周志华 著

MACHINE LEARNING

机器学习

清华大学出版社

本课件版权所有@LAMDA，为本书教学目的可免费使用，
其他目的的需征得本书作者同意

本章课件致谢：
郭翔宇
王魏

第十二章：计算学习理论

纲要

- 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”
  - 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
- 稳定性

关注的问题

- 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？

一些概念及记号

- 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 $\hat{E}(h)$.

一些概念及记号

- 误差参数ϵ

  ϵ 为 $E(h)$的上限，即 $E(h) \leq \epsilon$.

  $\Rightarrow$ 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$h_1, h_2 \in \mathcal{X} \rightarrow \mathcal{Y}$，通过 “不合” 度量它们的差别

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

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.

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。

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”(separable), 也称“一致的”（consistent).

  - 不可分的(separable)

    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”(non-separable), 也称“不一致的”（non-consistent).

- 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.

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$0 < \epsilon, \delta < 1$,所有$c \in \mathcal{C}$和分布$\mathcal{D}$,若存在学习算法$\mathfrak{L}$，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$\mathfrak{L}$能从假设空间$\mathcal{H}$中PAC辨识概念类$\mathcal{C}$.

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

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, $0 < \epsilon, \delta < 1$, 对所有分布 $\mathcal{D}$,若存在学习算法 $\mathcal{L}$ 和多项式函数 $\mathrm{poly}(\cdot,\cdot,\cdot,\cdot)$, 使得对于任何 $m \geq \mathrm{poly}(1/\epsilon, 1/\delta, \mathrm{size}(\boldsymbol{x}), \mathrm{size}(c))$, $\mathcal{L}$ 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。

什么是“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 $\text{poly}(1/\epsilon, 1/\delta, \text{size}(\boldsymbol{x}), \text{size}(c))$，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。

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 $m \geq \mathrm{poly}(1/\epsilon, 1/\delta, \mathrm{size}(\boldsymbol{x}), \mathrm{size}(c))$ 中最小的m，称为学习算法$\mathcal{L}$的样本复杂度。

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

- 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”.

- 假设空间 $\mathcal{H}$ 的复杂度是影响学习任务难度的重要因素之一
  - 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$\mathcal{H} \neq \mathcal{C}$时.

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.

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来对它们的有优劣做进一步区分.

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$\mathfrak{L}$以概率$1-\delta$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$O(\frac{1}{m})$.

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12.1

若$\mathcal{H}$为有限假设空间$0 < \delta < 1$,则对任意$h \in \mathcal{H}$，有

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

有限假设空间

不可分情况

定理12.1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12.1实际上指出了一种通用的学习原则：

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$\mathfrak{L}$为满足经验风险最小化原则的算法.

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，$0 < \epsilon, \delta < 1$，对所有分布 $\mathcal{D}$，若存在学习算法 $\mathfrak{L}$ 和多项式函数 $\mathrm{poly}(\cdot,\cdot,\cdot,\cdot)$，使得对于任何 $m \geq \mathrm{poly}(1/\epsilon, 1/\delta, \mathrm{size}(\boldsymbol{x}), \mathrm{size}(c))$，$\mathfrak{L}$ 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12.1说明有限假设集是不可知PAC可学习的

- 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
  - $\mathbb{R}^d$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）

VC维

记号引入

给定假设空间$\mathcal{H}$和示例集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$, $\mathcal{H}$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。

VC维

概念引入

- 增长函数(growth function)
- 对分(dichotomy)
- 打散(shattering)

VC维

定义 增长函数(growth function)

对所有$m \in \mathbb{N}$, 假设空间$\mathcal{H}$的增长函数$\prod_{\mathcal{H}}(\cdot)$为:

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

- 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.

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**12.2**

对假设空间$\mathcal{H}$, $m \in \mathbb{N}$, $0 < \epsilon < 1$和任意$h \in \mathcal{H}$有

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

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).

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”。

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

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

- $\mathrm{VC}(\mathcal{H}) = d$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。

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$\{h_{[a,b]} : a, b \in \mathbb{R}, a \le b\}, \mathcal{X} = \mathbb{R}$.

对 $x \in \mathcal{X}$, 若 $x \in [a, b]$, 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.

