机器学习

李成龙

安徽大学人工智能学院
“多模态认知计算”安徽省重点实验室
合肥综合性国家科学中心人工智能研究院

内容安排

- 什么是机器学习
- 机器如何学习
- 如何让机器学习的更好
- 为什么机器能学习

内容安排

- **为什么机器能学习**
  - 计算学习理论

本节目录

- **概述**
- **可学习性**
  - PAC可学习
  - VC维

本节目录

- **概述**
- **可学习性**
  - PAC可学习
  - VC维

概述

- **关注问题**
  - 怎样刻画"学习"这个过程
  - 什么样的问题是"可学习的"
  - 什么样的问题是"易学习的"
  - 对于给定的学习算法，能否在理论上预测其性能
  - 理论结果如何指导现实问题的算法设计

概述

- **概念与记号**
  - 样例集：独立同分布样本, 仅考虑二分类问题
    $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\},\boldsymbol{x}_i\in\mathcal{X},y_i\in\mathcal{Y}=\{-1,+1\}.$
  - h为从$\mathcal{X}$到$\mathcal{Y}$的一个映射
    - 泛化误差：分类器的期望误差
      $E(h;\mathcal{D})=P_{x\sim\mathcal{D}}(h(\boldsymbol{x})\neq y)$
    - 经验误差：分类器在给定样例集上的平均误差
      $$\hat{E}(h;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(h(\boldsymbol{x}_i)\neq y_i)$$
  - 由于D是$\mathcal{D}$的独立同分布采样, 因此h的经验误差的期望等于其泛化误差
  - 在上下文明确时, 将$E(h;\mathcal{D})$和$\hat{E}(h;D)$分别简记为$E(h)$和$\hat{E}(h)$

概述

- **概念与记号**
  - 误差参数ε
    - ε为$E(h)$的上限, 即$E(h) \leq \epsilon$.
    - 表示预先设定的学得模型所应满足的误差要求
  - 一致性
    - 若h在数据集D上的经验误差为0, 则称h与D一致, 否则不一致
  - 不合(disagreement)
    - 对于任意两个映射$h_1, h_2 \in \mathcal{X} \to \mathcal{Y}$, 通过"不合"度量它们的差别

$$d(h_1, h_2) = P_{x \sim \mathcal{D}}(h_1(\boldsymbol{x}) \neq h_2(\boldsymbol{x}))$$

本节目录

- 概述
- 可学习性
  - PAC可学习
  - VC维

可学习性

- **什么是“学习”**
  - 概念(concept)
    - 概念是从样本空间$\mathcal{X}$到标记空间$\mathcal{Y}$的映射, 它决定示例$\boldsymbol{x}$的真实标记y
    - 目标概念：如果对任何样例$(\boldsymbol{x}, y)$均有$c(\boldsymbol{x}) = y$成立, 则称c为目标概念
    - 概念类(concept class)：所有我们希望学得的目标概念所构成的集合称为”概念类”, 用符号$\mathcal{C}$表示

可学习性

- **什么是“学习”**
  - 假设空间(hypothesis space)
    - 给定学习算法$\mathcal{L}$, 它所考虑的所有可能概念的集合, 用符号$\mathcal{H}$表示
    - 由于学习算法事先并不知道概念类的真实存在, 因此$\mathcal{H}$和$\mathcal{C}$通常是不同的, 学习算法会把自认为可能的目标概念集中起来构成$\mathcal{H}$
    - 对于$h \in \mathcal{H}$,由于并不能确定它是否真的是目标概念, 因此称为“假设”。显然, h也是从样本空间$\mathcal{X}$到标记空间$\mathcal{Y}$的映射
    - 学习过程可以视为$\mathcal{L}$在$\mathcal{H}$中进行的搜索过程

可学习性

- **什么是“学习”**
  - 可分的与不可分的
    - 可分的(separable)：若目标概念$c \in \mathcal{H}$, 即$\mathcal{H}$中存在假设能将所有的示例完全正确分开(按照与真实标记一致的方式), 则称该问题对学习算法$\mathcal{L}$是“可分的”, 也称“一致的”（consistent)
    - 不可分的(inseparable)：若目标概念$c \notin \mathcal{H}$, 则$\mathcal{H}$中不存在任何假设能将所有的示例完全正确分开，则称该问题对学习算法$\mathcal{L}$是“不可分的”, 也称“不一致的”（non-consistent)

