基础（60~70s）→ 神经网络（80s）→ 神经网络替代方法（90s）

第四个方面是算法复杂度理论的发展。激发算法复杂度理论的是两个基本问题：

- ① 归纳推理的本质是什么？（What is the nature of inductive inference?）
- ② 随机性的本质是什么？（What is the nature of randomness?）

Kolmogorov对随机性的理解：对于一个长度为 l 的很长的数字串，如果找不到一个程序能够生成该数字串并且该程序的复杂度低于 l ，那么就可认为这串数字是随机的。这里算法的复杂度由实现这个算法的节目的最短长度来衡量。

算法复杂度思想对于如何基于有限样本（有限数量的经验数据）去估计依赖关系（ x 和 y 的依赖关系或者样本与密度函数的依赖关系）



感知机的诞生（60s）→ 学习理论基础（60~70s）→ **神经网络（80s）** → 神经网络替代方法（90s）

在M-P模型上稍加修改就形成了神经网络：将原先的Sign函数替换为连续的Sigmoid型函数，如 $\tanh(u)$ 或 $\sigma(u)$ 。这一改动最大的意义在于可导。即可以利用BP算法去同时修正网络中所有参数的值。虽然基于梯度找到的可能是局部极小值点，但当时人们仍然觉得学习过程应用分析的主要问题已经解决了。

（但是，对于大规模的神经网络是否具有更好的泛化能力，当时仍然存疑）

1984年提出了概率近似正确（Probably Approximately Correct, PAC）理论



感知机的诞生（60s）→ 学习理论基础（60~70s）→ 神经网络（80s）→ 神经网络替代方法（90s）

人们用很大精力研究了其它方法，如径向基函数模型、SVM等。并且，统计学习理论中比较艰深的部分重新开始吸引学者的注意。

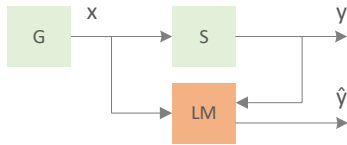
除了完成学习过程的一般分析外，人们还开始研究最优算法（对任意数量样本能得到的最高泛化能力）的合成。

- ① 前言背景
- ② 学习问题表示
- ③ 学习过程一致性
- ④ 学习过程收敛速度的界



学习理论中的学习问题是指基于有限的观察样本找到正确的依赖关系，所以学习问题也是一个函数估计问题。

- 产生器 (G)，从固定但未知的概率分布函数 $F(\mathbf{x})$ 中生成独立同分布的向量 $\mathbf{x} \in \mathbb{R}^n$
- 监督器 (S)，对每个 \mathbf{x} 对应一个 \mathbf{y} ，映射关系是确定但未知的 $F(\mathbf{y}|\mathbf{x})$
- 学习机 (LM)，一个函数集 $f(\mathbf{x}, \alpha), \alpha \in \Lambda$ ，其中 Λ 是函数参数集



学习问题就是从函数集 $f(\mathbf{x}, \alpha), \alpha \in \Lambda$ 中选择一个函数¹，它是对监督器的最佳近似。

¹例如从全体线性函数中选择一个



假设训练数据是由独立同分布的样本构成

$$(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_l, \mathbf{y}_l)$$

要得到对监督器最好的逼近，首先应能度量在给定 \mathbf{x} 时监督器的响应 \mathbf{y} 与学习机 $f(\mathbf{x}, \alpha)$ 之间的损失或差异（loss or discrepancy）。理论上损失的期望为

$$R(\alpha) = \int L(\mathbf{y}, f(\mathbf{x}, \alpha)) dF(\mathbf{x}, \mathbf{y}), \quad \alpha \in \Lambda$$

因为 $R(\alpha)$ 的取值随函数 $f(\mathbf{x}, \alpha)$ 的变化而变化，所以它被称为**风险泛函（risk functional）**。学习目标就是**最小化**风险泛函。注意，这里的 $F(\mathbf{x}, \mathbf{y})$ 是分布函数而非密度函数

$$F(\mathbf{x}, \mathbf{y}) = F(\mathbf{y}|\mathbf{x})F(\mathbf{x})$$

若分布函数对应的密度函数为 $\phi(\mathbf{x}, \mathbf{y})$ 则有

$$R(\alpha) = \int L(\mathbf{y}, f(\mathbf{x}, \alpha)) dF(\mathbf{x}, \mathbf{y}) = \int L(\mathbf{y}, f(\mathbf{x}, \alpha)) \phi(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}$$



对于三类常见的问题，其损失函数可分别写作

- 模式识别（分类）问题，假设监督器输出 $\mathbf{y} = \{0, 1\}$ ，此时 $f(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$ 就是指示函数集，损失函数可写为

$$L(\mathbf{y}, f(\mathbf{x}, \alpha)) = \begin{cases} 0, & \text{if } \mathbf{y} = f(\mathbf{x}, \alpha) \\ 1, & \text{if } \mathbf{y} \neq f(\mathbf{x}, \alpha) \end{cases}$$

- 对回归问题，损失函数可写为

$$L(\mathbf{y}, f(\mathbf{x}, \alpha)) = (\mathbf{y} - f(\mathbf{x}, \alpha))^2$$

- 对于密度估计问题，学习机候选函数为密度函数集 $p(\mathbf{x}, \alpha)$, $\alpha \in \Lambda$ ，损失函数可写为

$$L(p(\mathbf{x}, \alpha)) = -\log p(\mathbf{x}, \alpha)$$



三类问题仍有不同写法，可以进一步将三种写法统一成一种。令 $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ ，损失函数统一表示为 $Q(\mathbf{z}, \alpha)$ （ Q 可以是三种损失函数中的任意一种），那么风险泛函可表示为

$$R(\alpha) = \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}), \quad \alpha \in \Lambda$$

理论上，找到使 $R(\alpha)$ 取值最小的 α ，就相当于找到了泛化能力最优的模型。但是，因为 $F(\mathbf{z})$ （ $F(\mathbf{x}, \mathbf{y})$ ）是未知的，所以这个积分是不可求的，直接找到最小的 $R(\alpha)$ 也就是不可能的。现实中只能基于经验数据（观察数据）最小化风险泛函，把基于经验数据的风险泛函称为经验风险泛函并用 $R_{emp}(\alpha)$ 表示

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha), \quad \alpha \in \Lambda$$

通过最小化 R_{emp} 找到最优 α 的这一原则（方法）就称作经验风险最小化（empirical risk minimization）归纳原则，简称ERM原则。



具体而言，学习理论需要研究四个问题

- ① 一个基于ERM归纳原则的学习过程具有一致性的条件（充分必要条件）是什么？
- ② 学习过程收敛的速度有多快（收敛速度是否有界）？
- ③ 如何控制学习过程的收敛速度（泛化能力）？
- ④ 如何构造能够控制泛化能力的算法？

这四个问题对应的回答就构成了学习理论的四个部分

- ① 学习过程一致性理化
- ② 学习过程收敛速度的非渐近理论
- ③ 控制学习过程泛化能力的理论
- ④ 构造学习算法的理论



学习问题相对传统统计理论的两个主要转变：

- ① 待估函数集从特定函数集变为更为宽泛的函数集
- ② 理论背景从趋于无穷样本变为基于有限数量样本

最小化经验风险的归纳原则实质上就是极大似然归纳原则（二者的一致性还未说明），但极大似然有很大的局限性。假设某个密度函数是有两个高斯分量的混合高斯分布，混合系数均为 $\frac{1}{2}$ 。为了简化问题，假设其中一个分量的均值为0方差为1，另一个分量的均值为 a ，方差为 σ^2

$$p(x, a, \sigma) = \frac{1}{2\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\} + \frac{1}{2\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$



