淡蓝小点技术系列：学习理论精炼介绍

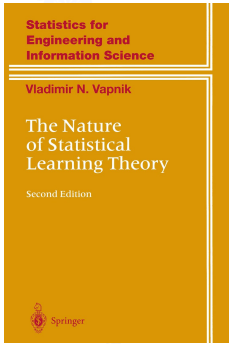
淡蓝小点Bluedotdot

微信：bluedotdot_cn

2024 年 5 月 12 日

前言背景

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



前言背景

- 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。

前言背景

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。

前言背景

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。

前言背景

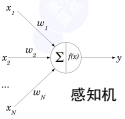
学习理论的主要内容包括：

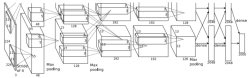
1. 统计推断一致性的充分必要条件及其相关概念
2. 学习机学习过程中收敛速度的界
3. 小样本下对泛化能力的控制（针对小样本的归纳推理原则）
4. 对各类问题的归纳推理（如模式识别、函数估计等）

名词解释

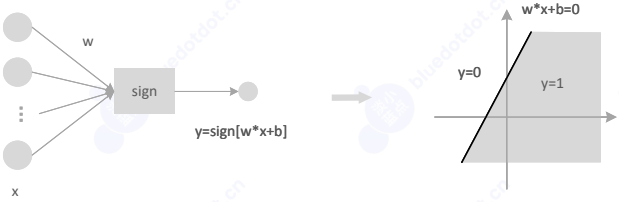
- 学习理论：learning theory
- 依赖关系：dependency
- 函数集：set of functions
- 泛化能力：generalization
- 收敛：converge
- 一致性：consistency
- 归纳推理：inductive inference
- 原则：principle
- 非平凡：non-trivial
- 渐近性：asymptotic

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



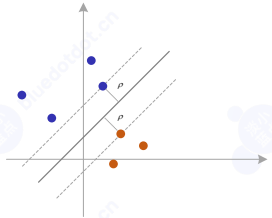
Perceptron 感知机

- 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：$\sup_{w} \min_{i} y_i(z_i \cdot w) > \rho$
3. 感知机经过充分的训练



那么最多经过$N \leq [\frac{R^2}{\rho^2}]$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$k(k=1,2,\cdots)$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

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

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 erros 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.）。

前言背景

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$A : \mathcal{X} \to \mathcal{Y}$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$\delta$是一个任意小量。如$f$和$f_\delta$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

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

即使困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程序的最短长度来衡量。

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

前言背景

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。

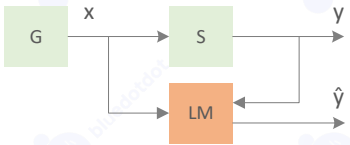
目录

学习问题表示

学习问题表示

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$f(\mathbf{x}, \boldsymbol{\alpha}), \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$，其中$\boldsymbol{\Lambda}$是函数参数集



学习问题就是从函数集$f(\mathbf{x}, \boldsymbol{\alpha}), \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$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

学习问题表示

假设训练数据是由独立同分布的样本构成

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

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$f(\mathbf{x}, \boldsymbol{\alpha})$之间的损失或差异（loss or discrepancy）。理论上损失的期望为

$$R(\boldsymbol{\alpha}) = \int L(\mathbf{y}, f(\mathbf{x}, \boldsymbol{\alpha}))\mathrm{d}F(\mathbf{x}, \mathbf{y}), \quad \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$$

因为$R(\boldsymbol{\alpha})$的取值随函数$f(\mathbf{x}, \boldsymbol{\alpha})$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

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

学习问题表示

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$f(\mathbf{x}, \boldsymbol{\alpha}), \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$就是指示函数集，损失函数可写为

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

- 对回归问题，损失函数可写为

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

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

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

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$Q(\mathbf{z},\boldsymbol{\alpha})$（$Q$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

$$R(\boldsymbol{\alpha})=\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z}),\quad \boldsymbol{\alpha}\in\boldsymbol{\Lambda}$$

理认上，找到使$R(\boldsymbol{\alpha})$取值最小的$\boldsymbol{\alpha}$，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$R(\boldsymbol{\alpha})$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$R_{emp}(\boldsymbol{\alpha})$表示

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

通过最小化R_{emp}找到最优$\boldsymbol{\alpha}$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。

学习问题表示

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

学习问题表示

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成更为宽泛的函数集
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

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

假设训练数据为$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_l\}$，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

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

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$f(\mathbf{x},\boldsymbol{\alpha}=\boldsymbol{\alpha}_0)=f_0(\mathbf{x})$，那么有

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

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$f_0(\mathbf{x})$。

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

所以

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

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

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

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$p_0(\mathbf{t})$

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

令

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

注意这一项与$p(\mathbf{t}, \boldsymbol{\alpha})$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$R^*(\boldsymbol{\alpha})$

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

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

学习过程一致性

学习过程一致性

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$R_{emp}(\boldsymbol{\alpha})$找到的$\boldsymbol{\alpha}$，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$R_{emp}(\boldsymbol{\alpha})$趋向最优$R(\boldsymbol{\alpha}^*)$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$\boldsymbol{\alpha}$使$R_{emp}(\boldsymbol{\alpha})$取值最小（或极小），不能确保该$\boldsymbol{\alpha}$同样使$R(\boldsymbol{\alpha})$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求密度函数（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

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

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

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

假设$Q(\mathbf{z}, \boldsymbol{\alpha}_l)$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$Q(\mathbf{z}, \boldsymbol{\alpha}), \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$函数集和分布函数$F(\mathbf{z})$是一致的

$$R(\boldsymbol{\alpha}_l) \xrightarrow[l \to \infty]{P} \inf_{\boldsymbol{\alpha} \in \boldsymbol{\Lambda}} R(\boldsymbol{\alpha})$$

$$R_{emp}(\boldsymbol{\alpha}_l) \xrightarrow[l \to \infty]{P} \inf_{\boldsymbol{\alpha} \in \boldsymbol{\Lambda}} R(\boldsymbol{\alpha})$$

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$R(\boldsymbol{\alpha})$，所以也需要第一个式子。但这两个式子可以改写成一个式子。

学习过程一致性

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

$$A \leq \int Q(\mathbf{z}, \boldsymbol{\alpha}) \mathrm{d}F(\mathbf{z}) \leq B, \quad (A \leq R(\boldsymbol{\alpha}) \leq B)$$

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

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

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

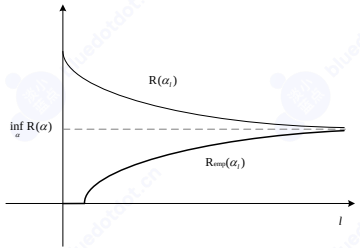
[8] 因为式子中有取上确界的操作

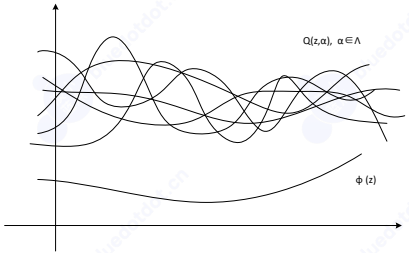
学习过程一致性

平凡一致性（trivial consistency）：假设对于函数集$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$，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

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

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。





学习过程一致性

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

$$\mathbf{\Lambda}(c) = \{\boldsymbol{\alpha} : \int Q(\mathbf{z}, \boldsymbol{\alpha}) \mathrm{d}F(\mathbf{z}) > c, \quad \boldsymbol{\alpha} \in \mathbf{\Lambda}\}$$