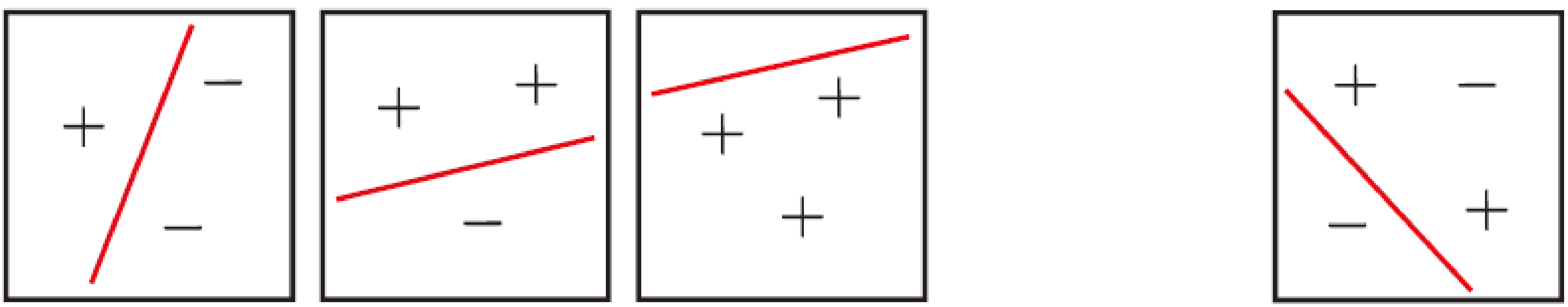
VC维

- 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，$\mathcal{X}=\mathbb{R}^2$.

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，二维实平面上所有线性划分构成的假设空间 $\mathcal{H}$ 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

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

(a) 示例集大小为 3

(b) 示例集大小为 4

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$m \in \mathbb{N}$有

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

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

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

VC维

基于VC维的泛化误差界:

定理12.3

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

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

证明:

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

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

代入中定理12.2, 于是定理12.3得证.

VC维

VC维的泛化误差界:

定理12.3

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

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

- 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

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

VC维

类似定理12.1，从定理12.3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12.4

任何VC维有限的假设空间$\mathcal{H}$都是(不可知)PAC可学习的.

- 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“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.

Rademacher复杂度

给定训练集$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$，

则假设h的经验误差为

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

Rademacher复杂度

给定训练集$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$,

则假设h的经验误差为

$$\hat{E}(h)=\frac{1}{2}-\frac{1}{2m}\sum_{i=1}^{m}y_i h(\boldsymbol{x}_i)$$

- 其中$\frac{1}{m}\sum_{i=1}^{m}y_i h(\boldsymbol{x}_i)$体现了预测值$h(\boldsymbol{x}_i)$与样例真实标记$y_i$之间的一致性.
- 若对于所有的$i\in\{1,2,\cdots,m\}$, 都有$h(\boldsymbol{x}_i)=y_i$,则$\frac{1}{m}\sum_{i=1}^{m}y_i h(\boldsymbol{x}_i)$取最大值1.
- 经验误差最小的假设是

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

Rademacher复杂度

- 经验误差最小的假设是

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

- 若假设标签y_i受到随机因素的影响，不再是$\boldsymbol{x}_i$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.

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$\boldsymbol{\sigma}$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.

Rademacher复杂度

定义 **Rademacher复杂度(Rademacher complexity)**

函数空间$\mathcal{F}$关于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$\mathcal{F} : \mathcal{Z} \rightarrow \mathbb{R}$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Z中的相关性。

Rademacher复杂度

定义 **Rademacher复杂度(Rademacher complexity)**

函数空间$\mathcal{F}$关于Z上分布$\mathcal{D}$的经验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.

Rademacher复杂度

定理12.5

对实值函数空间$\mathcal{F}:\mathcal{Z}\to[0,1]$, 根据分布$\mathcal{D}$从$\mathcal{Z}$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12.5中的函数空间$\mathcal{F}$是区间$[0,1]$上的实值函数，因此只适合回归问题。

Rademacher复杂度

定理12.6 ：基于Rademacher复杂度的泛化误差界

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12.5只适合回归问题，定理12.6适合二分类问题。

Rademacher复杂度

定理12.3 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12.6 Rademacher复杂度

对假设空间$\mathcal{H}: \mathcal{X} \rightarrow \{-1, +1\}$,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12.3(基于VC维的泛化误差界)与分布无关、数据独立的；
- 定理12.6(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.

Rademacher复杂度

- Rademacher复杂度与增长函数之间的关系:

定理12.7

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12.6、定理12.7、推论12.2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.

- 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.

稳定性(Stability)

训练集的两种变化

给定$D = \{\boldsymbol{z}_1 = (\boldsymbol{x}_1, y_1), \boldsymbol{z}_2 = (\boldsymbol{x}_2, y_2), \cdots, \boldsymbol{z}_m = (\boldsymbol{x}_m, y_m)\}, \boldsymbol{x}_i \in \mathcal{X}$是来自分布$\mathcal{D}$的独立同分布示例，$y_i = \{-1, +1\}$. 对假设空间$\mathcal{H} : \mathcal{X} \to \{-1, +1\}$和学习算法$\mathfrak{L}$，令$\mathfrak{L}_D \in \mathcal{H}$表示基于训练集$D$从假设空间$\mathcal{H}$中学得的假设.