可学习性

- **什么是“学习”**
  - 对于给定训练集D, 我们希望基于学习算法$\mathcal{L}$学得的模型所对应的假设h尽可能接近目标概念c

为什么不是希望精确地学到目标概念c呢？

  - 机器学习过程受到很多因素的制约
    - 获得的训练集D往往仅包含有限数量的样例, 因此通常会存在一些在D上“等效”的假设, 学习算法无法区别这些假设
    - 从分布$\mathcal{D}$采样得到D的过程有一定的偶然性, 即便对同样大小的不同训练集, 学得结果也可能有所不同

可学习性

- **什么是“可学习的”**
  - 概率近似正确(Probably Approximately Correct, PAC)
    - 我们希望以比较大的把握学得比较好的模型, 即以较大概率学得误差满足预设上限的模型
    - 令δ表示置信度，上述要求形式化为：
    - **定义** PAC辨识(PAC Identify)：对0<ε, δ<1, 所有$c\in\mathcal{C}$和分布$\mathcal{D}$,若存在学习算法$\mathcal{L}$, 其输出假设$h\in\mathcal{H}$满足
      $$P(E(h) \leq \epsilon) \geq 1-\delta,$$
    - 则称学习算法$\mathcal{L}$能从假设空间$\mathcal{H}$中PAC辨识概念类$\mathcal{C}$

这样的学习算法$\mathcal{L}$能以较大概率(至少1-δ)学得目标概念c的近似(误差最多为ε)

可学习性

- **什么是“可学习的”**
  - 定义 PAC可学习(PAC Learnable)
    - 令m表示从分布$\mathcal{D}$中独立同分布采样得到的样例数目, 0<ε, δ<1, 对所有分布$\mathcal{D}$,若存在学习算法$\mathcal{L}$和多项式函数poly(·, ·, ·, ·), 使得对于任何m≥poly(1/ε,1/δ,size($\boldsymbol{x}$),size(c)), $\mathcal{L}$能从假设空间$\mathcal{H}$中PAC辨识概念类$\mathcal{C}$, 则称概念类$\mathcal{C}$对假设空间$\mathcal{H}$而言是PAC可学习的，有时也简称概念类$\mathcal{C}$是PAC可学习的
  - PAC可学习性描述的是概念类$\mathcal{C}$的性质，若考虑到对应学习算法$\mathcal{L}$的时间复杂度，则有：
  - 定义 PAC学习算法(PAC Learning Algorithm)
    - 若学习算法$\mathcal{L}$使概念类$\mathcal{C}$为PAC可学习的, 且$\mathcal{L}$的运行时间也是多项式函数poly(1/ε,1/δ,size($\boldsymbol{x}$),size(c)), 则称概念类$\mathcal{C}$是高效PAC可学习(efficiently PAC learnable)的, 称$\mathcal{L}$为概念类$\mathcal{C}$的PAC学习算法

可学习性

- **什么是“可学习的”**
  - 假定学习算法$\mathcal{L}$处理每个样本的时间为常数, 则$\mathcal{L}$的时间复杂度等价于其样本复杂度. 于是, 我们对算法时间复杂度的分析可变为对样本复杂度的分析
  - **定义** 样本复杂度(Sample Complexity)
    - 满足PAC学习算法$\mathcal{L}$所需的$m \geq \text{poly}(1/\varepsilon, 1/\delta, \text{size}(\boldsymbol{x}), \text{size}(c))$中最小的$m$, 称为学习算法$\mathcal{L}$的样本复杂度

可学习性

- **什么是“可学习的”**
  - PAC学习的意义
    - 给出了一个抽象地刻画机器学习能力的框架, 基于这个框架可以对很多重要问题进行理论探讨
      - 研究某任务在什么样的条件下可学得较好的模型
      - 某算法在什么样的条件下可进行有效的学习
      - 需要多少训练样例才能获得较好的模型
    - 把对复杂算法的时间复杂度的分析转为对样本复杂度的分析