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

$$\inf_{\boldsymbol{\alpha} \in \mathbf{\Lambda}(c)} R_{emp}(\boldsymbol{\alpha}) \xrightarrow[l \to \infty]{P} \inf_{\boldsymbol{\alpha} \in \mathbf{\Lambda}(c)} R(\boldsymbol{\alpha})$$

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

学习过程一致性

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$p(\mathbf{x}, \boldsymbol{\alpha}_0)$，给定的密度函数集$p(\mathbf{x}, \boldsymbol{\alpha}) \in \boldsymbol{\Lambda}$ 具有下列依概率收敛的性质

$$\inf_{\boldsymbol{\alpha} \in \boldsymbol{\Lambda}} \frac{1}{l} \sum_{i=1}^{l} (-\log p(\mathbf{x}_i, \boldsymbol{\alpha})) \xrightarrow[l \to \infty]{P} \inf_{\boldsymbol{\alpha} \in \boldsymbol{\Lambda}} \int (-\log p(\mathbf{x}, \boldsymbol{\alpha})) p(\mathbf{x}, \boldsymbol{\alpha}_0) \mathrm{d}\mathbf{x}$$

极大似然方法的关键定理：若密度函数集满足$0 < a \leq p(\mathbf{x}, \boldsymbol{\alpha}) \leq A < \infty, \quad \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$[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$p(\mathbf{x}, \boldsymbol{\alpha}_0)$一致单边收敛

$$\lim_{l \to \infty} P\{ \sup_{\boldsymbol{\alpha} \in \boldsymbol{\Lambda}} (\int \log p(\mathbf{x}, \boldsymbol{\alpha}) p(\mathbf{x}, \boldsymbol{\alpha}_0) \mathrm{d}\mathbf{x} - R_{emp}(\boldsymbol{\alpha})) > \epsilon \} = 0$$

[10] 完全有界

学习过程一致性

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$\boldsymbol{\xi}^l$），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$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$的双边经验过程（two-sided empirical process）

$$\boldsymbol{\xi}^l=\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}}|\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})|,\quad l=1,2,\cdots$$

要讨论的是双边经验过程在什么条件下依概率收敛到0

$$\lim_{l\to\infty}P\{\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}}|\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})|>\epsilon\}=0,\quad\forall\epsilon>0$$

另一个随机过程是单边经验过程（one-sided empirical process），它由随机变量$\boldsymbol{\xi}_+^l$的序列构成

$$\boldsymbol{\xi}_+^l = \sup_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}} \left(\int Q(\mathbf{z},\boldsymbol{\alpha}) \mathrm{d}F(\mathbf{z}) - \frac{1}{l}\sum_{i=1}^{l} Q(\mathbf{z}_i,\boldsymbol{\alpha}) \right)_+, \quad l = 1, 2, \cdots$$

其中

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

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

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

学习过程一致性

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

$$\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}}\left(\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})_+\right)>\epsilon\overset{?}{=}\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}}\left(\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})\right)>\epsilon$$

原因在于：$\boldsymbol{\xi}_+^l$的定义中有取上确界（sup）的操作。设$\Delta R^l=intQ(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})$

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

学习过程一致性

若函数集$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\Lambda$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$\boldsymbol{\xi}^l$变为

$$\boldsymbol{\xi}^l=|\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})|$$

根据辛钦大数定律

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

此时双边经验过程收敛性一定满足

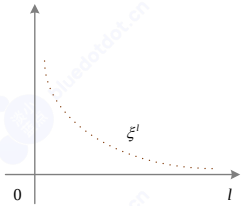
$$\lim_{l\to\infty}P\{\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}}|\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})|>\epsilon\}=0\Rightarrow\lim_{l\to\infty}P\{|\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})|>\epsilon\}=0$$

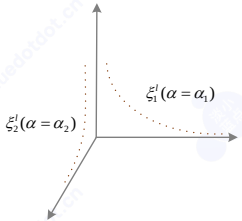
随着观察数量l的增加，随机变量序列$\boldsymbol{\xi}^l$收敛于0

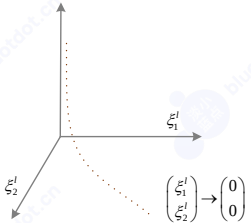
学习过程一致性

若函数集$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\Lambda$中有有限个元素（如$N$个），那么双边经验过程的随机变量$\boldsymbol{\xi}^l$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于应于一个维度，向量空间中的大数定律说明所有维度都会依概率收敛。







学习过程一致性

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$\boldsymbol{\xi}^l$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$\boldsymbol{\xi}^l$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。

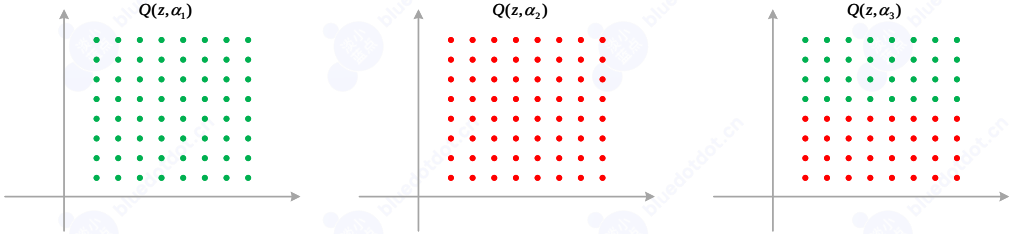
在传统的统计学中，只考虑了双边一致收敛（大数定律），但是并没有考虑单边一致收敛。单边一致收敛此处重要是因为根据关键定理它是ERM归纳原则一致性的充要条件。

学习过程一致性

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$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$的熵。

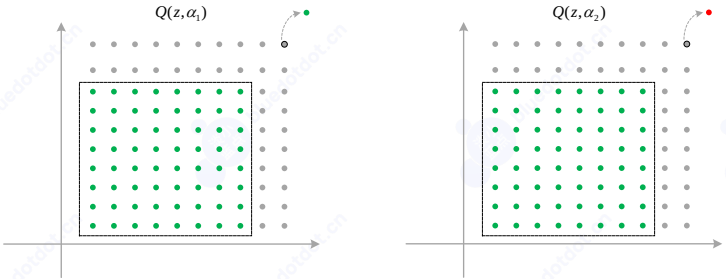
以二分类问题为例来讨论（即任意$\mathbf{z}_i$只能标记为1或者0），设$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$是一个指示函数集，有样本$\{\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_l\}$，用$N^{\boldsymbol{\Lambda}}(\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_l)$表示该指示函数集中的函数能把给定的所有样本分成多少种不同的结果。我们用这个量来表征该函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（$\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\boldsymbol{\alpha}_3$），它们对样本的分类结果如下所示，那么此时$N^{\boldsymbol{\Lambda}}(\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_l)=3$



学习过程一致性

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$N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_l)$的取值就是多少了？当然不是。因为$\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_l\}$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。

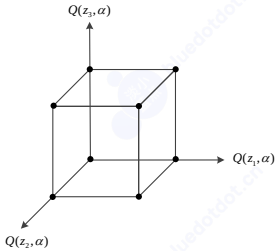


学习过程一致性

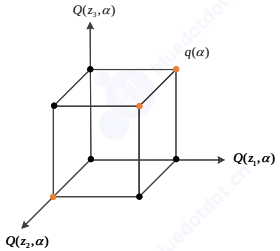
设$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$是函数$\boldsymbol{\alpha}$在泛化数据上的判定结果（对于判定问题只能是0或1），设