- $D^{\setminus i}$表示移除D中第i个样例得到的集合

$$D^{\setminus i} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_{i-1}, \boldsymbol{z}_{i+1}, \cdots, \boldsymbol{z}_m\},$$

- D^i表示替换D中第i个样例得到的集合

$$D^i = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_{i-1}, \boldsymbol{z}'_i, \boldsymbol{z}_{i+1}, \cdots, \boldsymbol{z}_m\},$$

其中$\boldsymbol{z}'_i = (\boldsymbol{x}'_i, y'_i), \boldsymbol{x}'_i$服从分布$\mathcal{D}$并独立于$D$.

稳定性(Stability)

损失函数

$\ell(\mathcal{L}_D(\boldsymbol{x}), y) : \mathcal{Y} \times \mathcal{Y} \longmapsto \mathbb{R}^+$ 刻画假设$\mathcal{L}_D$的预测标记$\mathcal{L}_D(\boldsymbol{x})$与真实标记 y 之间的差别，简记为$\ell(\mathcal{L}_D, \boldsymbol{z})$.

- 泛化损失

$$\ell(\mathcal{L}, D) = \mathbb{E}_{x \in \mathcal{X}, z=(\boldsymbol{x}, y)}[\ell(\mathcal{L}_D, \boldsymbol{z})].$$

- 经验损失

$$\hat{\ell}(\mathcal{L}, D) = \frac{1}{m} \sum_{i=1}^{m} \ell(\mathcal{L}_D, \boldsymbol{z}_i).$$

- 留一(leave-one-out)损失：

$$\ell_{loo}(\mathcal{L}, D) = \frac{1}{m} \sum_{i=1}^{m} \ell(\mathcal{L}_{D^{\setminus i}}, \boldsymbol{z}_i).$$

稳定性(Stability)

定义 **算法的均匀稳定性(uniform stability)**

对任何 $\boldsymbol{x} \in \mathcal{X}, \boldsymbol{z} = (\boldsymbol{x}, y)$，若学习算法$\mathfrak{L}$满足

$$|\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z})| \leq \beta, i = 1, 2, \cdots, m,$$

则称 $\mathfrak{L}$关于损失函数ℓ满足β−均匀稳定性.

- 若算法 $\mathfrak{L}$关于损失函数ℓ满足β−均匀稳定性，则有

$$\begin{aligned} & |\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^i}, \boldsymbol{z})| \\ \leq \; & |\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z})| + |\ell(\mathfrak{L}_{D^i}, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z})| \\ \leq \; & 2\beta \end{aligned}$$

也就是说，移除示例的稳定性包含替换示例的稳定性.

稳定性(Stability)

- 若损失函数ℓ有界

  对所有D和$\boldsymbol{z}=(\boldsymbol{x},y)$有$0\leq l(\mathfrak{L}_D,\boldsymbol{z})\leq M$,则有

定理12.8

给定从分布$\mathcal{D}$上独立同分布采样得到的大小为m的示例集D,若学习算法$\mathfrak{L}$满足关于损失函数ℓ的β–均匀稳定性，且损失函数ℓ的上界为M，同时$0<\delta<1$,则对任意$m\geq 1$,以至少$1-\delta$ 的概率有

$$\ell(\mathfrak{L},\mathcal{D})\leq\hat{\ell}(\mathfrak{L},D)+2\beta+(4m\beta+M)\sqrt{\frac{\ln(1/\delta)}{2m}},$$

$$\ell(\mathfrak{L},\mathcal{D})\leq\ell_{loo}(\mathfrak{L},D)+\beta+(4m\beta+M)\sqrt{\frac{\ln(1/\delta)}{2m}}.$$

稳定性(Stability)

- 定理12.8给出了基于稳定性分析推导出的学习算法$\mathcal{L}$学得假设的泛化误差界.
  - 经验损失与泛化损失之间差别的收敛率为$\beta\sqrt{m}$;
    若$\beta = O(\frac{1}{m})$则可保证收敛率为$O(\frac{1}{\sqrt{m}})$.
  - 该收敛率与基于VC维和Rademacher复杂度得到的收敛率一致.

稳定性(Stability)

- 学习算法的稳定性分析关注的是$|\hat{\ell}(\mathfrak{L}, D) - \ell(\mathfrak{L}, \mathcal{D})|$;
- 假设空间复杂度分析所关注的是$\sup_{h\in\mathcal{H}} |\hat{E}(h) - E(h)|$.

因此，稳定性分析不必考虑假设空间中所有可能的假设，只需根据分析算法自身的特性(稳定性)来讨论输出假设$\mathfrak{L}_D$的泛化误差界.

稳定性与可学习性之间有什么关系呢？

稳定性(Stability)

首先必须假设$\beta\sqrt{m} \to 0$, 这样才能保证稳定的学习算法具有一定泛化能力, 即经验损失收敛于泛化损失, 否则可学习性无从谈起。对于满足ERM原则的学习算法，有如下定理:

定理12.9

若学习算法$\mathcal{L}$是ERM且稳定的, 则假设空间$\mathcal{H}$可学习.

- 学习算法的稳定性能导出假设空间的可学习性.
- 稳定性和假设空间可通过损失函数ℓ联系起来.

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”
  - 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
- 稳定性