假设训练数据为 $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I\}$, 那么对于任意给定的常数 A , 总能找到一个 $\{a, \sigma\}$, 使得似然大于 A 。例如可令 $\sigma = \sigma_0, a = \mathbf{x}_1$, 则

$$\begin{aligned} L(a = \mathbf{x}_1, \sigma = \sigma_0) &= \sum_{i=1}^I \ln p(\mathbf{x}_i; a = \mathbf{x}_1, \sigma = \sigma_0) \\ &= \ln \frac{1}{2\sigma_0\sqrt{2\pi}} + \sum_{i=2}^I \ln \frac{1}{2\sigma_0\sqrt{2\pi}} \exp\left\{-\frac{(\mathbf{x}_i - \mathbf{x}_1)^2}{2\sigma_0^2}\right\} + \sum_{i=2}^I \ln \frac{1}{2\sqrt{2\pi}} \exp\left\{-\frac{x_i^2}{2}\right\} \\ &> \ln \frac{1}{2\sigma_0\sqrt{2\pi}} + \sum_{i=2}^I \ln \frac{1}{2\sqrt{2\pi}} \exp\left\{-\frac{x_i^2}{2}\right\} \\ &= -\ln \sigma_0 - \sum_{i=2}^I \frac{x_i^2}{2} - I \ln 2\sqrt{2\pi} > A \end{aligned}$$

后面两项为确定值, 只需要选择恰当的 σ_0 , 就能使似然大于 A 。这说明这种情况下, 极大似然得不到有效解。



但是最小化风险泛函仍然是有道理的。例如对于回归问题，假设监督器为 $f(\mathbf{x}, \alpha = \alpha_0) = f_0(\mathbf{x})$ ，那么有

$$\begin{aligned} R(\alpha) &= \int (\mathbf{y} - f(\mathbf{x}, \alpha))^2 dF(\mathbf{x}, \mathbf{y}) \\ &= \int (\mathbf{y} - f_0(\mathbf{x}) + f_0(\mathbf{x}) - f(\mathbf{x}, \alpha))^2 dF(\mathbf{x}, \mathbf{y}) \\ &= \int (\mathbf{y} - f_0(\mathbf{x}))^2 dF(\mathbf{x}, \mathbf{y}) + \int 2(\mathbf{y} - f_0(\mathbf{x}))(f_0(\mathbf{x}) - f(\mathbf{x}, \alpha)) dF(\mathbf{x}, \mathbf{y}) + \int (f_0(\mathbf{x}) - f(\mathbf{x}, \alpha))^2 dF(\mathbf{x}, \mathbf{y}) \end{aligned}$$

注意这里的第二项，假设 $F(\mathbf{x}, \mathbf{y})$ 对应的密度函数为 $\phi(\mathbf{x}, \mathbf{y})$ ，因为 \mathbf{y} 是由 $F(\mathbf{y}|\mathbf{x})$ 生成的，求 \mathbf{y} 基于 $\phi(\mathbf{y}|\mathbf{x})$ 的期望时正好等于 $f_0(\mathbf{x})$ 。

$$\begin{aligned} \int (\mathbf{y} - f_0(\mathbf{x}))(f_0(\mathbf{x}) - f(\mathbf{x}, \alpha)) dF(\mathbf{x}, \mathbf{y}) &= \int (f_0(\mathbf{x}) - f(\mathbf{x}, \alpha)) \left\{ \int (\mathbf{y} - f_0(\mathbf{x})) \phi(\mathbf{y}|\mathbf{x}) d\mathbf{y} \right\} \phi(\mathbf{x}) d\mathbf{x} \\ &= 0 \end{aligned}$$



所以

$$R(\alpha) = \int (\mathbf{y} - f_0(\mathbf{x}))^2 dF(\mathbf{x}, \mathbf{y}) + \int (f_0(\mathbf{x}) - f(\mathbf{x}, \alpha))^2 dF(\mathbf{x}, \mathbf{y})$$

因为第一项和 $f(\mathbf{x}, \alpha)$ 无关，所以风险泛函就等价于

$$R^*(\alpha) = \int (f_0(\mathbf{x}) - f(\mathbf{x}, \alpha))^2 dF(\mathbf{x}, \mathbf{y})$$

最小化风险泛函找到的就是和 $f_0(\mathbf{x})$ 最接近的函数



对于密度估计问题，假设 $F(\mathbf{t})$ 的密度函数为 $p_0(\mathbf{t})$

$$\begin{aligned} R(\alpha) &= - \int \ln p(\mathbf{t}, \alpha) dF(\mathbf{t}) \\ &= - \int p_0(\mathbf{t}) \ln p(\mathbf{t}, \alpha) d\mathbf{t} \end{aligned}$$

令

$$C = \int \ln p_0(\mathbf{t}) dF(\mathbf{t}) = \int p_0(\mathbf{t}) \ln p_0(\mathbf{t}) d\mathbf{t}$$

注意这一项与 $p(\mathbf{t}, \alpha)$ 无关，所以它可看作常数，因此最小化风险泛函等价于最小化 $R^*(\alpha)$

$$\begin{aligned} R^*(\alpha) &= - \int p_0(\mathbf{t}) \ln p(\mathbf{t}, \alpha) d\mathbf{t} + \int p_0(\mathbf{t}) \ln p_0(\mathbf{t}) d\mathbf{t} \\ &= - \int \ln \frac{p(\mathbf{t}, \alpha)}{p_0(\mathbf{t})} p_0(\mathbf{t}) d\mathbf{t} \\ &= \text{KL}(p_0(\mathbf{t}) \| p(\mathbf{t}, \alpha)) \end{aligned}$$

最小化风险泛函相当于找到跟 $p_0(\mathbf{t})$ 最接近的 $p(\mathbf{t}, \alpha)$ 。



既然最小化经验风险是不可靠的而最小化期望风险是合理的，那基于经验数据最小化 $R_{emp}(\alpha)$ 找到的 α ，能不能作为学习结果使用了？什么时候能，什么时候又不能了？²

- 学习过程的一致性核心就是回答：对一个经验风险最小化的学习过程，它何时能够取得较小的实际风险（泛化风险），何时又不能？或者说，研究经验风险最小化学习过程一致性的充要条件。
- 假设误差函数是二次的，它关于待估参数是凸的，通过求梯度等于零得到的点必然是全局最优，那我们为什么还需要学习理论、需要讨论学习过程的一致性了？优化理论是否足以确保学习过程最终的正确性？³

²其它问题如 $R_{emp}(\alpha)$ 趋向最优 $R(\alpha^*)$ 的速度有多快？它们之间的差是否有界可控？这个界跟样本集大小有怎样的关系？

³不能，全局（或局部）最优点只是确保 α 使 $R_{emp}(\alpha)$ 取值最小（或极小），不能确保该 α 同样使 $R(\alpha)$ 取最小



两点补充

- 用有限数量信息解决问题的基本原则：在解决一个特定问题时，要避免把解决一个更一般的问题作为其中间步骤。
密度估计是全能问题⁴，但密度估计是一个不适定问题。根据这一原则，并非所有情况都需要密度函数（判定模型和生成模型）。
- 除了ERM外，还有其它归纳原则，例如随机逼近推理原则

$$\alpha(k+1) = \alpha(k) - \gamma_k \text{grad}_{\alpha} Q(z_k, \alpha(k))$$

经验风险最小化原则和随机逼近原则是两种不同的归纳原则⁵，它们对应不同的学习理论（主要是一致性）。虽然除了ERM外，还有很多其它归纳原则，但总体而言，ERM通常更具鲁棒性（它能很好地利用经验数据，不依赖先验信息，并且实现方法也很明确）。

⁴知道了密度函数相当于知道了一切，可以解决所有问题，例如可以判定 $p(C|x)$ ，可以生成样本 $p(x)$

⁵虽然它们的目标都是最小化 R_{emp}



一致的定义：已知 $R(\alpha)$ 表示风险泛函， $R_{emp}(\alpha)$ 表示经验风险泛函，它们的定义为

$$R(\alpha) = \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}), \quad R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha)$$

假设 $Q(\mathbf{z}, \alpha_l)$ 使得 R_{emp} 取最小值，如果下面两个序列依概率收敛到同一个极限，我们就说ERM原则对于 $Q(\mathbf{z}, \alpha), \alpha \in \Lambda$ 函数集和分布函数 $F(\mathbf{z})$ 是一致的

$$R(\alpha_l) \xrightarrow[l \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha)$$

$$R_{emp}(\alpha_l) \xrightarrow[l \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha)$$

这两个式子各是什么意思？现实中是利用经验数据学习模型的，只有第二个式子行不行？⁶

⁶不行！理论上只有第二个就行，但在实际分析计算中要用到 $R(\alpha)$ ，所以也需要第一个式子。但这两个式子可以改写成 $\inf_{\alpha \in \Lambda} R(\alpha)$ 一个式子。



