

# Understanding Machine Learning

Yuyang Zhang

2017年7月24日

## 目录

<b>1</b>	<b>why can machine learn</b>	<b>4</b>
1.1	基本符号 . . . . .	4
1.1.1	输入 . . . . .	4
1.1.2	输出 . . . . .	4
1.1.3	数据生成模型 . . . . .	4
1.1.4	衡量标准 . . . . .	4
1.2	经验风险最小化 . . . . .	5
1.2.1	过拟合 . . . . .	5
1.3	归纳偏好 . . . . .	5
1.4	为什么可以学习到东西 . . . . .	6
1.5	思路总结 . . . . .	8
<b>2</b>	<b>Probably Approximately Correct</b>	<b>10</b>
2.1	PAC学习理论 . . . . .	10
2.1.1	PAC可学习 . . . . .	10
2.1.2	采样复杂度 . . . . .	10
2.2	泛化PAC理论 . . . . .	10
2.3	不可知PAC学习 . . . . .	11
2.4	学习问题建模 . . . . .	11
2.4.1	广义损失函数 . . . . .	12
2.5	思路总结 . . . . .	12
<b>3</b>	<b>Learning via Uniform Convergence</b>	<b>13</b>
3.1	学习的一致收敛性 . . . . .	13
3.2	有限假设是不可知PAC可学习的 . . . . .	13
<b>4</b>	<b>附录1: 不等式证明</b>	<b>15</b>
4.1	Markov Inequality . . . . .	15
4.2	引理1 . . . . .	15
4.3	Chebyshev Inequality . . . . .	15
4.4	引理2 . . . . .	16
4.5	Chernoff Bound . . . . .	16
4.6	引理3 . . . . .	17

目录	3
4.7 Hoeffding Inequality . . . . .	18

# 1 why can machine learn

## 1.1 基本符号

### 1.1.1 输入

Domain Set: 一个任意集合 $\mathcal{X}$ ，也作领域集。可以理解为所有样本的集合，其中每个样本通常以一个能够表征其特征的向量表示。

Label Set: 标签集 $\mathcal{Y}$ 。样本所属于的类别，通常二分类问题，标签集为 $\{0, 1\}$ 或者是 $\{-1, +1\}$ 。

Training Data: 训练数据 $\mathcal{S} = \{(x_1, y_1), \dots, (x_m, y_m)\}$ ，也叫训练集，样本集。

### 1.1.2 输出

Predicting Rule: 预测规则， $h : \mathcal{X} \rightarrow \mathcal{Y}$ 。该规则是一个由样本集到标签集的映射，可以理解为预测器(predictor)，假设(hypothesis)，分类器(classifier)，等等。

### 1.1.3 数据生成模型

我们假定样本 $\mathcal{S}$ 是由概率分布 $\mathcal{D}$ 生成，并且根据一个标记函数 $f : \mathcal{X} \rightarrow \mathcal{Y}$ ，来标记样本的类别，对于任意 $i = 1, \dots, m$ 都有  $y_i = f(x_i)$ ，然而我们并不知道概率分布 $\mathcal{D}$ 与标记函数 $f$ ，我们的目标就是找一个合适的假设 $h$ ，令其与标记函数 $f$ 可以对样本做出相同的标记。

### 1.1.4 衡量标准

我们定义分类误差为：未能成功预测随机数据点正确标签的概率，即对于随机的一个 $x \in \mathcal{X}$ ， $h(x) \neq f(x)$ 的概率。

定义 $\mathcal{A} \subseteq \mathcal{X}$ 为一个领域子集， $\mathcal{A}$ 中的任意实例 $x \in \mathcal{A}$ 的出现概率由 $\mathcal{D}(\mathcal{A})$ 所决定。通常，我们称 $\mathcal{A}$ 为一个事件， $\mathcal{A} = \{x \in \mathcal{X} : \pi(x) = 1\}$ ，其中 $\pi : \mathcal{X} \rightarrow \{0, 1\}$ ，表示样本是否被观测到。我们也将 $\mathcal{D}(\mathcal{A})$ 写作 $\mathbb{P}_{x \sim \mathcal{D}}[\pi(x)]$ 。此时我们可以定义假设 $h$ 的错误率为：

$$\mathcal{L}_{\mathcal{D}, f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] \stackrel{\text{def}}{=} \mathcal{D}(\{x | h(x) \neq f(x)\})$$

其中误差的测量是基于概率分布 $\mathcal{D}$ 和标记函数 $f$ 的， $\mathcal{L}_{\mathcal{D},f}(h)$ 也称为泛化误差，**损失**或者 $h$ 的真实误差。