可学习性

- **假设空间**
  - 假设空间$\mathcal{H}$的复杂度是影响可学习性的重要因素之一
    - 一般而言, $\mathcal{H}$越大, 其包含任意目标概念的可能性越大, 但从中找到某个具体概念的难度也越大
    - $|\mathcal{H}|$有限时, 我们称$\mathcal{H}$为“有限假设空间”, 否则称为“无限假设空间”
    - 恰PAC可学习(properly PAC learnable)：假设空间$\mathcal{H}$包含了学习算法$\mathcal{L}$所有可能输出的假设, 在PAC学习中假设空间与概念类完全相同, 即$\mathcal{H} = \mathcal{C}$
    - 直观地看, 这意味着学习算法的能力与学习任务“恰好匹配”, 即所有候选假设都来自概念类
    - 然而在现实应用中我们对概念类$\mathcal{C}$通常一无所知, 设计一个假设空间与概念类恰好相同的学习算法通常是不切实际的
  - 研究的重点：当假设空间与概念类不同的情形，即$\mathcal{H} \neq \mathcal{C}$

可学习性

- **有限假设空间**
  - 可分情况
    - 目标概念c属于假设空间$\mathcal{H}$, 即$c \in \mathcal{H}$

给定包含m个样例的训练集D，如何找出满足误差参数的假设呢？

  - 一种简单的学习策略
    - 由于c存在于假设空间$\mathcal{H}$中, 因此任何在训练集D上出现标记错误的假设肯定不是目标概念c
    - 保留与D一致的假设, 剔除与D不一致的假设
    - 若训练集D足够大, 则可不断借助D中的样例剔除不一致的假设, 直到$\mathcal{H}$中仅剩下一个假设为止, 这个假设就是目标概念c

可学习性

- **有限假设空间**
  - 通常情形下, 由于训练集规模有限, 假设空间$\mathcal{H}$中可能存在不止一个与D一致的“等效”假设, 对这些假等效假设, 无法根据D来对它们的优劣做进一步区分

到底需要多少样例才能学得目标概念c的有效近似呢？

  - 训练集D的规模使得学习算法$\mathcal{L}$以概率1-δ找到目标假设的ε近似, 则

$$m \geq \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right).$$

  - 可分情况下的有限假设空间$\mathcal{H}$都是PAC可学习的, 输出假设h的泛化误差随样例数目的增多而收敛到0, 收敛速率为$O(\frac{1}{m})$

可学习性

- **有限假设空间**
  - 不可分情况
    - 对于较困难的学习问题, 目标概念c不属于假设空间$\mathcal{H}$, 即假定对于任何$h \in \mathcal{H}, \hat{E}(h) \neq 0$, $\mathcal{H}$中的任何一个假设都会在训练集上出现或多或少的错误
  - **定理1：** 若$\mathcal{H}$为有限假设空间, $0 < \delta < 1$,则对任意$h \in \mathcal{H}$, 有

$$P\left(|E(h) - \hat{E}(h)| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2m}}\right) \geq 1 - \delta.$$

    - 定理1表明在有限假设集的情况下，当样本大小m足够大时， h的经验误差是其泛化误差很好的近似。此时尽管$c \notin \mathcal{H}$，若能找到$\mathcal{H}$中泛化误差最小的假设也不失为一个较好的选择。定理1实际上指出了一种通用的学习原则：

可学习性

- **有限假设空间**
  - 经验风险最小化（Empirical Risk Minimization，ERM）原则
    - 令h表示学习算法$\mathfrak{L}$输出的假设, 若h满足

$$\hat{E}(h) = \min_{h' \in \mathcal{H}} \hat{E}(h'),$$

则称h为满足经验风险最小化原则的算法

可学习性

- **有限假设空间**
  - 在$c \notin \mathcal{H}$时，可以把PAC学习的定义做如下推广
  - 定义 不可知PAC可学习(agnostic PAC Learnable)
    - 令m表示从分布$\mathcal{D}$中独立同分布采样得到的样例数目, 0<ε, δ<1, 对所有分布$\mathcal{D}$, 若存在学习算法$\mathcal{L}$和多项式poly(·, ·, ·, ·), 使得对于任何m≥poly(1/ε,1/δ,size($\boldsymbol{x}$),size(c)), $\mathcal{L}$能从假设空间$\mathcal{H}$中输出满足下式的假设

$$P(E(h) - \boxed{\min_{h' \in \mathcal{H}} E(h')} \leq \epsilon) \geq 1 - \delta,$$