学习理论中的关键定理：设函数集 $Q(\mathbf{z}, \alpha), \alpha \in \Lambda$ 满足以下性质：

$$A \leq \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) \leq B, \quad (A \leq R(\alpha) \leq B)$$

那么ERM原则具有一致性的充分必要条件是经验风险 $R_{emp}(\alpha)$ 在函数集上依概率一致收敛于期望风险 $R(\alpha)$ ⁷：

$$\lim_{l \rightarrow \infty} P\left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \epsilon \right\} = 0, \quad \forall \epsilon > 0$$

这种一致收敛也被称为单边一致（uniform one-sided）收敛。ERM的一致性等价于单边一致收敛，这就将一致性问题转化为了一致收敛的问题，而且单边一致收敛考虑的是最坏情况。

为什么这里没有绝对值符号？⁸

⁷ 关键定理的意义在于将学习过程的一致性问题转换成了一致收敛问题

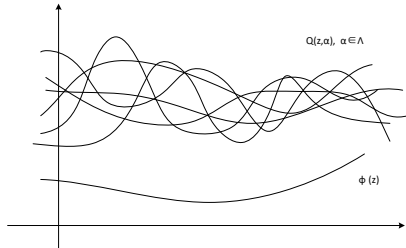
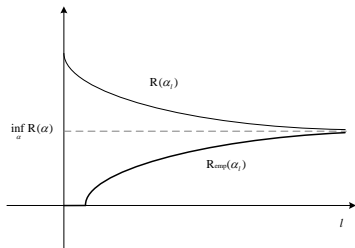
⁸ 因为式子中有取上确界的操作



平凡一致性 (trivial consistency): 假设对于函数集 $Q(\mathbf{z}, \alpha), \alpha \in \Lambda$, ERM方法不一致。现对函数集做拓展, 向函数集中加入一个函数 $\phi(\mathbf{z})$, 要求

$$\inf_{\alpha \in \Lambda} Q(\mathbf{z}, \alpha) > \phi(\mathbf{z}), \quad \forall \mathbf{z}$$

拓展后函数集具有一致性, 但它的最小值总是取到 $\phi(\mathbf{z})$ 上面的值, 因此这种一致性是无意义的。





平凡一致性基于特定元素，但我们想要的一致性函数集的一种一般化性质。非平凡一致性定义：设 $c \in (-\infty, +\infty)$ ， Λ 有子集 $\Lambda(c)$ ，其中

$$\Lambda(c) = \{\alpha : \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) > c, \quad \alpha \in \Lambda\}$$

我们说ERM对于函数集 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ 和分布函数 $F(\mathbf{z})$ 具有非平凡一致性，如果下式满足

$$\inf_{\alpha \in \Lambda(c)} R_{emp}(\alpha) \xrightarrow{P} \inf_{\alpha \in \Lambda(c)} R(\alpha)$$

为什么这里只有一个收敛式？⁹

⁹可以证明，在非平凡一致性的情况下第一个式子会被自动满足



极大似然的非平凡一致性定义：极大似然方法是非平凡一致的，如果对于任意一个密度函数 $p(\mathbf{x}, \alpha_0)$ ，给定的密度函数集 $p(\mathbf{x}, \alpha) \in \Lambda$ 具有下列依概率收敛的性质

$$\inf_{\alpha \in \Lambda} \frac{1}{l} \sum_{i=1}^l (-\log p(\mathbf{x}_i, \alpha)) \xrightarrow[l \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} \int (-\log p(\mathbf{x}, \alpha)) p(\mathbf{x}, \alpha_0) d\mathbf{x}$$

极大似然方法的关键定理：若密度函数集满足 $0 < a \leq p(\mathbf{x}, \alpha) \leq A < \infty$, $\alpha \in \Lambda^{10}$ ，极大似然方法非平凡一致的充分必要条件是该函数集上的风险函数对于任意密度函数 $p(\mathbf{x}, \alpha_0)$ 一致单边收敛

$$\lim_{l \rightarrow \infty} P\left\{ \sup_{\alpha \in \Lambda} \left(\int \log p(\mathbf{x}, \alpha) p(\mathbf{x}, \alpha_0) d\mathbf{x} - R_{emp}(\alpha) \right) > \epsilon \right\} = 0$$

¹⁰完全有界



因为单边一致收敛是ERM一致的充要条件，为了更好的研究单边一致收敛，引入两个随机过程，它们也被称为经验过程：双边经验过程和单边经验过程。

设有随机变量序列（这里的随机变量是 ξ^l ），称这一随机变量序列为依赖于概率分布 $F(\mathbf{z})$ 及函数空间 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ 的双边经验过程（two-sided empirical process）

$$\xi^l = \sup_{\alpha \in \Lambda} \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right|, \quad l = 1, 2, \dots$$

要讨论的是双边经验过程在什么条件下依概率收敛到0

$$\lim_{l \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right| > \epsilon \right\} = 0, \quad \forall \epsilon > 0$$



另一个随机过程是单边经验过程 (one-sided empirical process), 它由随机变量 ξ_+^l 的序列构成

$$\xi_+^l = \sup_{\alpha \in \Lambda} \left(\int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right)_+, \quad l = 1, 2, \dots$$

其中

$$(u)_+ = \begin{cases} u, & \text{if } u > 0 \\ 0, & \text{otherwise} \end{cases}$$

对应要讨论的是单边经验过程在什么条件下依概率收敛到0

$$\lim_{l \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} \left(\int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right) > \epsilon \right\} = 0, \quad \forall \epsilon > 0$$



前面已经说明过，单边一致收敛就是ERM一致性的充要条件！后面会看到，双边一致收敛对于构造单边一致收敛的条件很重要。

为什么 $\{\xi_+^l\}$ 的收敛性就是单边一致收敛性？（ ξ_+^l 中有取正，而单边一致收敛式中没有取正）

$$\sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right)_+ > \epsilon \stackrel{?}{=} \sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right) > \epsilon$$

原因在于： ξ_+^l 的定义中有取上确界（sup）的操作。设 $\Delta R^l = \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha)$

- 若 ΔR^l 随着 l 的增加取值总为正，那么 $\sup_{\alpha \in \Lambda} \Delta R_+^l = \sup_{\alpha \in \Lambda} \Delta R^l$
- 若 ΔR^l 随着 l 的增加取值有正有负，那么 $\sup_{\alpha \in \Lambda} \Delta R^l$ 一定为正，那么 $\sup_{\alpha \in \Lambda} \Delta R_+^l = \sup_{\alpha \in \Lambda} \Delta R^l$
- 若 ΔR^l 当 l 大于某值后总为负或总为0或有负有0，那么 $\sup_{\alpha \in \Lambda} \Delta R^l$ 不可能大于 ϵ ，所以一致单边收敛性成立



若函数集 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ 中只有一个元素 (即只有一个函数), 那么双边经验过程的随机变量 ξ^l 变为

$$\xi^l = \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right|$$

根据辛钦大数定律

$$\lim_{n \rightarrow \infty} P\left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| < \epsilon\right) = 1$$

此时双边经验过程收敛性一定满足

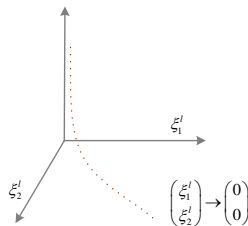
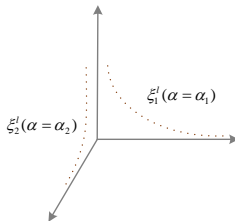
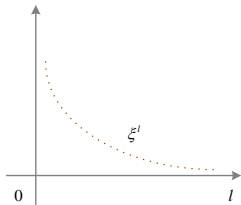
$$\lim_{l \rightarrow \infty} P\left\{ \sup_{\alpha \in \Lambda} \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right| > \epsilon \right\} = 0 \Rightarrow \lim_{l \rightarrow \infty} P\left\{ \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right| > \epsilon \right\} = 0$$

随着观察数量 l 的增加, 随机变量序列 ξ^l 收敛于 0



若函数集 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ 中有有限个元素（如 N 个），那么双边经验过程的随机变量 ξ^l 仍然依概率收敛于0。