## 1.2 经验风险最小化

机器学习的过程都是基于训练集 $\mathcal{S}$ 的，训练集 $\mathcal{S}$ 由未知分布 $\mathcal{D}$ 从领域集 $\mathcal{X}$ 中采样得出，并由标记函数 $f$ 标记，机器学习的输出是一个基于训练集 $\mathcal{S}$ 的假设， $h_{\mathcal{S}} : \mathcal{X} \rightarrow \mathcal{Y}$ 。

由于我们并不知道分布 $\mathcal{D}$ 与标记函数 $f$ ，所以我们只能根据训练集 $\mathcal{S}$ 来判断我们所选择假设的表现，定义训练误差为：

$$\mathcal{L}_{\mathcal{S}} \stackrel{\text{def}}{=} \frac{|\{x_i | h(x_i) \neq y_i, i = 1, \dots, m\}|}{m}$$

训练误差也称作经验误差和经验风险。

因为训练集 $\mathcal{S}$ 是领域 $\mathcal{X}$ 的一个子集，所以训练样本是真实世界的一个缩影，正如概率统计的一个核心思想，通过样本反映总体。所以我们认为利用样本集寻找一个较好的假设是可行的，即最小化训练误差 $\mathcal{L}_{\mathcal{S}}(h)$ ，这称之为经验风险最小化，ERM(Experience Risk Minimize)。

### 1.2.1 过拟合

一个假设在训练集上效果优异，但在真实世界中表现却糟糕，这种现象称之为过拟合。当我们过度追求经验风险最小化原则时，我们就有可能面临过拟合的风险。

## 1.3 归纳偏好

虽然经验风险最小化会有过拟合的风险，相比于抛弃这个原则，我们更愿意去修正这个原则，考虑我们对于假设的归纳偏好。我们通常的解决方案是根据我们定好的归纳偏好，在有限的假设空间中去搜索所要用的假设，这些假设的集合成为假设类，记为 $\mathcal{H}$ ，则我们的学习过程可以记为：

$$ERM_{\mathcal{H}}(\mathcal{S}) \in \arg \max_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{S}}(h)$$

我们常见的正则化，就是一种归纳偏好， $L1$ 正则化表明我们的归纳偏好是更喜欢参数稀疏的假设。

### 1.4 为什么可以学习到东西

本节旨在说明：当拥有足够多样本时，在有限假设空间 $\mathcal{H}$ 中，经验风险最小化 $ERM_{\mathcal{H}}$ 原则不会出现过拟合，即我们通过训练样本，可以找到一个足够好的假设 $h_s$ ，在真实世界中表现也足够好。

**定义1.1 可实现性假设** 存在 $h^* \in \mathcal{H}$ ，使得 $\mathcal{L}_{\mathcal{D},f}(h^*) = 0$ 。

该假设意味着，对于随机样本集 $\mathcal{S}$ ，由概率分布 $\mathcal{D}$ 采样，由标记函数 $f$ 标记，以概率1使得 $L_{\mathcal{S}}(h^*) = 0$ ，其中样本集 $\mathcal{S}$ 中的样本是独立同分布的。我们定义 $h_{\mathcal{S}}$ 为对 $\mathcal{S}$ 利用 $ERM_{\mathcal{H}}$ 得到的结果：

$$h_{\mathcal{S}} \in \arg \max_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{S}}(h)$$

因为样本集 $\mathcal{S}$ 仍是领域集 $\mathcal{X}$ 的子集，是根据分布随机得到的实例集合，会有一定概率使得采样得到的样本不具有代表性，并不能反映真实的总体情况，所以根据样本集 $\mathcal{S}$ 得到的假设 $h_{\mathcal{S}}$ 并不一定准确，在真实世界中的表现有可能很差。因此，我们选择一定程度的容忍，容忍会有一定几率采样到不具有代表性的样本，一般来说，我们定义采样得到不具有代表性样本的概率之多 $\delta$ ，则 $1 - \delta$ 为置信参数。同时对于假设的预测效果，我们能容忍的损失上限为 $\epsilon$ ，称为精度参数，如果 $\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon$ ，那么这是一个差的假设，反之，则是一个好的假设。

对于我们的样本集 $\mathcal{S}$ ，最好的假设 $h_{\mathcal{S}}$ 仍有 $\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon$ ，则这组样本采样失败，因为我们的 $ERM_{\mathcal{H}}$ 无法从 $\mathcal{S}$ 中学到有用的东西。当样本集中所有样本 $(x_1, \dots, x_m)$ ，都不具有代表性，我们记为：

$$\{x_i | x_i \in \mathcal{S}, i = 1, \dots, m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}$$