    - 则称假设空间$\mathcal{H}$是不可知PAC可学习的
  - 定理1说明有限假设集是不可知PAC可学习的

可学习性

- **无限假设空间**
  - 现实学习任务所面临的通常是无限假设空间
    - 实数域中的所有区间
    - $\mathcal{R}^d$空间中的所有线性超平面
  - 欲研究此种情形下的可学习性，需使用$|\mathcal{H}|$之外的方法度量假设空间的复杂性
    - VC维（ Vapnik-Chervonenkis dimension）
    - Rademacher复杂度（Rademacher Complexity）

可学习性

- **VC维**
  - 记号引入
    - 给定假设空间$\mathcal{H}$和示例集$D=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, $\mathcal{H}$中每个假设h都能对D中示例赋予标记, 标记结果可表示为
$$h|_D = \{(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \cdots, h(\boldsymbol{x}_m))\}.$$
    - 随着m的增大, $\mathcal{H}$中所有假设对D中的示例所能赋予标记的可能结果数也会增大
    - 例如, 对于二分类问题：若D中只有两个示例，则赋予标记的可能结果只有4种; 若D中有3个示例，则可能结果有8种

可学习性

- **VC维**
  - 概念引入
    - 增长函数(growth function)
    - 对分(dichotomy)
    - 打散(shattering)

可学习性

- **VC维**
  - 定义 增长函数(growth function)
    - 对所有$m \in N$, 假设空间$\mathcal{H}$的增长函数$\prod_{\mathcal{H}}(\cdot)$为增长函数(growth function)

$$\prod_{\mathcal{H}}(m) = \max_{\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdot,\boldsymbol{x}_m\}\subseteq\mathcal{X}} \left|\{(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \cdots, h(\boldsymbol{x}_m)) \,|h \in \mathcal{H}\}\right|.$$

    - 增长函数表示假设空间对m个示例所能赋予标记的最大可能结果数
    - $\mathcal{H}$对示例所能赋予标记的可能结果数越大, $\mathcal{H}$的表示能力越强, 对学习任务的适应能力也越强
    - 增长函数表述了假设空间$\mathcal{H}$的表示能力, 由此反映出假设空间的复杂度

可学习性

- **VC维**
  - 利用增长函数来估计经验误差与泛化误差之间的关系
  - **定理2** 对假设空间 $\mathcal{H}, m \in \mathbb{N}, 0 < \epsilon < 1$ 和任意 $h \in \mathcal{H}$ 有

$$P(|E(h) - \hat{E}(h)| > \epsilon) \leq 4 \prod_{\mathcal{H}} (2m) \exp(-\frac{m\epsilon^2}{8}).$$

有限假设空间: $P\left(|E(h) - \hat{E}(h)| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2m}}\right) \geq 1 - \delta.$

可学习性

- **VC维**
  - 假设空间$\mathcal{H}$中不同的假设对于D中示例赋予标记的结果可能相同, 也可能不同
  - 尽管$\mathcal{H}$可能包含无穷多个假设, 但是其对D中示例赋予标记的可能结果是有限的: 对于m个示例，最多有2^m个可能结果(二分类)

可学习性

- **VC维**
  - 对分(dichotomy)
    - 对二分类问题来说, $\mathcal{H}$中的假设对D中示例赋予标记的每种可能结果称为对D的一种“对分”
  - 打散(shattering)
    - 若假设空间$\mathcal{H}$能实现示例集D上的所有对分，即 $\prod_{\mathcal{H}}(m) = 2^m$ ，则称示例集D能被假设空间$\mathcal{H}$“打散”