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

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$N^{\boldsymbol{\Lambda}}(\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_l)$就是顶点集合的子集。例如，设只有三个样本$\{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3\}$



理想情况下所有可能的判定结果



函数集在样本集上的判定结果

学习过程一致性

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

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

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

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

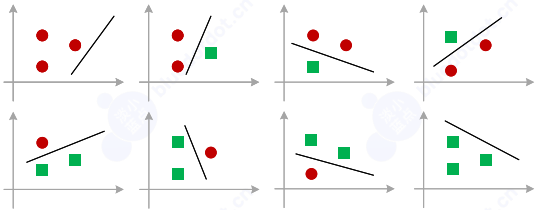
$H^{\boldsymbol{\Lambda}}(l)$称为指示函数集$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$在数量为$l$的样本集上的熵，它反映的是指示函数集在$l$样本集上的多样性。或者说，它反应的是对于$l$个样本，函数空间能对这$l$个样本的多少种取值情况进行区分。这是对函数空间表现能力的一种度量。

注意，$H^{\boldsymbol{\Lambda}}(l)$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）。

学习过程一致性

假设$l = 3$并且这三个样本不在一条直线上，设函数集$\mathbf{\Lambda}$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

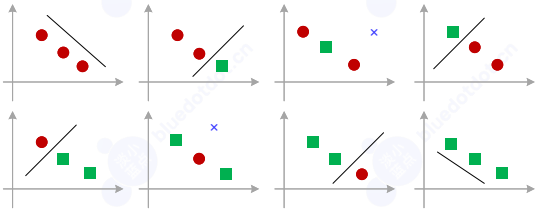
$$N^{\mathbf{\Lambda}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = 8, \quad H^{\mathbf{\Lambda}}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = \log_2(\mathbf{N}) = 3$$



学习过程一致性

若$l=3$并且这三个样本恰好在一条直线上，设函数集$\mathbf{\Lambda}$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

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

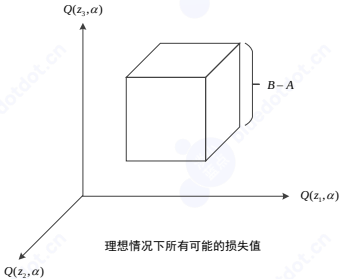


因为只有这两种情况，假设出现这两种情况的概率各为0.5，所以

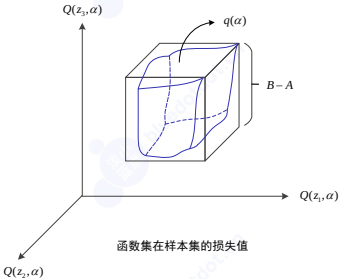
$$H^{\mathbf{\Lambda}}(l)=\mathrm{E}\ln N^{\mathbf{\Lambda}}(\mathbf{z}_1,\cdots,\mathbf{z}_l)=0.5*3+0.5*2.58=2.79$$

学习过程一致性

将指示函数拓展到一般实函数。设$A \leq Q(\mathbf{z}, \boldsymbol{\alpha}) \leq B, \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$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l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。



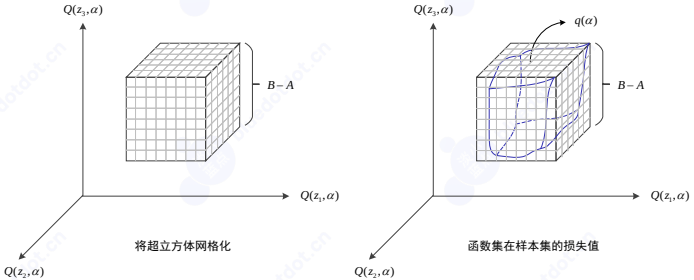
理想情况下所有可能的损失值



函数集在样本集的损失值

学习过程一致性

将超立方体离散化为有限最小$\epsilon-$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$N = N^{\mathbf{\Lambda}}(\epsilon; \mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_l)$ 是向量集$q(\boldsymbol{\alpha}), \boldsymbol{\alpha} \in \mathbf{\Lambda}$的最小$\epsilon-$网格的元素数目。



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

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

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

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

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

学习过程一致性

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$\epsilon-$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

$$
\begin{aligned}
N^{\mathbf{\Lambda}}(\epsilon; \mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_l) &= N^{\mathbf{\Lambda}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_l) \\
H^{\mathbf{\Lambda}}(\epsilon; \mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_l) &= H^{\mathbf{\Lambda}}(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_l) \\
H^{\mathbf{\Lambda}}(\epsilon; l) &= H^{\mathbf{\Lambda}}(l)
\end{aligned}
$$

这表明，对有界实函数集的一般结论也适用于指示函数集。

学习过程一致性

一致双边收敛的定义

$$\lim_{l\to\infty} P\{\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}} |\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z}) - \frac{1}{l}\sum_{i=1}^{l} Q(\mathbf{z}_i,\boldsymbol{\alpha})| > \epsilon\} = 0, \quad \forall\epsilon > 0$$

对于有界实函数集一致双边收敛的充分必要条件

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

对于指示函数集一致双边收敛的充分必要条件[11]

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

（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

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

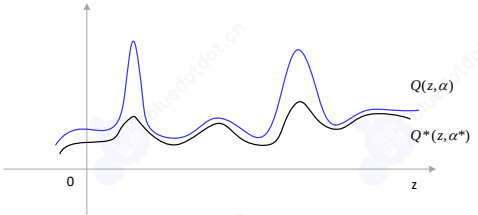
一致双边收敛定义去绝对值符号后为

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

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）。

学习过程一致性

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



$$Q(\mathbf{z},\boldsymbol{\alpha})-Q^*(\mathbf{z},\boldsymbol{\alpha}^*)\geq 0,\quad \forall\mathbf{z}$$

$$\int(Q(\mathbf{z},\boldsymbol{\alpha})-Q^*(\mathbf{z},\boldsymbol{\alpha}^*))\mathrm{d}F(\mathbf{z})\leq\delta,\quad \delta>0$$

可简单理解为：Q^*是Q的下界，并且Q和Q^*是非常接近的

学习过程一致性

完全有界函数集$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$，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$Q^*(\mathbf{z},\boldsymbol{\alpha}^*),\boldsymbol{\alpha}^*\in\boldsymbol{\Lambda}^*$，使得$Q$和$Q^*$满足前面的关系式，并且$Q^*$在$l$个样本上其$\epsilon-$熵（VC熵）满足下面不等式

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

已知根据关键定里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）

学习过程一致性

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

- 退火的VC熵：$H^{\mathbf{\Lambda}}_{ann}(l)=\ln \mathrm{E}[N^{\mathbf{\Lambda}}(\mathbf{z}_1,\cdots,\mathbf{z}_l)]$
- 生长函数：$G^{\mathbf{\Lambda}}(l)=\ln \sup_{\mathbf{z}_1,\cdots,\mathbf{z}_l} N^{\mathbf{\Lambda}}(\mathbf{z}_1,\cdots,\mathbf{z}_l)$（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l总有

$$H^{\mathbf{\Lambda}}(l)\le H^{\mathbf{\Lambda}}_{ann}(l)\le G^{\mathbf{\Lambda}}(l)$$