则采样失败的概率上界为：

$$\mathcal{D}^m(\{x_i | x_i \in \mathcal{S}, i = 1, \dots, m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\})$$

因为每个样本都是不具有代表性的样本，所以是采样失败的概率的上界。设 $\mathcal{H}_B$ 为差的假设的集合：

$$\mathcal{H}_B = \{h | \mathcal{L}_{\mathcal{D},f}(h) > \epsilon, h \in \mathcal{H}\}$$

同时设：

$$M = \{x | x \in \mathcal{S}, \exists h \in \mathcal{H}_B, \mathcal{L}_{\mathcal{S}}(h) = 0\}$$

为误导集，误导集使差的假设在训练样本 $\mathcal{S}$ 上表现良好，但在真实世界中表现较差。因为有可实现假设

$$\exists h^*, \mathcal{L}_{\mathcal{D},f}(h^*) = 0$$

且有

$$h_{\mathcal{S}} \in \arg \max_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{S}}(h)$$

产生 $\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon$  是因为样本集不好，样本集采样失败，所以，当且仅当 $\mathcal{S} \subseteq M$ 时，才会出现 $\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon$ ，我们将其表示为：

$$\{x_i | x_i \in \mathcal{S}, i = 1, \dots, m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\} \subseteq M$$

所以我们有

$$\mathcal{D}^m(\{x_i | x_i \in \mathcal{S}, i = 1, \dots, m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}) \leq \mathcal{D}^m(M)$$

而 $M$ 又可以写作：

$$M = \bigcup_{h \in \mathcal{H}_B} \{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\}$$

则：

$$\mathcal{D}^m(\{x_i | x_i \in \mathcal{S}, i = 1, \dots, m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}) \leq \mathcal{D}^m(\bigcup_{h \in \mathcal{H}_B} \{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\})$$

由 $P(A \cup B) \leq P(A) + P(B)$ 得：

$$\mathcal{D}^m(\bigcup_{h \in \mathcal{H}_B} \{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\}) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\})$$

其中我们将 $\mathcal{S}$ 拆开，

$$\mathcal{D}^m(\{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\}) = \mathcal{D}^m(\{x_i | h(x_i) = f(x_i), x_i \in \mathcal{S}, i = 1, \dots, m\})$$

$$\mathcal{D}^m(\{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\}) = \prod_{i=1}^m \mathcal{D}(\{x_i | h(x_i) = f(x_i), x_i \in \mathcal{S}\})$$

根据 $1 - \epsilon \leq e^{-\epsilon}$ ，对于等号右边的连乘的每一项都有：

$$\mathcal{D}(\{x_i | h(x_i) = y_i, x_i \in \mathcal{S}\}) = 1 - \mathcal{L}_{\mathcal{D},f}(h) \leq 1 - \epsilon$$

对于所有的  $h \in \mathcal{H}_B$ :

$$\mathcal{D}^m(\{x|x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\}) \leq (1 - \epsilon)^m \leq e^{-\epsilon m}$$

可得:

$$\mathcal{D}^m(\{x_i|x_i \in \mathcal{S}, i = 1, \dots, m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}) \leq |\mathcal{H}_B|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m}$$

所以, 若我们对采样失败概率的容忍大于采样失败的概率上限, 我们认为最后得到的  $h_{\mathcal{S}}$  学到了有用的东西的, 记为:

$$|\mathcal{H}|e^{-\epsilon m} \leq \delta$$

借此我们可以推断出我们需要拥有的样本数量为:

$$m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$$

**推论1.1** 设  $\mathcal{H}$  为一个有限假设集合,  $\delta \in (0, 1)$ ,  $\epsilon > 0$ , 当

$$m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$$

成立时, 从而对于任何标记函数  $f$ 、任何分布  $\mathcal{D}$ , 可实现性假设最少以  $1 - \delta$  的概率, 对于每个  $ERM$  假设  $h_{\mathcal{S}}$ , 有以下不等式成立:

$$\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) \leq \epsilon$$

上述推论表明, 对于足够大的  $m$ , 由  $ERM_{\mathcal{H}}$  规则生成的有限假设会以  $1 - \delta$  概率得到小于误差  $\epsilon$  的近似正确解, 概率在于容忍采样失败概率  $\delta$ , 近似在于容忍误差  $\epsilon$ , 即我们的算法, 可能 (容忍了失败概率) 会学到令我们基本满意 (容忍了一定误差) 的东西。而概率近似正确, 即 PAC 理论, 将在第二章详述。

## 1.5 思路总结

1. 为什么不能有效学习? 为什么会  $\mathcal{L}_{\mathcal{D},f}(h) > \epsilon$ ?
2. 因为样本不具有代表性, 学到的东西有问题, 采样失败。
3. 怎么解决?