可学习性

- **VC维**
  - 定义 VC维(Vapnik-Chervonenkis dimension)
    - 假设空间$\mathcal{H}$的VC维是能被$\mathcal{H}$打散的最大示例集的大小, 即

$$\mathrm{VC}(\mathcal{H}) = \max\{m : \prod_{\mathcal{H}}(m) = 2^m\}.$$

  - $\mathrm{VC}(\mathcal{H}) = d$ 意味着存在一个大小为d的示例集能被$\mathcal{H}$打散，并且所有大小超过d的示例集都无法被$\mathcal{H}$打散

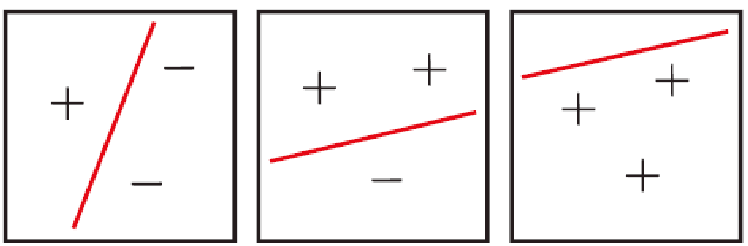
可学习性

- **VC维**
  - **例1** 实数域中的区间$[a, b]$
    - 令$\mathcal{H}$表示实数域中所有闭区间构成的集合$\{h_{[a,b]} : a, b \in \mathbb{R}, a \leq b\}, \mathcal{X} = \mathbb{R}.$
    - 对$\boldsymbol{x} \in \mathcal{X}$, 若$\boldsymbol{x} \in [a, b]$, 则$h_{[a,b]}(x) = +1$, 否则$h_{[a,b]}(x) = -1.$
    - 令x_1=0.5, x_2=1.5, 则假设空间$\mathcal{H}$中存在假设$\{h_{[0,1]}, h_{[0,2]}, h_{[1,2]}, h_{[2,3]}\}$

    将$\{x_1, x_2\}$打散，所以假设空间$\mathcal{H}$的VC维至少为2
    - 对任意大小为3的示例集$\{x_3 , x_4 , x_5\}$, 不妨设$x_3 < x_4 < x_5$, 则$\mathcal{H}$中不存在任何假设$h_{[a,b]}$能实现对分结果$\{(x_3, +), (x_4, -), (x_5, +)\}$
    - 于是, $\mathcal{H}$的VC维为2

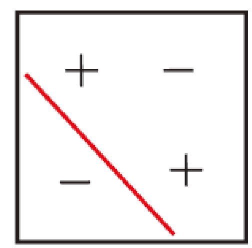
可学习性

- **VC维**
  - **例2** 二维实平面的线性划分
    - 令$\mathcal{H}$表示二维实平面上所有线性划分构成的集合: $\mathcal{X}=\mathcal{R}^2$
    - 由下图可知, 存在大小为3的示例集可被$\mathcal{H}$打散, 但不存在大小为4的示例集可被$\mathcal{H}$打散
    - 于是, $\mathcal{H}$的VC维为3



存在这样的集合，其 $2^3=8$ 种对分均可被线性划分实现

(a) 示例集大小为 3



对任何集合，其 $2^4=16$ 种对分中至少有一种不能被线性划分实现

(b) 示例集大小为 4

可学习性

- **VC维**
  - VC维与增长函数之间的定量关系
    - **Sauer引理** 若假设空间$\mathcal{H}$的VC维为d，则对任意$m \in N$有

$$\prod_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

    - 由Sauer引理可以计算出增长函数的上界
    - **推论** 若假设空间$\mathcal{H}$的VC维为d，则对任意整数$m \geq d$有

$$\prod_{\mathcal{H}}(m) \leq \left(\frac{e \cdot m}{d}\right)^{d}.$$

可学习性

- **VC维**
  - 基于VC维的泛化误差界
    - **定理3** 若假设空间$\mathcal{H}$的VC维为$m>d$，则对任意$m > d, 0 < \delta < 1$和$h \in \mathcal{H}$

$$P\left(E(h) - \hat{E}(h) \leq \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}\right) \geq 1 - \delta.$$