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

学习过程一致性

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

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

- 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

$$\lim_{l\to\infty}\frac{H^{\Lambda}_{ann}(l)}{l}=0$$

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

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

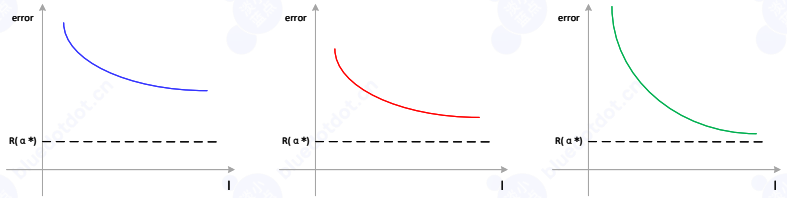
- ERM原则是一致的且有快的收敛速度，并且这种收敛不依赖于任何特定的概率分布的充要条件：

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

学习过程收敛速度的界

学习过程收敛速度的界

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
- 基于生成长函数的与分布无关的界

[12]不可构造是指不可实际计算

学习过程收敛速度的界

对于指示函数（判定问题），设$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$是一个指示函数集，$H^{\boldsymbol{\Lambda}}(l)$是对应的VC熵，$H^{\boldsymbol{\Lambda}}_{ann}(l)$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

$$P\{\sup_{\boldsymbol{\alpha}\in\Lambda}|\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})|>\epsilon\}\leq 4\exp\{(\frac{H^{\boldsymbol{\Lambda}}_{ann}(2l)}{l}-\epsilon^2)l\}$$

$$P\{\sup_{\boldsymbol{\alpha}\in\Lambda}\frac{\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\mathbf{z}_i,\boldsymbol{\alpha})}{\sqrt{\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})}}>\epsilon\}\leq 4\exp\{(\frac{H^{\boldsymbol{\Lambda}}_{ann}(2l)}{l}-\frac{\epsilon^2}{4})l\}$$

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

$$\lim_{l\to\infty}\frac{H^{\boldsymbol{\Lambda}}_{ann}(l)}{l}=0$$

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。

学习过程收敛速度的界

指示函数集，$R(\boldsymbol{\alpha}) = \int Q(\mathbf{z}, \boldsymbol{\alpha}) \mathrm{d}F(\mathbf{z})$表示的是$\{\mathbf{z} : Q(\mathbf{z}, \boldsymbol{\alpha}) = 1\}$的概率，而$R_{emp}(\boldsymbol{\alpha})$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

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

于是便有人提出了相对一致收敛

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

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$R(\boldsymbol{\alpha})$取值较小时

$$\sqrt{R(\boldsymbol{\alpha})} \to \sqrt{R(\boldsymbol{\alpha})(1 - R(\boldsymbol{\alpha}))}$$

所以可用$\sqrt{R(\boldsymbol{\alpha})}$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$H_{ann}^{\mathbf{\Lambda}}(l) \leq G^{bm\Lambda}(l)$，所以只需要用$G^{bm\Lambda}(l)$替换掉$H_{ann}^{\mathbf{\Lambda}}(l)$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）。

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

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

同样如果下式被满足，那么这两个界是非平凡的

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

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

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

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

学习过程收敛速度的界

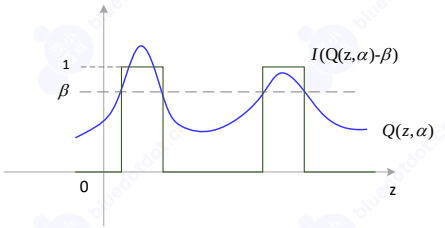
收敛速度上界可以推广到实函数集上。设$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$是一个实函数集并且

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

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$Q(\mathbf{z},\boldsymbol{\alpha})$构建指示函数

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

θ函数表示对于任意给定的$\boldsymbol{\alpha}^*,\beta^*$



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

学习过程收敛速度的界

设$H^{\mathbf{\Lambda},\mathcal{B}}(l)$表示对应的指示函数集的VC熵，相应的$H_{ann}^{\mathbf{\Lambda},\mathcal{B}}(l)$和$G^{\mathbf{\Lambda},\mathcal{B}}(l)$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$A \leq Q(\mathbf{z},\boldsymbol{\alpha}) \leq B, \boldsymbol{\alpha} \in \mathbf{\Lambda}$，其中$A, B$均是有限值，那么下面不等式成立

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

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

$$P\{\sup_{\boldsymbol{\alpha}\in\Lambda}\frac{\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z}) - \frac{1}{l}\sum_{i=1}^{l} Q(\mathbf{z}_i,\boldsymbol{\alpha})}{\sqrt{\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})}} > \epsilon\} \leq 4\exp\{(\frac{H_{ann}^{\mathbf{\Lambda},\mathcal{B}}(2l)}{l} - \frac{\epsilon^2}{4B})l\}$$

（3）若函数集是非负（但不一定有界）的：设$0 \leq Q(\mathbf{z}, \boldsymbol{\alpha}), \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$，将$Q(\mathbf{z}, \boldsymbol{\alpha})$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

$$m_p(\boldsymbol{\alpha}) = \sqrt[p]{\int Q^p(\mathbf{z}, \boldsymbol{\alpha}) \mathrm{d}F(\mathbf{z})}$$

那么下面的界成立

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

其中

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

这里三个式子定义的界是非平凡的，如果满足

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

学习过程收敛速度的界

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：$-\infty < A \leq Q(\mathbf{z}, \boldsymbol{\alpha}) \leq B, \boldsymbol{\alpha} \in \boldsymbol{\Lambda} < +\infty$

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

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

$$P\{\sup_{\boldsymbol{\alpha}\in\Lambda} \frac{\int Q(\mathbf{z}, \boldsymbol{\alpha}) \mathrm{d}F(\mathbf{z}) - \frac{1}{l}\sum_{i=1}^{l} Q(\mathbf{z}_i, \boldsymbol{\alpha})}{\sqrt{\int Q(\mathbf{z}, \boldsymbol{\alpha}) \mathrm{d}F(\mathbf{z})}} > \epsilon\} \leq 4 \exp\{(\frac{G_{ann}^{\boldsymbol{\Lambda},\mathcal{B}}(2l)}{l} - \frac{\epsilon^2}{4B})l\}$$

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

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

同样满足前面的条件时，这三个界是非平凡的

学习过程收敛速度的界

再回答两个问题：

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

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

[14]第三种情况公式更复杂，此处节省时间和空间

（1）设$A \leq Q(\mathbf{z}, \boldsymbol{\alpha}) \leq B, \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$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$Q(\mathbf{z}, \boldsymbol{\alpha}), \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$中所有函数成立）

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

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

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

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

（2）设$0 \leq Q(\mathbf{z}, \boldsymbol{\alpha}) \leq B, \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$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）

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

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

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

学习过程收敛速度的界

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？

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

- 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？

学习过程收敛速度的界

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!