这种情况相当于 N 维向量空间中的大数定律，函数集中的每个函数对应于一个维度，向量空间中的大数定律说明所有维度都会依概率收敛。





如果函数集 $Q(\mathbf{z}, \alpha), \alpha \in \Lambda$ 中有无限个元素情况会有质的区别，即 ξ^l 序列不一定收敛于0。因此我们要回答的问题就是：

- 函数集 $Q(\mathbf{z}, \alpha), \alpha \in \Lambda$ 和概率分布函数 $F(\mathbf{z})$ 具有何种性质时，随机变量序列 ξ^l 依概率收敛到0？

相当于我们需要泛函空间的大数定律，若泛函空间的大数定律存在则该函数集双边一致收敛。泛函空间的大数定律可看作是传统大数定律在泛函空间的推广。

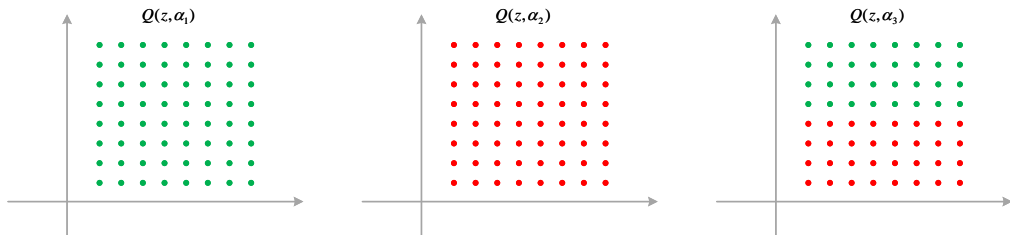
在传统的统计学中，只考虑了双边一致收敛（大数定律），但是并没有考虑单边一致收敛。单边一致收敛此处重要是因为根据关键定理它是ERM归纳原则一致性的充要条件。



一致单边收敛和一致双边收敛的充要条件都建立在一个新的概念上：在有 l 个样本的样本集上，函数集 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ 的熵。

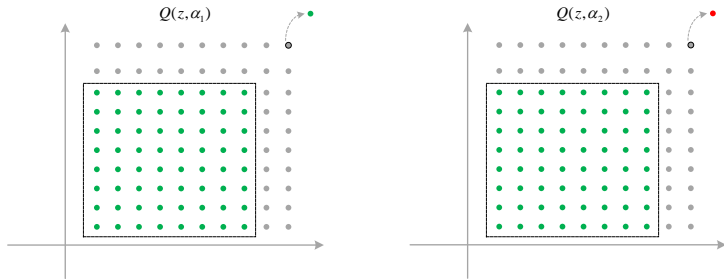
以二分类问题为例来讨论（即任意 \mathbf{z}_i 只能标记为1或者0），设 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ 是一个指示函数集，有样本 $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l\}$ ，用 $N^\Lambda(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$ 表示该指示函数集中的函数能把给定的所有样本分成多少种不同的结果。我们用这个量来表征该函数集在给定样本集上的多样性。

假设函数集有3个不同的函数（ $\alpha_1, \alpha_2, \alpha_3$ ），它们对样本的分类结果如下所示，那么此时 $N^\Lambda(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l) = 3$





因为集合中是不会有重复元素的，所以函数集中也不会有相同的函数，那是否意味着有多少个函数 $N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$ 的取值就是多少了？当然不是。因为 $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l\}$ 只是全体数据的一部分，下图中有两个不同的函数，但它们对样本集的分类结果是一样的。

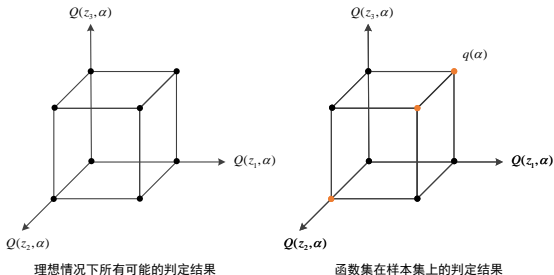




设 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ 是函数 α 在泛化数据上的判定结果（对于判定问题只能是0或1），设

$$q(\alpha) = (Q(\mathbf{z}_1, \alpha), Q(\mathbf{z}_2, \alpha), \dots, Q(\mathbf{z}_l, \alpha)), \quad \alpha \in \Lambda$$

是函数集在样本集上判定结果的集合。那么所有可能的判定结果就是 l 维超立方体的顶点数，而 $N^\Lambda(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$ 就是顶点集合的子集。例如，设只有三个样本 $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$





对于特定的函数集 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, 因为样本集中的样本是独立同分布的, 所以 $N^\Lambda(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$ 是一个随机变量。称 $H^\Lambda(\mathbf{z}_1, \dots, \mathbf{z}_l)$ 为随机熵

$$H^\Lambda(\mathbf{z}_1, \dots, \mathbf{z}_l) = \ln N^\Lambda(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$$

如果考虑概率分布函数 $F(\mathbf{z}_1, \dots, \mathbf{z}_l)$ 及所有数量为 l 的样本集, 则 $\ln N^\Lambda(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$ 的期望记为

$$H^\Lambda(l) = E[\ln N^\Lambda(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)]$$

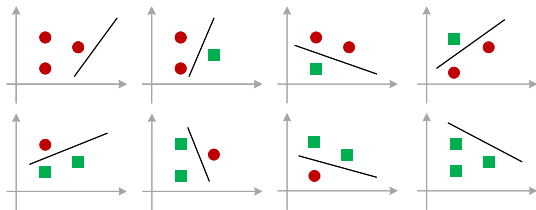
$H^\Lambda(l)$ 称为指示函数集 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ 在数量为 l 的样本集上的熵, 它反映的是指示函数集在 l 样本集上的多样性。或者说, 它反应的是对于 l 个样本, 函数空间能对这 l 个样本的多少种取值情况进行区分。这是对函数空间表现能力的一种度量。

注意, $H^\Lambda(l)$ 同时跟三方面有关: 指示函数集、概率分布函数及样本数量 (但它跟某个具体的样本集无关, 因为它要求样本集的期望)。



假设 $I = 3$ 并且这三个样本不在一条直线上，设函数集 Λ 为线性函数集，因为线性函数可以将所有的情况区分开，所以此时

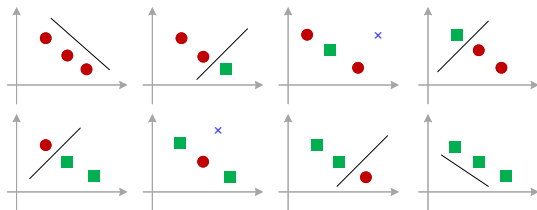
$$N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = 8, \quad H^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = \log_2(N) = 3$$





若 $l = 3$ 并且这三个样本恰好一条直线上，设函数集 Λ 为线性函数集，此时线性函数只能将六种情况区分开，所以此时

$$N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = 6, \quad H^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = \log_2(N) \approx 2.58$$

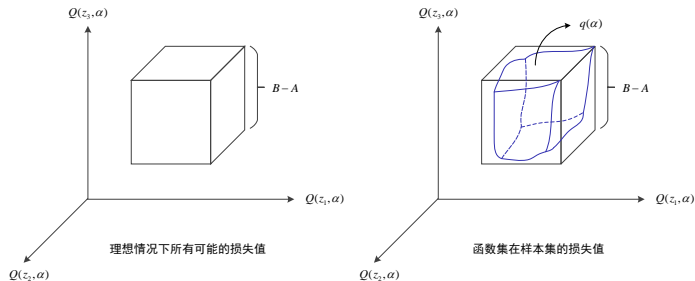


因为只有这两种情况，假设出现这两种情况的概率各为0.5，所以

$$H^{\Lambda}(l) = \mathbb{E} \ln N^{\Lambda}(\mathbf{z}_1, \dots, \mathbf{z}_l) = 0.5 * 3 + 0.5 * 2.58 = 2.79$$