证明：令

$$4\prod_{\mathcal{H}}(2m)\exp(-\frac{m\epsilon^2}{8}) \leq 4(\frac{2em}{d})^d\exp(-\frac{m\epsilon^2}{8}) = \delta$$

解得：

$$\epsilon = \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}},$$

代入中定理2, 于是定理3得证

可学习性

- **VC维**

$$P\left(E(h)-\hat{E}(h)\leq\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}}\right)\geq 1-\delta.$$

- 上式的泛化误差界只与样例数目m有关, 收敛速率为$O\left(\frac{1}{\sqrt{m}}\right)$.
- 上式的泛化误差界与数据分布$\mathcal{D}$及样例集D无关
- 因此, 基于VC维的泛化误差界是

分布无关(distribution-free) & 数据独立(data-independent)

可学习性

- **VC维**
  - 类似定理1，从定理3易看出，当假设空间的VC维有限且样本大小m足够大时，h的经验误差 $\hat{E}(h)$ 是其泛化误差 $E(h)$ 的较好近似，因此对于满足经验风险最小化原则的学习算法$\mathcal{L}$，有下述定理
    - **定理4** 任何VC维有限的假设空间$\mathcal{H}$都是(不可知)PAC可学习的

可学习性

- **Rademacher复杂度**
  - 基于VC维的泛化误差界是分布无关、数据独立的, 这使得基于 VC维的可学习性分析结果具有一定的“普适性”
  - 但由于没有考虑数据自身，因此得到的泛化误差界通常比较“松”
  - 能否将数据的分布也考虑进来
    - Rademacher复杂度(Rademacher complexity)：另一种刻画假设空间复杂度的途径，与VC维不同的是，它在一定程度上考虑了数据分布

可学习性

- **Rademacher复杂度**
  - 给定训练集$D=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$，则假设$h$的经验误差为

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

  - 经验误差最小的假设是

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

可学习性

- **Rademacher复杂度**
  - 若假设标签y_i受到随机因素的影响，不再是x_i的真实标记. 则应该选择$\mathcal{H}$中事先已经考虑了随机噪声影响的假设

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

σ_i为Rademacher随机变量：以0.5的概率取值-1, 0.5的概率取值+1

可学习性

- **Rademacher复杂度**
  - 考虑$\mathcal{H}$中所有的假设, 取期望可得

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

    其中 $\boldsymbol{\sigma}=\{\sigma_1,\sigma_2,\cdots,\sigma_m\}$.
  - 上式的取值范围是[0, 1], 体现了假设空间$\mathcal{H}$的表达能力
    - 当$|\mathcal{H}|=1$时, $\mathcal{H}$中仅有一个假设, 则期望值为0
    - 当$|\mathcal{H}|=2^m$且$\mathcal{H}$能打散D时, 对任意σ总有一个假设使得

$$h(\boldsymbol{x}_i)=\sigma_i(i=1,2,\cdots,m)$$

      此时可计算出期望值为1

可学习性

- **Rademacher复杂度**
  - 定义 Rademacher复杂度(Rademacher complexity)
    - 函数空间$\mathcal{F}$关于$\mathcal{Z}$的经验Rademacher复杂度

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

其中$\mathcal{F}: \mathcal{Z} \to \mathbb{R}$ 为实值函数空间, $Z = \{z_1, z_2, \cdots, z_m\}$, 其中$z_i \in \mathcal{Z}$.

    - 经验Rademacher复杂度衡量了函数空间$\mathcal{F}$与随机噪声在集合$\mathcal{Z}$中的相关性

可学习性

- **Rademacher复杂度**
  - 定义 Rademacher复杂度(Rademacher complexity)
    - 函数空间$\mathcal{F}$关于$\mathcal{Z}$上分布$\mathcal{D}$的Rademacher复杂度