4. 降低采样失败概率至我们可以容忍的范围，小于 $\delta$ 。
5. 采样失败的概率上限小于 $\delta$ 。
6.  $\sup \mathcal{D}^m(\{x_i | x_i \in \mathcal{S}, i = 1, \dots, m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}) = |\mathcal{H}|e^{-\epsilon m} \leq \delta$
7.  $m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$
8. 当样本足够充足时，学习到的假设概率近似正确。

## 2 Probably Approximately Correct

### 2.1 PAC学习理论

#### 2.1.1 PAC可学习

接着上一章继续说，在经验风险最小化准则下，对于一个有限假设类，如果有足够多的训练样本，则我们输出的学习算法在真实世界中，是概率近似正确的。我们做以下定义：

**定义2.1 PAC可学习** 若存在一个函数  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathcal{N}$  和一个学习算法，使得对于给定的  $\epsilon, \delta \in (0, 1)$  和任一分布  $\mathcal{D}$ 、任一标记函数  $f : \mathcal{X} \rightarrow \{0, 1\}$ ，可实现假设成立时，那么当样本数量  $m \geq m_{\mathcal{H}}(\epsilon, \delta)$  时，算法将以不小于  $1 - \delta$  的概率返回一个使

$$\mathcal{L}_{\mathcal{D}, f}(h) \leq \epsilon$$

的假设  $h$ 。

#### 2.1.2 采样复杂度

函数  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathcal{N}$  决定了假设的采样复杂度，即可以学到东西时所需要的样本数量。采样复杂度不仅依赖于  $\epsilon$  和  $\delta$ ，同样还依赖于假设空间中的假设数量：

$$m_{\mathcal{H}} = \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$$

其与假设数量的对数成正比。

**推论2.1** 当采样复杂度满足：

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$

则任一有限类假设是PAC可学习的。

$m_{\mathcal{H}}(\epsilon, \delta)$  被称为假设结合  $\mathcal{H}$  的采样复杂度的最小函数。由上述定义我们可知，一个假设是否是PAC可学习的，还依赖于假设空间的大小，而假设空间则与VC维相关，将会在后几章说明。

### 2.2 泛化PAC理论

为了让该理论更贴近实际，则我们考虑将理论的前提约束放宽，进行泛化。

1. 去掉可实现性假设。
2. 考虑多分类、回归等问题。

### 2.3 不可知PAC学习

当我们放弃可实现假设时，我们称之为不可知，具体可以理解为：完全一样的样本，却又不同的标记，这种情况下，我们的算法如何学到东西。

从此处开始，为了简写，我们将  $\mathcal{D}$  定义为  $\mathcal{X} \times \mathcal{Y}$  上的概率分布，即将之前  $\mathcal{D}$  和  $f$  简写在一起，作为领域集和标签集的联合概率分布，定义真实误差为：

$$\mathcal{L}_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathbb{P}_{(x,y) \sim \mathcal{D}}[h(x) \neq y] \stackrel{\text{def}}{=} \mathcal{D}(\{(x,y) : h(x) \neq y\})$$

在面对不可知的情况下，最好的预测器为贝叶斯预测器：

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \mathbb{P}[y = 1|x] \geq \frac{1}{2} \\ 0 & \text{else} \end{cases}$$

我们易证贝叶斯预测器是最优的：对于每个领域集的实例  $x \in \mathcal{X}$  都选择数量较多的类别，即每个实例的分类正确概率都大于  $1/2$ ，没有其他分类器可以达到比它更低的错误率，对于任意预测器  $g$  都有  $\mathcal{L}_{\mathcal{D}}(f_{\mathcal{D}}) \leq \mathcal{L}_{\mathcal{D}}(g)$ ，所以该预测器最优。但我们并不知道实际的概率分布  $\mathcal{D}$ ，所以我们并没法使用这样的预测器。