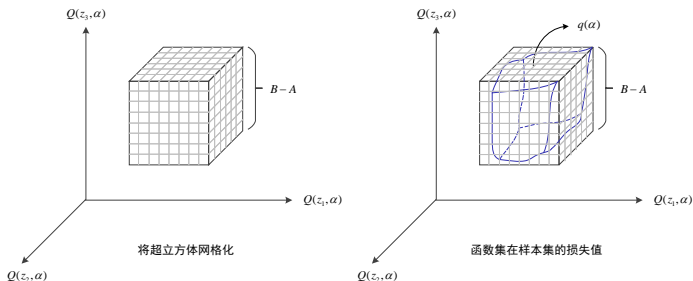


将指示函数拓展到一般实函数。设 $A \leq Q(\mathbf{z}, \alpha) \leq B, \alpha \in \Lambda$ 是一个有界损失函数集，那么所有数据在函数集下的损失值集合是一个 I 维的超立方体。而函数集在特定数据集上的损失为该超立方体内的子集。





将超立方体离散化为有限最小 ϵ -网络（简单来说就是将整个超立方体以 ϵ 为边长进行切分），那么 $N = N^{\Lambda}(\epsilon; \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$ 是向量集 $q(\alpha), \alpha \in \Lambda$ 的最小 ϵ -网络的元素数目。





类似的，因为样本集 $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l\}$ 是独立同分布的，所以 $N^\Lambda(\epsilon; \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$ 是一个随机变量。令

$$H^\Lambda(\epsilon; \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l) = \ln N^\Lambda(\epsilon; \mathbf{z}_1, \dots, \mathbf{z}_l)$$

它被称为函数集 $A \leq Q(\mathbf{z}, \alpha) \leq B, \alpha \in \Lambda$ 在样本集 $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l\}$ 上的随机VC熵。设概率分布函数为 $F(\mathbf{z}_1, \dots, \mathbf{z}_l)$ ，随机VC熵的期望

$$H^\Lambda(\epsilon; l) = \mathbb{E}[H^\Lambda(\epsilon; \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)]$$

称为函数集 $A \leq Q(\mathbf{z}, \alpha) \leq B, \alpha \in \Lambda$ 在 l 个样本上的VC熵。



实函数集的VC熵就是指示函数集熵的推广。对于指示函数集，也可以看作是将超立方体切分成 $\epsilon < 1$ 的最小 ϵ -网络，那么 $q(\alpha)$ 占据的网络数量就是超立方体顶点集的子集（因为 $\epsilon < 1$ 确保了超立方体的 l 个顶点各自属于不同的网络）。相当于有

$$N^{\Lambda}(\epsilon; \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l) = N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$$

$$H^{\Lambda}(\epsilon; \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l) = H^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_l)$$

$$H^{\Lambda}(\epsilon; l) = H^{\Lambda}(l)$$

这表明，对有界实函数集的一般结论也适用于指示函数集。



一致双边收敛的定义

$$\lim_{l \rightarrow \infty} P\left\{ \sup_{\alpha \in \Lambda} \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right| > \epsilon \right\} = 0, \quad \forall \epsilon > 0$$

对于有界实函数集一致双边收敛的充分必要条件

$$\lim_{l \rightarrow \infty} \frac{H^\Lambda(\epsilon, l)}{l} = 0, \quad \forall \epsilon > 0$$

对于指示函数集一致双边收敛的充分必要条件¹¹

$$\lim_{l \rightarrow \infty} \frac{H^\Lambda(l)}{l} = 0$$

(指示函数集是实函数集的特殊情况)

¹¹这一条件是Vapnik和Chervonenkis在1968年得到的，它对有界实函数集的推广是在1981年完成的



一致单边收敛定义

$$\lim_{l \rightarrow \infty} P\{\sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \epsilon\} = 0, \quad \forall \epsilon > 0$$

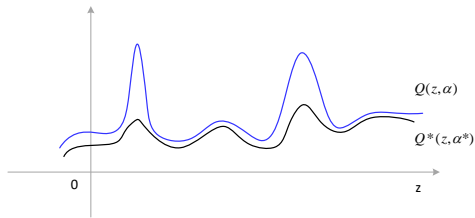
一致双边收敛定义去绝对值符号后为

$$\lim_{l \rightarrow \infty} P\{[\sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \epsilon] \text{ or } [\sup_{\alpha \in \Lambda} (R_{emp}(\alpha) - R(\alpha)) > \epsilon]\} = 0, \quad \forall \epsilon > 0$$

可见双边收敛中包含了单边收敛，所以双边收敛是单边收敛的充分条件（剩下的情况是最大化经验风险时的一致性）。



设有有界实函数集 $Q(z, \alpha), \alpha \in \Lambda$ 和一个新的函数集 $Q^*(z, \alpha^*), \alpha^* \in \Lambda^*$, 若对于概率分布函数 $F(z)$ 具有如下性质: 对 $Q(z, \alpha), \alpha \in \Lambda$ 中的任意函数, 在 $Q^*(z, \alpha^*), \alpha^* \in \Lambda^*$ 中都存在一个函数使得



$$Q(z, \alpha) - Q^*(z, \alpha^*) \geq 0, \quad \forall z$$

$$\int (Q(z, \alpha) - Q^*(z, \alpha^*)) dF(z) \leq \delta, \quad \delta > 0$$

可简单理解为: Q^* 是 Q 的下界, 并且 Q 和 Q^* 是非常接近的



完全有界函数集 $Q(\mathbf{z}, \alpha), \alpha \in \Lambda$ ，经验均值单边一致收敛于期望均值的充分必要条件是：对于任意正的 δ, η, ϵ ，存在一个函数集 $Q^*(\mathbf{z}, \alpha^*), \alpha^* \in \Lambda^*$ ，使得 Q 和 Q^* 满足前面的关系式，并且 Q^* 在 l 个样本上其 ϵ -熵（VC熵）满足下面不等式

$$\lim_{l \rightarrow \infty} \frac{H^{\Lambda^*}(\epsilon, l)}{l} < \eta$$

已知根据关键定理一致单边收敛等价于ERM的一致性，所以这一条件也是ERM一致性的充要条件（但这只是一致双边收敛的必要非充分条件，因为它只确保最小化经验风险时的一致性）



以指示函数集为例， I 个样本的VC熵为 $H^\Lambda(I) = \mathbb{E}[\ln N^\Lambda(\mathbf{z}_1, \dots, \mathbf{z}_I)]$ ，在VC熵的基础上再构造两个概念：

- 退火的VC熵： $H_{ann}^\Lambda(I) = \ln \mathbb{E}[N^\Lambda(\mathbf{z}_1, \dots, \mathbf{z}_I)]$
- 生长函数： $G^\Lambda(I) = \ln \sup_{\mathbf{z}_1, \dots, \mathbf{z}_I} N^\Lambda(\mathbf{z}_1, \dots, \mathbf{z}_I)$ （有些资料的定义没有取对数，因此这里也可以称为对数生长函数）

对于任意的 I 总有

$$H^\Lambda(I) \leq H_{ann}^\Lambda(I) \leq G^\Lambda(I)$$

左边的不等式由琴生不等式便可证明，右边的不等式因为生长函数中取的是上确界

注意 $H^\Lambda(I)$ 和 $H_{ann}^\Lambda(I)$ 是和 $F(\mathbf{z})$ 有关的，而 $G^\Lambda(I)$ 和 $F(\mathbf{z})$ 无关



学习理论的三个里程碑:

- ERM原则一致性的充分条件, 它确保随着 l 的增大ERM能收敛到泛化误差(期望风险)上

$$\lim_{l \rightarrow \infty} \frac{H^\Lambda(l)}{l} = 0$$

- ERM原则收敛的渐近速度快的充分条件, 它保证了收敛有快的渐近速度

$$\lim_{l \rightarrow \infty} \frac{H_{ann}^\Lambda(l)}{l} = 0$$

所谓收敛渐近速度快是指对于任何的 $l > l_0$, 都有下面的指数界成立

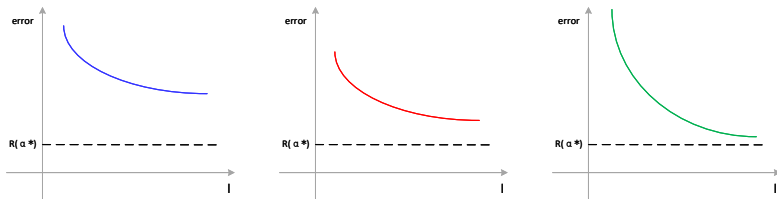
$$P\{(R(\alpha_l) - R(\alpha_0)) > \epsilon\} < e^{-c\epsilon^2 l}, \quad c > 0$$