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

    - 基于Rademacher复杂度可得关于函数空间$\mathcal{F}$的泛化误差界

可学习性

- **Rademacher复杂度**
  - **定理5** 对实值函数空间 $\mathcal{F}: \mathcal{Z} \to [0,1]$, 根据分布$\mathcal{D}$从Z中独立同分布采样得到示例 $Z = \{z_1, z_2, \cdots, z_m\}, z_i \in \mathcal{Z}, 0 < \delta < 1,$ 对任意 $f \in \mathcal{F}$, 以至少$1-\delta$的概率有

$$\mathbb{E}[f(z)] \leq \frac{1}{m}\sum_{i=1}^{m} f(z_i) + 2R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}},$$

$$\mathbb{E}[f(z)] \leq \frac{1}{m}\sum_{i=1}^{m} f(z_i) + 2\hat{R}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

  - 定理5中的函数空间是区间 $[0,1]$ 上的实值函数, 因此只适合回归问题

可学习性

- **Rademacher复杂度**
  - **定理 6**　对假设空间$\mathcal{H}: \mathcal{X} \to \{-1, +1\}$,根据分布$\mathcal{D}$从$\mathcal{X}$中独立同分布采样得到示例集$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}, \boldsymbol{x}_i \in \mathcal{X}, 0 > \delta < 1,$ 对任意$h \in \mathcal{H}$,以至少$1 - \delta$ 的概率有

$$E(h) \leq \hat{E}(h) + R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}},$$

$$E(h) \leq \hat{E}(h) + \hat{R}_D(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

  - 定理5只适合回归问题, 定理6适合二分类问题

可学习性

- **Rademacher复杂度**

定理 3 VC维的泛化误差界

若假设空间$\mathcal{H}$的VC维为d，则对任意 $m > d$，$0 < \delta < 1$ 和 $h \in \mathcal{H}$ 有

$$P\left(E(h) - \hat{E}(h) \leq \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}\right) \geq 1 - \delta.$$

定理 6 Rademacher复杂度

对假设空间$\mathcal{H}: \mathcal{X} \to \{-1, +1\}$,根据分布$\mathcal{D}$从$\mathcal{X}$中独立同分布采样得到示例集$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}, \boldsymbol{x}_i \in \mathcal{X}, 0 > \delta < 1$, 对任意 $h \in \mathcal{H}$, 以至少 $1-\delta$的概率有

$$E(h) \leq \hat{E}(h) + R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}},$$

$$E(h) \leq \hat{E}(h) + \hat{R}_D(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

定理3(基于VC维的泛化误差界)　与分布无关、数据独立的

定理6(基于Rademacher复杂度的泛化误差界)　与分布$\mathcal{D}$或样本D有关.

基于Rademacher复杂度的泛化误差界依赖于具体学习问题的数据分布,类似于为该问题“量身定制”的, 因此它通常比基于VC维的泛化误差界要更紧一些

可学习性

- **Rademacher复杂度**
  - Rademacher复杂度与增长函数之间的关系
    - **定理7** 假设空间$\mathcal{H}$的Rademacher复杂度为$R_m(\mathcal{H})$ 与增长函数$\prod_{\mathcal{H}}(m)$满足：

$$R_m(\mathcal{H}) \leq \sqrt{\frac{2\ln \prod_{\mathcal{H}}(m)}{m}}.$$

    - 由定理6、定理7、推论2可得

$$E(h) \leq \hat{E}(h) + \sqrt{\frac{2d\ln\frac{em}{d}}{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

    - 从Rademacher复杂度和增长函数能推导出基于VC维的泛化误差界

可学习性

- **稳定性**
  - 无论基于VC维和Rademacher复杂度来分析泛化性能, 得到的结果均与具体的学习算法无关, 这使得人们能够脱离具体的学习算法来考虑学习问题本身的性质
  - 但另一方面, 为了获得与算法有关的分析结果, 则需另辟蹊径
  - 稳定性(stability)分析是这方面值得关注的一个方向
    - 考察算法在输入(训练集)发生变化时，输出是否发生较大的变化

总结

- **概述**
  - 关注的问题
  - 一些概念及记号
- **可学习性**
  - 什么是“学习”
  - 什么是“可学习的”（PAC）
  - 假设空间复杂性对可学习性的影响
    - 有限假设空间
    - 无限假设空间：基于VC维的分析
    - 无限假设空间：基于Rademacher复杂度的分析