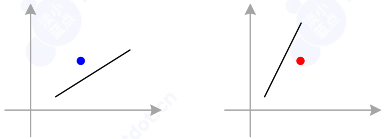
以指示函数集为例，$Q(\mathbf{z}, \boldsymbol{\alpha}), \boldsymbol{\alpha} \in \boldsymbol{\Lambda}$的VC维是指：能够被集合中的函数完全打散的样本集$\{\mathbf{z}_1, \cdots, \mathbf{z}_h\}$的最大样本数$h$（或者说$h$个样本能够被集合中的函数完全分类为$2^h$种情况的最大$h$值）[15]

VC维（h值）的大小跟两个因素有关：函数集、样本的维度数

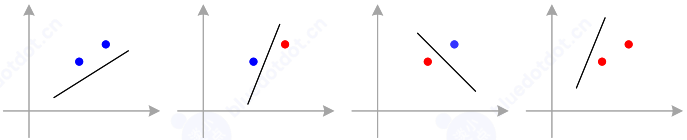
[15]注意：h个样本可以有多种不同的分布情况，VC维的定义中只要求有h个样本能被完全分类为2^h种情况即可，并不要求所有分布下的h个样本都能被分类成2^h种情况；另外，很明显VC维只能是非负整数

学习过程收敛速度的界

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

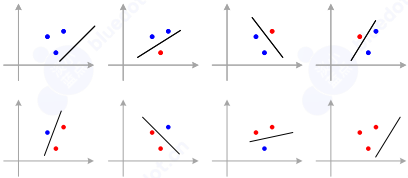


若样本集中有两个样本，同样可以用不同的直线将其分成2^2次方种情况

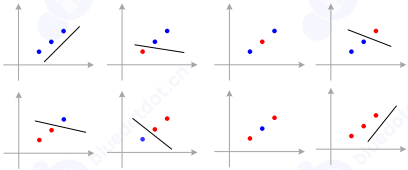


学习过程收敛速度的界

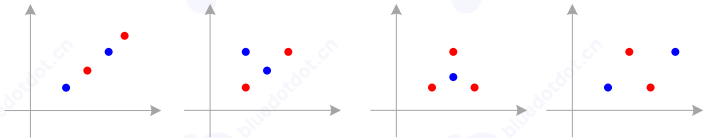
若样本集中有三个样本（$2^3 = 8$），则有些分布下能将样本完全打散，有些分布下则不能



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



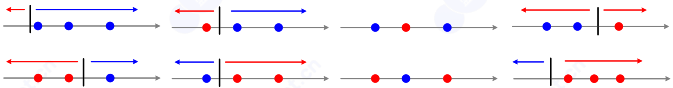
所以，对于二维变量的线性函数集，其VC维等于3，因为3是能完全分开的最大样本数量

学习过程收敛速度的界

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

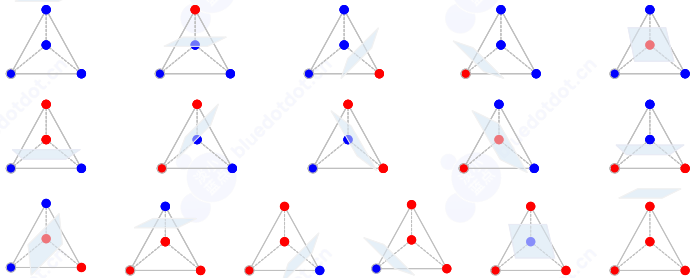
有两个样本点时，也可以被一个分隔点完全分开

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]。

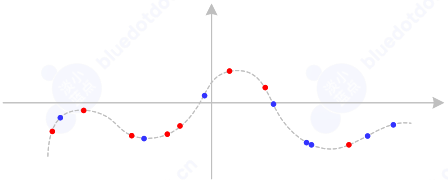


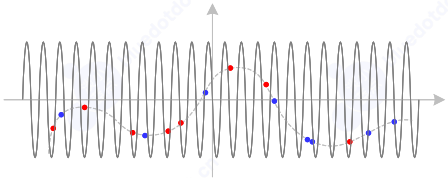
[16]线性函数VC维等于数据维度数加1，但这一结论对非线性函数一般不成立

学习过程收敛速度的界

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$\boldsymbol{\alpha}$，实函数$\sin(\boldsymbol{\alpha} z)$就能将任意多样本完全分开[17]

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





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

学习过程收敛速度的界

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$G^{\mathbf{\Lambda}}(l) = \ln \sup_{\mathbf{z}_1,\cdots,\mathbf{z}_l} N^{\mathbf{\Lambda}}(\mathbf{z}_1,\cdots,\mathbf{z}_l)$，而$N^{\mathbf{\Lambda}}(\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_l)$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，所以$N^{\mathbf{\Lambda}}$的最大取值就是2^l（将所有的情况分开），所以$G^{\mathbf{\Lambda}}(l)$取值的上限是

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

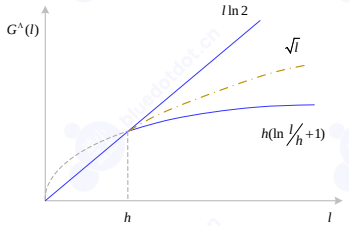
另外，人们还发现生长函数总是受下面不等式的约束（h为一个特殊的正整数[18]）

$$G^{\mathbf{\Lambda}}(l) \leq h(\ln \frac{l}{h} + 1)$$

[18]后面会看到就是函数集的VC维

学习过程收敛速度的界

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，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$\ln 2, \ln 2^2, \ln 2^3$

学习过程收敛速度的界

因为有下列不等式

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

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，$G^{\mathbf{\Lambda}}(l)$是与分布无关的

学习过程收敛速度的界

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$G^{\Lambda,\mathcal{B}}$，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

学习过程收敛速度的界

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。

控制学习过程的泛化能力

控制学习过程的泛化能力

控制学习器泛化（推广）能力是指：构造一种利用小样本训练集最小化风险泛函的归纳原则[23]

前面给出了完全有界非负函数集的学习器泛化（推广）能力的界[24]：

$$R(\alpha_l) \leq R_{emp}(\alpha_l) + \frac{B\mathcal{E}}{2}[1 + \sqrt{1 + \frac{4R_{emp}(\alpha_l)}{B\mathcal{E}}}]$$

若函数集$Q(z, \alpha)$是有限的，包含$\alpha_1, \alpha_2, \cdots, \alpha_N$共$N$个元素，则

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

若函数集$Q(z, \alpha), \alpha \in \Lambda$是无限的但VC维是有限的记为$h$，则

$$\mathcal{E} = 4\frac{h(\ln \frac{2l}{h} + 1) - \ln(\frac{\eta}{4})}{l}$$

[23]通俗解释：如何使得基于小样本训练集找到的模型α_l，它的泛化误差（也称实际风险）是较小的，也就是有界的；前面介绍过这些界以$1 - \eta$的概率成立

[24]这里只是以有界非负函数集为例；这里的B就是函数集的上界

控制学习过程的泛化能力

什么样的样本集是小样本集？
若样本集的样本数量为l，如果比值l/h（样本数与学习器VC维之比）较小，例如$l/h < 20$，我们就认为样本数是少的，即小样本集

l, h的取值与泛化能力之间的关系：

- 当l/h较大时（即l较大h较小），$\mathcal{E}$的取值相对较小，此时界的右边第二项取值较小，泛化误差和经验误差的差较小
- 当l/h较小时（即l较小h较大），$\mathcal{E}$的取值相对较大，此时界的右边第二项取值较大，泛化误差和经验误差的差较大

泛化误差的上界由两项共同决定[25]，第一项为经验误差（$R_{emp}(\alpha_l)$），第二项称为置信范围（其取值由l和h决定），好的归纳原则应同时控制这两项的取值，使它们的和达到最小[26]

[25] 上界也称为保证风险guaranteed risk
[26] 前面的ERM原则在小样本下是不够的

控制学习过程的泛化能力

SRM：structural risk minimization 结构风险最小化，也称为有序风险最小化

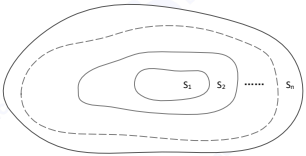
设函数$Q(\mathbf{z},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$的集合$S$具有一种特殊的结构——容许结构（admissible structure），它是由一系列嵌套的函数子集$S_k=\{Q(\mathbf{z},\alpha),\alpha\in\boldsymbol{\Lambda}_k\}$组成，它们满足

$$S_1\subset S_2\subset\cdots\subset S_n\cdots$$

结构中的元素满足下面两个性质：

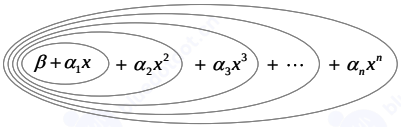
1. 每个函数集S_k的VC维h_k是有限的，因此有$h_1\leq h_2\leq\cdots\leq h_n\cdots$
2. 任何S_k或者是完全有界的函数集$0\leq Q(\mathbf{z},\boldsymbol{\alpha})\leq B_k,\boldsymbol{\alpha}\in\boldsymbol{\Lambda}_k$，或者对于某对$(p,\tau_k)$满足

$$\sup_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}_k}\frac{(\int Q^p(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z}))^{1/p}}{\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})}\leq\tau_k,\quad p>2$$