- ERM原则是一致的且有快的收敛速度, 并且这种收敛不依赖于任何特定的概率分布的充要条件:

$$\lim_{l \rightarrow \infty} \frac{G^\Lambda(l)}{l} = 0$$



讨论一致收敛速度的界，通常考虑的是上界（下界虽然也存在但它不像上界那么重要）。这里所谓的“速度的界”，不是计算速度（计算复杂度），而是随着 l 的增加，经验风险和期望风险之间的差的界。



首先介绍两种非构造性（只能用于分析不能用于实际计算）的收敛速度界¹²

- 基于退火熵函数的依赖于分布的界
- 基于生成函数的与分布无关的界

¹²不可构造是指不可实际计算



对于指示函数（判定问题），设 $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$ 是一个指示函数集， $H^\Lambda(l)$ 是对应的VC熵， $H_{ann}^\Lambda(l)$ 是退火熵， $G^{bm\Lambda}(l)$ 是生长函数，学习过程的收敛速度有如下两个基于退火熵的基本界

$$P\left\{\sup_{\alpha \in \Lambda} \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right| > \epsilon\right\} \leq 4 \exp\left\{\left(\frac{H_{ann}^\Lambda(2l)}{l} - \epsilon^2\right)l\right\}$$

$$P\left\{\sup_{\alpha \in \Lambda} \frac{\left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right|}{\sqrt{\int Q(\mathbf{z}, \alpha) dF(\mathbf{z})}} > \epsilon\right\} \leq 4 \exp\left\{\left(\frac{H_{ann}^\Lambda(2l)}{l} - \frac{\epsilon^2}{4}\right)l\right\}$$

这两个界是非平凡的（即对于任意的 $\epsilon > 0$ ，右边会随 l 的增加逐渐趋近于0），如果下式被满足

$$\lim_{l \rightarrow \infty} \frac{H_{ann}^\Lambda(l)}{l} = 0$$

这两个界限定了期望风险和经验风险之差的上限，所以满足这两个界的函数集就可认为是可以快速收敛的。



指示函数集, $R(\alpha) = \int Q(\mathbf{z}, \alpha) dF(\mathbf{z})$ 表示的是 $\{\mathbf{z} : Q(\mathbf{z}, \alpha) = 1\}$ 的概率, 而 $R_{emp}(\alpha)$ 表示的是此类事件在样本集上的频率。所以第一个界估计的是概率和频率之间偏差的模的双边一致收敛速度。

但是很明显, 概率和频率之间偏差的大小会受方差的影响, 方差越大越容易取到上界值。以伯努利分布为例, 它的方差的定义为

$$\sigma^2 = p(1 - p) = R(\alpha)(1 - R(\alpha))$$

于是便有人提出了相对一致收敛

$$P\left\{\sup_{\alpha \in \Lambda} \frac{|R(\alpha) - R_{emp}(\alpha)|}{\sqrt{R(\alpha)(1 - R(\alpha))}} \geq \epsilon\right\} < \Phi(\epsilon, l)$$

但是这种形式下很难写出右边的具体形式。一种更简单的写法是, 注意到当 $R(\alpha)$ 取值较小时

$$\sqrt{R(\alpha)} \rightarrow \sqrt{R(\alpha)(1 - R(\alpha))}$$

所以可用 $\sqrt{R(\alpha)}$ 去替换上面的分母, 这样就有了第二个界。第二个界会比第一个界好很多。



因为已知 $H_{ann}^\Lambda(l) \leq G^{bm\Lambda}(l)$, 所以只需要用 $G^{bm\Lambda}(l)$ 替换掉 $H_{ann}^\Lambda(l)$ 就能得到基于生长函数的界 (这样的界跟具体的分布函数无关)。

$$P\left\{\sup_{\alpha \in \Lambda} \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right| > \epsilon\right\} \leq 4 \exp\left\{\left(\frac{G^\Lambda(2l)}{l} - \epsilon^2\right)l\right\}$$

$$P\left\{\sup_{\alpha \in \Lambda} \frac{\left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right|}{\sqrt{\int Q(\mathbf{z}, \alpha) dF(\mathbf{z})}} > \epsilon\right\} \leq 4 \exp\left\{\left(\frac{G^\Lambda(2l)}{l} - \frac{\epsilon^2}{4}\right)l\right\}$$

同样如果下式被满足, 那么这两个界是非平凡的

$$\lim_{l \rightarrow \infty} \frac{G^\Lambda(l)}{l} = 0$$

这个条件就是前面提到过的学习理论的第三个里程碑, 它也是双边过程快速一致收敛 (第一个界) 的充要条件。注意, 这一式子如果不被满足, 不仅意味着收敛速度不快, 而且可能存在 $F(\mathbf{z})$ 使得双边过程不收敛。



以下式为例，对界的不等式略作分析

$$P\{\sup_{\alpha \in \Lambda} |\int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha)| > \epsilon\} \leq 4 \exp\{(\frac{G^\Lambda(2l)}{l} - \epsilon^2)l\}$$

若 ϵ 取值越大，表示经验风险和期望风险的差越大； ϵ 取值越小，表示经验风险和期望风险的差越小。理想情况下，应该是 ϵ 取值越大 右边 的取值越小，表示二者出现较大误差的概率越小。很明显， G^Λ 取值越小 右边 的取值越小

这说明：函数集的生长函数越小（函数越简单），经验风险收敛到期望风险的可能性越大；函数集的生长函数越大（函数越复杂），最小化经验风险所得的结果在训练集上误差较小但在泛化数据上误差较大的可能性较大¹³。

¹³其实相当于过拟合的可能性越大



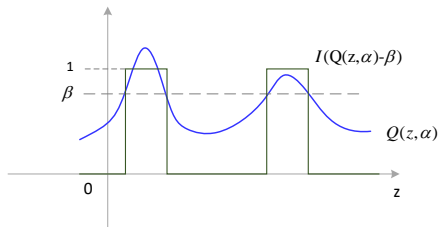
收敛速度上界可以推广到实函数集上。设 $Q(z, \alpha), \alpha \in \Lambda$ 是一个实函数集并且

$$A = \inf_{\alpha, z} Q(z, \alpha) \leq Q(z, \alpha) \leq \sup_{\alpha, z} Q(z, \alpha) = B$$

这里的 A, B 可以是无穷，所以这里不要求函数集是完全有界的。用 \mathcal{B} 表示开区间 (A, B) ，基于 $Q(z, \alpha)$ 构建指示函数

$$I(z, \alpha, \beta) = \theta(Q(z, \alpha) - \beta), \alpha \in \Lambda, \beta \in \mathcal{B}$$

θ 函数表示对于任意给定的 α^*, β^*



$$I(z, \alpha^*, \beta^*) = \begin{cases} 1, & \text{if } Q(z, \alpha^*) \geq \beta^* \\ 0, & \text{if } Q(z, \alpha^*) < \beta^* \end{cases}$$



设 $H^{\Lambda, B}(l)$ 表示对应的指示函数集的 VC 熵, 相应的 $H_{ann}^{\Lambda, B}(l)$ 和 $G^{\Lambda, B}(l)$ 分别是退火熵和生长函数。分三种情况对实函数集的收敛界做推广。

(1) 若函数集是完全有界的: 设 $A \leq Q(\mathbf{z}, \alpha) \leq B, \alpha \in \Lambda$, 其中 A, B 均是有限值, 那么下面不等式成立

$$P\left\{\sup_{\alpha \in \Lambda} \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right| > \epsilon\right\} \leq 4 \exp\left\{\left(\frac{H_{ann}^{\Lambda, B}(2l)}{l} - \frac{\epsilon^2}{(B-A)^2}\right)l\right\}$$