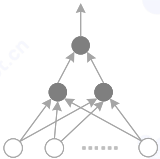


控制学习过程的泛化能力

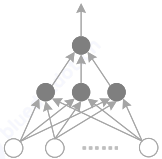
容许结构的两个例子：



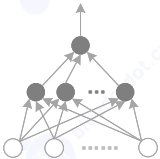
多项式函数集



2个隐元

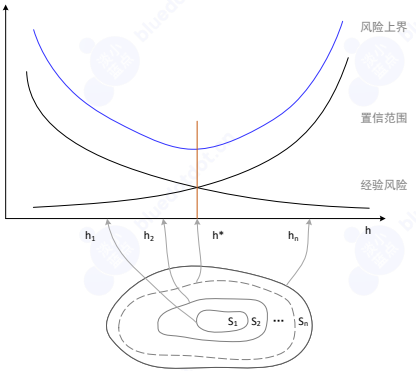


3个隐元



n个隐元

控制学习过程的泛化能力



SRM原则：在逼近精度和函数复杂度之间的折衷

- 当模型VC维增加时经验风险会变小，但置信范围会变大
- 在使保证风险最小的子集S_n中选择使经验风险最小的函数$Q(\mathbf{z}, \boldsymbol{\alpha}_l^n)$

很明显对于有l个样本的数据集，函数集S_n的序号n跟l，因此可记作$n = n(l)$

控制学习过程的泛化能力

收敛速度的渐近理论：SRM方法从$Q(\mathbf{z},\boldsymbol{\alpha}_l^{n(l)})$中得到的$\boldsymbol{\alpha}_l^{n(l)}$，其经验风险将收敛到风险泛函的最小值

$$R(\boldsymbol{\alpha}_0)=\inf_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}}\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})$$

并且收敛的渐近速度是

$$V(l)=r_{n(l)}+T_{n(l)}\sqrt{\frac{h_{n(l)}\ln l}{l}}$$

其中$r_{n(l)}$是函数逼近速度

$$r_n=\inf_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}_n}\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})-\inf_{\boldsymbol{\alpha}\in\boldsymbol{\Lambda}}\int Q(\mathbf{z},\boldsymbol{\alpha})\mathrm{d}F(\mathbf{z})$$

如果

$$\lim_{l\to\infty}\frac{T_{n(l)}^2 h_{n(l)}\ln l}{l}=0$$

控制学习过程的泛化能力

T_n根据容许结构中S_n有界或无界分两种情况：

- 若S_n中是完全有界函数$Q(\mathbf{z}, \alpha) \leq B_n$，则$T_n = B_n$
- 否则$T_n = \tau_n$

渐近理论告诉我们：可以根据极限式先验地找到$n(l)$

渐近速度：如果存在一个常数C，使得

$$V^{-1}(l)|\xi_l - \xi_0| \xrightarrow[l\to\infty]{P} C$$

那么就说随机变量$\xi_l, l = 1, 2, \cdots$以渐近速度$V(l)$收敛于ξ_0，很明显，$V(l)$（即r_n）取值越小收敛越快

控制学习过程的泛化能力

函数逼近速度是传统函数逼近理论（函数近似理论）研究的内容，人们研究了函数平滑性与其逼近速度之间的关系。假设某个函数存在的导数阶数为s，那么典型的逼近渐近速度有如下形式：

$$r_n = n^{-\frac{s}{N}}, \quad n = n(l)$$

其中N是输入变量所有空间维度数。这意味着，只有非常平滑（s很大）的函数才能在高维空间中有较高的渐近收敛速度

通常如果$r_n \leq n^{-1/2}$就认为收敛速度是高的，即收敛速度不超过$O(\frac{1}{\sqrt{n}})$时就认为是快的[27]

[27]很明显线性函数收敛是快的，因为它收敛速度为$O(1/\sqrt[N/2]{n})$

SRM是一种归纳原则，它的具体实现方法有很多种，常见的包括：

- L1正则化Lasso：

$$L(\mathbf{w}) = \mathrm{Loss}(\mathbf{w}) + \lambda \sum_{i=1}^{n} |w_i|$$

- L2正则化岭回归：

$$L(\mathbf{w}) = \mathrm{Loss}(\mathbf{w}) + \lambda \sum_{i=1}^{n} w_i^2$$

- BIC贝叶斯信息准则：

$$L(\mathbf{w}) = \frac{1}{2} M \ln N - \ln p(D|\mathbf{w}_{ML})$$[28]

- AIC信息准则：

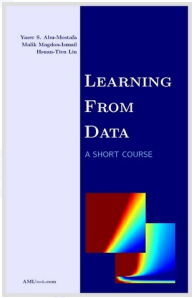
$$L(\mathbf{w}) = M - \ln p(D|\mathbf{w}_{ML})$$

[28] M是参数数量，N是训练数据量；BIC和AIC在不同资料中定义略有不同，此处定义来自PRML

概率近似正确PAC

概率近似正确PAC

PAC部分主要参考了以下文献



概率近似正确PAC

学习理论中另一重要部分是PAC：概率近以正确（Probably Approximately Correct）[29]

设学习问题是从X到Y的映射，我们称任何从X到Y的映射为概念（concept），用c表示；若c能准确的对任意样本(x, y)做映射即$c(x) = y$，则称c为“目标概念”（target concept），所以目标概念构成的集合称为“概念类”（concept classs），用C表示[30]

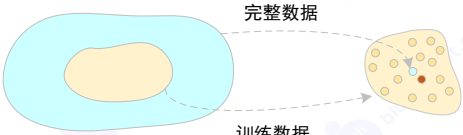
用H表示假设空间（hypothesis space，即所有候选模型的集合），所以目标概念c有可能存在于H中也有可能不存在于H中。若$c \in H$则称该问题是可分的（separable），若H中不包含任何目标概念c，则称该问题是不可分的（non-seperable）

[29]本部分内容主要参考西瓜书《机器学习》

[30]通常一个映射规则可以由多个不同的函数来表示，例如一个神经网络，只要交换每一层中神经元参数的位置，就能得到一个新的网络但其映射规则保持不变

概率近似正确PAC

一般情况下，由于训练集是有限的并且我们并不知道真实的概念类C是什么样的，因此想要从H中恰好找到一个c是十分困难的。通常，我们的期望是在训练集的作用下，能够利用某种学习算法L从H中找到一个h，使得h尽量和c相似



以较大概率学习出与目标概念误差在可接受范围内的模型，就是所谓概率近似正确

PAC讨论的是：从假设空间中找到一个较好的h的难度；或者说从假设空间中找到的h，它的泛化误差的大小是多少；或者说至少需要多少个样本才能找到一个可接受的h[31]