(2) 若函数集是完全有界且非负的: 设 $0 \leq Q(\mathbf{z}, \alpha) \leq B, \alpha \in \Lambda$, 其中 B 是有限值, 那么下面不等式成立

$$P\left\{\sup_{\alpha \in \Lambda} \frac{\left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right|}{\sqrt{\int Q(\mathbf{z}, \alpha) dF(\mathbf{z})}} > \epsilon\right\} \leq 4 \exp\left\{\left(\frac{H_{ann}^{\Lambda, B}(2l)}{l} - \frac{\epsilon^2}{4B}\right)l\right\}$$



(3) 若函数集是非负（但不一定有界）的：设 $0 \leq Q(\mathbf{z}, \alpha), \alpha \in \Lambda$ ，将 $Q(\mathbf{z}, \alpha)$ 看作一个随机变量，若该随机变量使得对某个 $p > 2$ ，其 p 阶矩开 p 次方存在

$$m_p(\alpha) = \sqrt[p]{\int Q^p(\mathbf{z}, \alpha) dF(\mathbf{z})}$$

那么下面的界成立

$$P\left\{\sup_{\alpha \in \Lambda} \frac{\int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha)}{\sqrt[p]{\int Q^p(\mathbf{z}, \alpha) dF(\mathbf{z})}} > a(p)\epsilon\right\} \leq 4 \exp\left\{\left(\frac{H_{ann}^{\Lambda, \mathcal{B}}(2l)}{l} - \frac{\epsilon^2}{4}\right)l\right\}$$

其中

$$a(p) = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2}\right)^{p-1}}$$

这里三个式子定义的界是非平凡的，如果满足

$$\lim_{l \rightarrow \infty} \frac{H_{ann}^{\Lambda, \mathcal{B}}(l)}{l} = 0$$



与指示函数集类似，将退火熵替换为生长函数就得到了与分布无关的界

(1) 对完全有界函数集: $-\infty < A \leq Q(\mathbf{z}, \alpha) \leq B, \alpha \in \Lambda < +\infty$

$$P\left\{\sup_{\alpha \in \Lambda} \left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right| > \epsilon\right\} \leq 4 \exp\left\{\left(\frac{G_{ann}^{\Lambda, B}(2l)}{l} - \frac{\epsilon^2}{(B-A)^2}\right)l\right\}$$

(2) 完全有界非负函数集: $0 \leq Q(\mathbf{z}, \alpha) \leq B, \alpha \in \Lambda < +\infty$

$$P\left\{\sup_{\alpha \in \Lambda} \frac{\left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right|}{\sqrt{\int Q(\mathbf{z}, \alpha) dF(\mathbf{z})}} > \epsilon\right\} \leq 4 \exp\left\{\left(\frac{G_{ann}^{\Lambda, B}(2l)}{l} - \frac{\epsilon^2}{4B}\right)l\right\}$$

(3) 非负实函数集 $0 \leq Q(\mathbf{z}, \alpha)$ 且存在 $p > 2$ 阶矩开 p 次方

$$P\left\{\sup_{\alpha \in \Lambda} \frac{\left| \int Q(\mathbf{z}, \alpha) dF(\mathbf{z}) - \frac{1}{l} \sum_{i=1}^l Q(\mathbf{z}_i, \alpha) \right|}{\sqrt[p]{\int Q^p(\mathbf{z}, \alpha) dF(\mathbf{z})}} > a(p)\epsilon\right\} \leq 4 \exp\left\{\left(\frac{G_{ann}^{\Lambda, B}(2l)}{l} - \frac{\epsilon^2}{4}\right)l\right\}$$

同样满足前面的条件时，这三个界是非平凡的



再回答两个问题：

- 假设函数空间中的某一项 $Q(\mathbf{z}, \alpha_l)$ 使得 $R_{emp}(\alpha_l)$ 取到最小值，那么此时期望风险 $R(\alpha_l)$ 是多少了？
- 此时 $R(\alpha_l)$ 距离 $\inf_{\alpha} R(\alpha)$ 有多大差距了？

以完全有界函数集和完全有界非负函数集为例给出这两个界，这两个界也被称为学习机推广能力的界。第三种情况非负不一定有界函数集此处忽略¹⁴。定义

$$\mathcal{E} = 4 \frac{G^{\Lambda, B}(2l) - \ln(\eta/4)}{l}$$

¹⁴第三种情况公式更复杂，此处节省时间和空间



(1) 设 $A \leq Q(\mathbf{z}, \alpha) \leq B, \alpha \in \Lambda$ 是完全有界函数集, 那么

- 下面两个式子成立 (以至少 $1 - \eta$ 的概率同时对 $Q(\mathbf{z}, \alpha), \alpha \in \Lambda$ 中所有函数成立)

$$P\{R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{(B - A)}{2} \sqrt{\mathcal{E}}\} \geq (1 - \eta)$$

$$P\{R_{\text{emp}}(\alpha) - \frac{(B - A)}{2} \sqrt{\mathcal{E}} \leq R(\alpha)\} \geq (1 - \eta)$$

- $R(\alpha_I)$ 和 $\inf_{\alpha \in \Lambda} R(\alpha)$ 之间的关系成立 (以至少 $1 - 2\eta$ 的概率成立)

$$P\{R(\alpha_I) - \inf_{\alpha \in \Lambda} R(\alpha) \leq (B - A) \sqrt{\frac{-\ln \eta}{2I}} + \frac{B - A}{2} \sqrt{\mathcal{E}}\} \geq 1 - 2\eta$$



(2) 设 $0 \leq Q(\mathbf{z}, \alpha) \leq B, \alpha \in \Lambda$ 是有界非负函数集, 那么

- 这一式子对函数空间中的所有函数成立 (以至少 $1 - \eta$ 的概率成立)

$$P\{R(\alpha) \leq R_{emp}(\alpha) + \frac{B\mathcal{E}}{2} [1 + \sqrt{1 + \frac{4R_{emp}(\alpha)}{B\mathcal{E}}}] \} \geq 1 - \eta$$

- $R(\alpha_l)$ 和 $\inf_{\alpha \in \Lambda} R(\alpha)$ 之间的关系成立 (以至少 $1 - 2\eta$ 的概率成立)

$$P\{R(\alpha_l) - \inf_{\alpha \in \Lambda} R(\alpha) \leq B\sqrt{\frac{-\ln \eta}{2l}} + \frac{B\mathcal{E}}{2} [1 + \sqrt{1 + \frac{4}{\mathcal{E}}}] \} \geq 1 - 2\eta$$



思考两个重要问题：

- 前面引入了VC熵和生长函数的概念，但是它们都是概念性的是无法具体计算的，那该怎么办了？

$$H^{\Lambda}(l) = \mathbb{E}[\ln N^{\Lambda}(\mathbf{z}_1, \dots, \mathbf{z}_l)], \quad G^{\Lambda}(l) = \ln \sup_{\mathbf{z}_1, \dots, \mathbf{z}_l} N^{\Lambda}(\mathbf{z}_1, \dots, \mathbf{z}_l)$$

- 如何量化一个模型（模型空间）对数据的表现能力（capacity）？参数数量？最高阶次数？可导次数？函数类型？



VC维是目前能够度量函数表现力（**capacity**）的最重要量化指标之一，而且理论上它是可计算的！

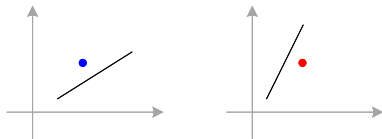
以指示函数集为例， $Q(\mathbf{z}, \alpha), \alpha \in \Lambda$ 的VC维是指：能够被集合中的函数完全打散的样本集 $\{\mathbf{z}_1, \dots, \mathbf{z}_h\}$ 的最大样本数 h （或者说 h 个样本能够被集合中的函数完全分类为 2^h 种情况的最大 h 值）¹⁵

VC维（ h 值）的大小跟两个因素有关：函数集、样本的维度数

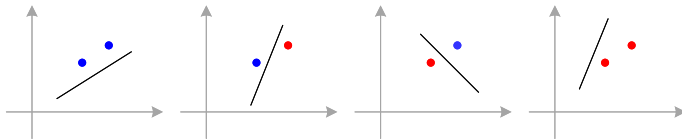
¹⁵注意： h 个样本可以有多种不同的分布情况，VC维的定义中只要求有 h 个样本能被完全分类为 2^h 种情况即可，并不要求所有分布下的 h 个样本都能被分类成 2^h 种情况；另外，很明显VC维只能是非负整数



假设样本是二维的（二维平面的点），函数集为全体线性函数。若样本集中只有一个样本，则不同的线性函数可以将该样本分成 2^1 种不同的情况

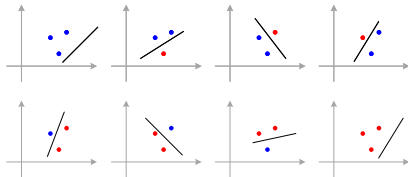


若样本集中有两个样本，同样可以用不同的直线将其分成 2^2 次方种情况

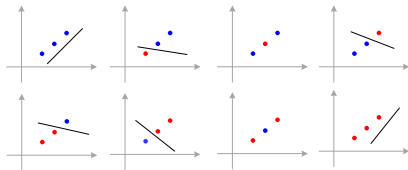




若样本集中有三个样本 ($2^3 = 8$), 则有些分布下能将样本完全打散, 有些分布下则不能

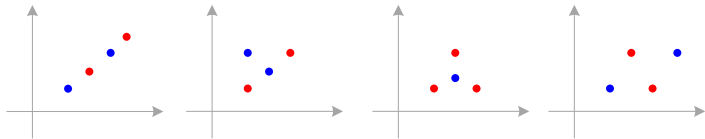


当样本分布于一条直线之上时, 不能完全分开





当有四个样本时，无论它们处于何种分布，总有一些情况是不能完全分开的



所以，对于二维变量的线性函数集，其VC维等于3，因为3是能完全分开的最大样本数量



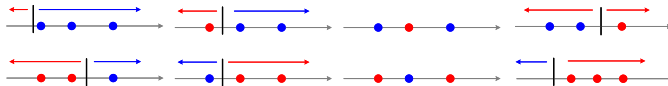
假设样本为一维的（即为一维直线上的点），那么它的分离超平面也只能是一个点。当 $l = 1$ 时，它可以被完全分开



有两个样本点时，也可以被一个分隔点完全分开



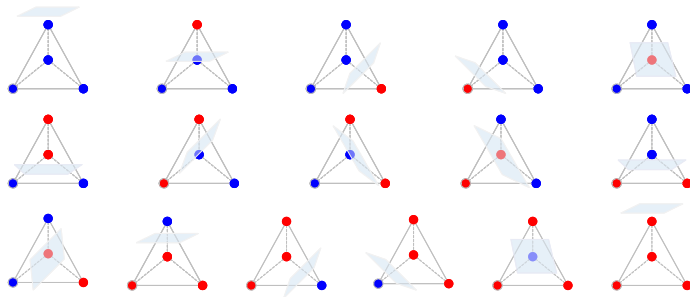
但是有三个样本点时，就不能被一个分隔点完全分开



所以一维数据线性函数集的VC维等于2



对于三维数据，重点看一下四个样本的情况。三维数据的分离超平面是二维平面，它可以将四个样本的情况完全分开，因此三维数据线性函数集的VC维是4（5个数据不能完全分开）¹⁶。

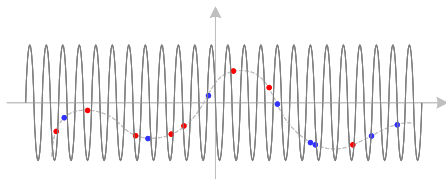
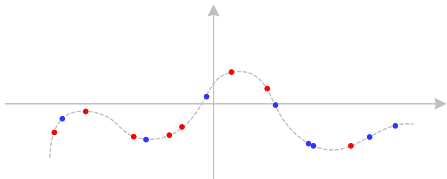


¹⁶线性函数VC维等于数据维度数加1，但这一结论对非线性函数一般不成立



假设用极坐标表示样本，即用一条线上的位置表示样本，那么可以看到，只要精心挑选系数 α ，实函数 $\sin(\alpha z)$ 就能将任意多样本完全分开¹⁷

$$f(z, \alpha) = \theta(\sin(\alpha z)), \quad \alpha \in \mathbb{R}^1$$



¹⁷ $\alpha = \pi \{ \sum_{i=1}^l (1 - \delta_i) 10^i + 1 \}$, δ_i 为 z_i 所属的类别，取值0或1



VC维是非常重要的概念，并且人们在1968年发现了VC维和生长函数之间的重要联系

已知生长函数 $G^{\Lambda}(I) = \ln \sup_{z_1, \dots, z_I} N^{\Lambda}(z_1, z_2, \dots, z_I)$ ，而 $N^{\Lambda}(z_1, z_2, \dots, z_I)$ 表示该指示函数集中的函数能把给定的所有样本分成多少种不同的结果，所以 N^{Λ} 的最大取值就是 2^I （将所有情况分开），所以 $G^{\Lambda}(I)$ 取值的上限是

$$G^{\Lambda}(I) = I \ln 2$$

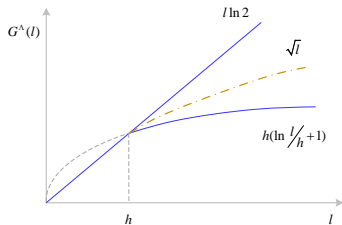
另外，人们还发现生长函数总是受下面不等式的约束（ h 为一个特殊的正整数¹⁸）

$$G^{\Lambda}(I) \leq h \left(\ln \frac{I}{h} + 1 \right)$$

¹⁸ 后面会看到就是函数集的VC维



如下图所示，生长函数要么是线性的 ($l \ln 2$)，要么是以对数函数 ($h(\ln \frac{l}{h} + 1)$) 为界的。例如，不可能有 $G^\Lambda(l) = c\sqrt{l}$ 。



- 若生长函数是线性的，则函数集VC维无穷大
- 若生长函数以参数为 h 的对数函数为界，则函数集VC维是有限的且等于 h
- 生长函数在 l 取值较小时可能是线性的，在 l 取值较大后可能受对数函数界约束

19 20

¹⁹ 生长函数的取值跟 l 的值有关，但VC维跟具体的 l 值无关，无论 l 多大总是能把样本打散，这意味着VC维的取值没有上限即无穷大

²⁰ 例如对于二维数据的线性函数集，当 $l = \{1, 2, 3\}$ 时，生长函数的取值分别为 $\ln 2, \ln 2^2, \ln 2^3$



因为有下列不等式

$$\frac{H^{\Lambda}(l)}{l} \leq \frac{H_{ann}^{\Lambda}(l)}{l} \leq \frac{G^{\Lambda}(l)}{l} \leq \frac{h(\ln \frac{l}{h} + 1)}{l}, \quad (l > h)$$

所以指示函数集的VC维有限是ERM（经验风险最小化）原则一致性的充分条件，并且很明显，VC维有限意味着收敛速度快²¹

VC维有限也是ERM学习机具有与分布无关（与 $F(\mathbf{z})$ 无关）的一致性的充分必要条件²²

²¹因为这意味着前面收敛速度的界是有限的

²²三个里程碑里的第三个， $G^{\Lambda}(l)$ 是与分布无关的



已知前面给出了完全有界函数集 $R(\alpha)$ 和 $R_{emp}(\alpha)$ 之间的误差的界 (Jump to)，只不过在那里用的是非构造性的 $G^{\mathbf{A}, \mathbf{B}}$ ，现在可利用VC维给出构造性的界

- 若函数集的VC维有限，令

$$\mathcal{E} = 4 \frac{h(\ln \frac{2l}{h} + l) - \ln(\frac{\eta}{4})}{l}$$

- 若函数集中的函数数量是有限的为 N ，令

$$\mathcal{E} = 2 \frac{\ln N - \ln \eta}{l}$$

将此处的 \mathcal{E} 代入前面界的表达式即可



函数空间的VC维和函数自由参数数量、函数复杂度之间有什么关系？

- 一般情况下，模型自由参数数量增加模型复杂度也会增加，模型空间的VC维也会增加
- 模型空间的VC维可能小于、等于、大于自由参数的数量
- 本质上，影响模型泛化能力的是模型的VC维而不是参数数量

VC维给我们克服“维度灾难”创造了一个很好的机会：用一个包含很多参数但却有较小VC维的函数集实现较好的泛化性。