[31] PAC是在前面学习理论基础之上发展出来的，因而它讨论的问题比一般学习问题更具体一些

PAC辨识（PAC Identify）：对于任意ε, δ[32]，$0 < \varepsilon, \delta < 1$，若$c \in H$，对于任意$c \in C$和所有的数据分布$D$ [33]，若学习算法L总能从H中找到h，并满足[34]

$$P(E(h) \leq \varepsilon) \leq 1 - \delta$$

则称学习算法L能从假设空间H中PAC辨识概念类C

[32]这两个量是提前设定好的，它们的取值可以是相互独立的；ε定义了精度，δ定义了置信度

[33]这里的L是指某种归纳原则下的具体实现，例如ERM下的最小平方误差，SRM下的岭回归等；注意这里是数据分布即泛化数据，而非训练集

[34]这里的$E(h)$表示找到的h在泛化数据上预测结果的误差

概率近似正确PAC

PAC可学习（PAC Learnable）：对于任意$0 < \varepsilon, \delta < 1$，若对于任意数据分布只需要给出$m$个独立同分布的数据，存在学习算法$L$能从假设空间$H$中PAC辨识概念类$C$，就称概念类$C$对假设空间$H$是PAC可学习的，其中

$$m \geq \mathrm{poly}(1/\varepsilon, 1/\delta, \mathrm{size}(x), \mathrm{size}(c))$$[35]

这里$\mathrm{size}(x)$表示数据复杂度，$\mathrm{size}(c)$表示目标概念复杂度[36]

- 数据复杂度通常跟数据的维度数、数据分布复杂性、特征间依赖关系、不同类别数据量的不均衡性等有关
- 目标概念复杂度反映的是目标概念本身的固有特性，它决定了学习出目标概念所需的数据量、计算时间、计算量等，它有多种不同的量化方式，例如VC维

[35] 表示所需训练数据量m是$1/\varepsilon, 1/\delta, \mathrm{size}(x), \mathrm{size}(c)$的多项式函数

[36] 注意，关于m的定义中并没有显示涉及到H的复杂度，这是因为我们要找到的是H中的c或者跟c非常接近的h，所以h本身复杂度对m的影响由$\mathrm{size}(c)$来表现

概率近似正确PAC

若学习算法$\mathcal{L}$使概念类C为PAC可学习的并且$\mathcal{L}$的运行时间也是多项式函数$\text{poly}(1/\varepsilon, 1/\delta, \text{size}(x), \text{size}(c))$的，则称概念类$C$是高效PAC可学习的（efficiently PAC learnable），称$\mathcal{L}$为概念类C的PAC学习算法

满足PAC学习算法$\mathcal{L}$所需的最小样本数m称为学习算法$\mathcal{L}$的样本复杂度[37]

[37]假定学习算法处理每个样本的时间为常数的，则$\mathcal{L}$的时间复杂度由样本数量决定，也就是由样本复杂度决定

若假设空间的函数是有限的并且$c \in H$，该如何确定样本复杂度了？

假设H中的h其在泛化数据上的误差至少是ε，即对于任意的数据对(x, y)有$P(h(x) \neq y) = E(h) \geq \varepsilon$，那么

$$\begin{aligned} P(h(x) = y) &= 1 - P(h(x) \neq y) \\ &= 1 - E(h) \\ &< 1 - \varepsilon \end{aligned}$$

对于有m个样本的训练集$\{(x_i, y_y)\}^m$，其预测完全正确的概率为

$$\begin{aligned} P(\{(x_i, y_i)\}^m) &= (1 - P(h(x) \neq y))^m \\ &< (1 - \varepsilon)^m \end{aligned}$$

若假设空间中共有$|H|$个函数并且极限情况下H中只有一个h能在$\{(x_i, y_i)\}^m$上都完全正确，那么恰巧找到h并且预测正确的概率为

$$\frac{1}{|H|} P(\{(x_i, y_i)\}^m) < (1-\varepsilon)^m$$

$$\begin{aligned} P(h \in H : \{(x_i, y_i)\}^m) &< |H|(1-\varepsilon)^m \\ &< |H| e^{-m\varepsilon} \end{aligned}$$

注意这里用到了$(1-\varepsilon)^m \approx e^{-m\varepsilon}$

已知根据泰勒展开式有

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

令$x = -m\varepsilon$则有

$$e^{-m\varepsilon} = 1 - m\varepsilon + \frac{(m\epsilon)^2}{2!} - \frac{(m\varepsilon)^3}{3!} + \cdots$$

对于$(1-\varepsilon)^m$可利用二项式定理展开

$$\begin{aligned}(1-\varepsilon)^m &= \sum_{k=0}^{m} \binom{m}{k}(-\varepsilon)^k \\ &= \binom{m}{0}(-\varepsilon)^0 + \binom{m}{1}(-\varepsilon)^1 + \binom{m}{2}(-\varepsilon)^2 + \cdots \\ &= 1 - m\varepsilon + \frac{m(m-1)}{2}\varepsilon^2 + \cdots\end{aligned}$$

所以，当ε取值较小时可认为有

$$e^{-m\varepsilon} \approx 1 - m\varepsilon \approx (1-\varepsilon)^m$$

概率近似正确PAC

根据前面PAC可辨识的定义，我们要求$P(E(h) \leq \varepsilon) \geq 1-\delta$，因此

$$
\begin{aligned}
|H|e^{-m\varepsilon} &\leq \delta \\
m &\geq \frac{1}{\varepsilon}(\ln|H| + \ln\frac{1}{\delta})
\end{aligned}
$$

在有些资料中[38]，推导时会用$|C|$替代$|H|$。即假设一共有$|C|$个目标概念，那么应该是

$$
\begin{aligned}
\frac{1}{|C|}P(\{(x_i, y_i)\}^m) &< (1-\varepsilon)^m \\
P(h \in H : \{(x_i, y_i)\}^m) &< |C|(1-\varepsilon)^m \\
&< |C|e^{-m\varepsilon}
\end{aligned}
$$

进而推导出

$$
m \geq \frac{1}{\varepsilon}(\ln|C| + \ln\frac{1}{\delta})
$$

[38] https://pages.cs.wisc.edu/ shuchi/courses/787-F07/scribe-notes/lecture25.pdf

对于H无限、但$c \notin H$的情况，该如何确定样本复杂度了？这种情况下首先介绍霍夫丁不等式（Hoeffding inequality）

若训练集包含m个独立同分布样本，并有$0 < \varepsilon, \delta < 1$，则对任意$h \in H$下列不等式置信度为$\delta$ [39]

$$P(\hat{E}(h) - E(h) \geq \varepsilon) \leq \exp(-2m\varepsilon^2)$$
$$P(E(h) - \hat{E}(h) \geq \varepsilon) \leq \exp(-2m\varepsilon^2)$$
$$P(|E(h) - \hat{E}(h)| \geq \varepsilon) \leq 2\exp(-2m\varepsilon^2)$$

以及基于霍夫丁不等式的推论[40]

$$\hat{E}(h) - \sqrt{\frac{\ln(2/\delta)}{2m}} \leq E(h) \leq \hat{E}(h) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

[39] $P(\cdot \leq \varepsilon)$形式置信度为$1-\delta$；$E(h)$表示$h$在泛化数据上的误差，$\hat{E}(h)$表示$h$在训练数据上的误差

[40] 我们不在此证明霍夫丁不等式，它涉及到概率论中的矩母函数，留待在《面向机器学习（深度学习、人工智能）的数学基础》中再证明

因为霍夫丁不等式本身置信度为δ，根据第三个不等式有

$$\delta \geq 2\exp(-2m\varepsilon^2)$$

$$\ln\frac{\delta}{2} \geq -2m\varepsilon^2$$

$$\frac{1}{2\varepsilon^2}\ln\frac{2}{\delta} \leq m$$

所以有

$$m \geq \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}$$

当我们给定精度为ε，置信度为δ时，我们至少需要m个样本

类似地，如果给定了置信区间δ和样本数m，那么误差精度最大为

$$\begin{aligned}\delta &\geq 2\exp(-2m\varepsilon^2)\\ \ln\frac{\delta}{2} &\geq -2m\varepsilon^2\\ \varepsilon^2 &\geq \frac{1}{2m}\ln\frac{2}{\delta}\\ \varepsilon &\geq \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}\end{aligned}$$

也就是说此时精度最高为$\sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}$

对精度ε、置信度δ和样本数量m的关系略作分析就能发现：提高置信度是“廉价”的，但提高精度是“昂贵”的

假设ε不变，现将置信度从δ提升为$\frac{1}{10}\delta$，那么所需样本量m'最小为

$$\begin{aligned} m' &\geq \frac{1}{2\varepsilon^2}\ln\frac{20}{\delta} \\ &= \frac{1}{2\varepsilon^2}(\ln\frac{2}{\delta}+\ln\frac{10}{\delta}) \\ &= \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}+\frac{1}{2\varepsilon^2}\ln\frac{10}{\delta} \\ &= m + C(\varepsilon) \end{aligned}$$

可见只需要增加$C(\varepsilon)$个样本就行，并且$C(\varepsilon)$关于提升度是对数的，而且它与δ本身无关。

假设δ不变，现将精度从ε提升为$\frac{1}{10}\varepsilon$，那么所需样本量m'最小为

$$\begin{aligned} m' &\geq \frac{1}{2(\frac{1}{10}\varepsilon)^2} \ln \frac{2}{\delta} \\ &= 100\frac{1}{2\varepsilon^2} \ln \frac{2}{\delta} \\ &= 100m \end{aligned}$$

可见将精度提升10倍所需样本量要增加100倍，所以提升精度是非常昂贵的

如果确定ε、δ、m不变，根据霍夫丁不等式要求$|E(h) - \hat{E}(h)| \geq \varepsilon$的概率越小越好，即要求总有

$$|E(h) - \hat{E}(h)| \leq \varepsilon$$
$$\hat{E}(h) - \varepsilon \leq E(h) \leq \hat{E}(h) + \varepsilon$$

将ε的临界值代进去就有[41]

$$\hat{E}(h) - \sqrt{\frac{1}{2m} \ln \frac{2}{\delta}} \leq E(h) \leq \hat{E}(h) + \sqrt{\frac{1}{2m} \ln \frac{2}{\delta}}$$

[41]霍夫丁不等式的推论

概率近似正确PAC

若假设空间共有$|H|$个模型，则前面的推导略有变化

$$P(\hat{E}(h) - E(h) \geq \varepsilon) \leq |H| \exp(-2m\varepsilon^2)$$
$$P(E(h) - \hat{E}(h) \geq \varepsilon) \leq |H| \exp(-2m\varepsilon^2)$$
$$P(|E(h) - \hat{E}(h)| \geq \varepsilon) \leq 2|H| \exp(-2m\varepsilon^2)$$

因为

$$\delta \geq 2|H| \exp(-2m\varepsilon^2)$$

所以此时有

$$m \geq \frac{1}{2\varepsilon^2} \ln \frac{2|H|}{\delta}, \quad \varepsilon \geq \sqrt{\frac{1}{2m} \ln \frac{2|H|}{\delta}}$$

$$\hat{E}(h) - \sqrt{\frac{1}{2m} \ln \frac{2|H|}{\delta}} \leq E(h) \leq \hat{E}(h) + \sqrt{\frac{1}{2m} \ln \frac{2|H|}{\delta}}$$

对于无限函数空间，我们首先回顾一下前文对生长函数的定义（Jump to）

$$G^{\mathbf{\Lambda}}(l) = \ln \sup_{\mathbf{z}_1,\cdots,\mathbf{z}_l} N^{\mathbf{\Lambda}}(\mathbf{z}_1,\cdots,\mathbf{z}_l)$$

在前面我们也解释过，有些地方定义生长函数时会去掉对数操作，所以$G^{\mathbf{\Lambda}}$也称为对数生长函数。现在用$m_H(N)$表示假设空间H（对应前文的函数空间$\mathbf{\Lambda}$）在N个数据上的（非对数）生长函数

$$m_H(N) = \sup_{\mathbf{z}_1,\cdots,\mathbf{z}_N} N^H(\mathbf{z}_1,\cdots,\mathbf{z}_N)$$

假设函数空间H的VC维为d_{VC}，很明显函数空间的生长函数值要么等于2^N要么等于$2^{d_{VC}}$，所以有

$$m_H(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i}$$

而本身又有[42]

$$\sum_{i=0}^{d_{VC}} \binom{N}{i} \leq N^{d_{VC}} + 1$$

所以有

$$m_H(N) \leq N^{d_{VC}} + 1$$

[42]可利用归纳法证明

无限函数空间的VC维泛化界为

$$E(h) \leq \hat{E}(h) + \sqrt{\frac{8}{N} \ln \frac{4m_H(2N)}{8}}$$

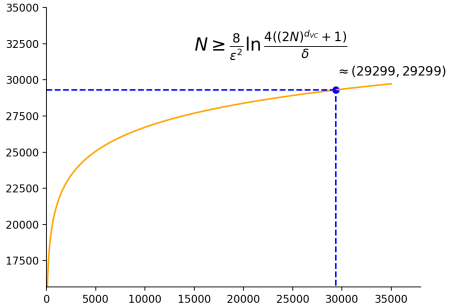
并且在给定ε和δ时有

$$N \geq \frac{8}{\varepsilon^2} \ln \frac{4m_H(2N)}{\delta}$$

$$N \geq \frac{8}{\varepsilon^2} \ln \frac{4((2N)^{d_{VC}} + 1)}{\delta}$$

注意，式子两边都是有N，所以它是关于N的隐式表达，在精度要求不高时可以用启发式方法求解

假设$d_{VC}=3$，如果设定$\varepsilon=0.1,\delta=0.1$，那我们大约需要至少30000个样本
（如果改变d_{VC}的值会发现，$d_{VC}=4$时$N\approx 40000$，$d_{VC}=5$时$N\approx 50000$）[43]



- $N=1000, bound(N)\approx 21193$
- $N=10000, bound(N)\approx 26719$
- $N=20000, bound(N)\approx 28383$
- $N=30000, bound(N)\approx 29356$

[43] N随d_{VC}大致呈线性增长

补充说明：

- $\sum_{i=0}^{d_{VC}} \binom{N}{i} \leq N^{d_{VC}} + 1$ 可利用归纳法证明
- 给定$\varepsilon = 0.1, \delta = 0.1$时$N$随$d_{VC}$线性增长且大约要几万个样本，而现实中统计学习的数据维往往比这小得多（$\frac{l}{h} < 20$），但这不代表学习结果是不可接受的，因为这里的的界是理论上的界，它放的很宽
- 无限函数空间VC维泛化界的计算用到了近似替换，将二次项系数和近似的用指数函数替代，就能得到给出的样子（8就是在近似替换中出现的）

完结

Congrats | 淡小蓝点 bluedotdot.cn

微信号：bluedotdot_cn

More：《PRML Page-by-page》、《面向机器学习（深度学习、人工智能）的数学基础》、《DDPM原理推导及代码实现》、《OpenAI编程基础》等