所以对于不可知问题，我们只能选择容忍（看来统计的机器学习就是一个容忍的过程Orz）：

**定义2.2 不可知PAC可学习** 若存在一个函数  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathcal{N}$  和一个学习算法，使得对于给定的  $\epsilon, \delta \in (0, 1)$  和任一分布  $\mathcal{D}$ ，当样本数量  $m \geq m_{\mathcal{H}}(\epsilon, \delta)$  时，算法将以不小于  $1 - \delta$  的概率返回一个使

$$\mathcal{L}_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h') + \epsilon$$

成立的假设  $h$ 。

### 2.4 学习问题建模

此节我们讨论学习问题建模，总的来说，除了二分类问题，学习任务还分为以下几种：

- 多分类
- 回归

虽然说学习任务分为多种，但主要区别只在于损失函数不同。

### 2.4.1 广义损失函数

给定任意集合 $\mathcal{H}$ 和定义域 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ ，令 $\ell$ 为 $\mathcal{H} \times \mathcal{Z}$ 到非负实数的映射，记为 $\ell: \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}_+$ ， $\ell$ 就是损失函数。现在我们重新定义真实误差和经验误差：

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(h) &\stackrel{\text{def}}{=} \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)] \\ \mathcal{L}_{\mathcal{S}}(h) &\stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^m \ell(h, z_i)\end{aligned}$$

同时我们定义常用的损失函数：

- 0-1损失

$$\ell_{0-1}(h, (x, y)) \stackrel{\text{def}}{=} \begin{cases} 0 & h(x) = y \\ 1 & h(x) \neq y \end{cases}$$

- 平方损失

$$\ell_{sq}(h, (x, y)) \stackrel{\text{def}}{=} (h(x) - y)^2$$

**定义2.3 广义损失函数下的不可知PAC可学习** 对于集合 $\mathcal{Z}$ 和损失函数 $\ell: \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}_+$ ，若存在一个函数 $m_{\mathcal{H}}: (0, 1)^2 \rightarrow \mathbb{N}$ 和一个学习算法，使得对于给定的 $\epsilon, \delta \in (0, 1)$ 和任一分布 $\mathcal{D}$ ，当样本数量 $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ 时，算法将以不小于 $1 - \delta$ 的概率返回一个使

$$\mathcal{L}_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h') + \epsilon$$

成立的假设 $h$ ，其中 $\mathcal{L}_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$ 。

## 2.5 思路总结

本章主要在于泛化PAC学习的试用范围，通过更改假设来让这个理论更具有普适性，更贴近生活现实。（通俗来说就是通过不断的容忍，放缩范围，看来搞统计的都是受啊XD）

### 3 Learning via Uniform Convergence

#### 3.1 学习的一致收敛性

我们希望我们所得到的假设集合 $\mathcal{H}$ 不仅仅只在训练集 $\mathcal{S}$ 上表现良好，同样希望它能在真实情况也有相似的表现，既不是好一些也不是差一些，是具有相近的性能，即

$$\forall h \in \mathcal{H}, \mathcal{L}_{\mathcal{S}}(h) \approx \mathcal{L}_{\mathcal{D}}(h)$$

当我们的学习到的假设集合 $\mathcal{H}$ 能达到上述期望时，我们称该学习到的假设集合 $\mathcal{H}$ 具有一致收敛性，同时，当经验误差和泛化误差的差异小于我们容忍限度 $\epsilon$ 时，我们称这时的训练样本为 $\epsilon$ 代表性的。

**定义3.1  $\epsilon$ 代表性样本** 当训练样本满足以下不等式时

$$\forall h \in \mathcal{H}, |\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| \leq \epsilon$$

这时训练样本 $\mathcal{S}$ 被称为 $\epsilon$ 代表性样本。

**引理3.1** 假设一个训练集 $\mathcal{S}$ 是 $\epsilon/2$ 代表性的，那么任何一个根据ERM准则输出的假设集合 $\mathcal{H}$ ，都满足

$$\mathcal{L}_{\mathcal{D}}(h_{\mathcal{S}}) \leq \min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h) + \epsilon$$

证明如下：对于 $\forall h \in \mathcal{H}$

$$\mathcal{L}_{\mathcal{D}}(h_{\mathcal{S}}) \leq \mathcal{L}_{\mathcal{S}}(h_{\mathcal{S}}) + \frac{\epsilon}{2} \leq \mathcal{L}_{\mathcal{S}}(h) + \frac{\epsilon}{2} \leq \mathcal{L}_{\mathcal{D}}(h) + \epsilon$$

其中第一和第三个不等式由 $\epsilon/2$ 代表性的样本保证，第二个不等式是对假设 $h$ 损失的放缩。

**定义3.2 一致收敛** 存在一个函数 $m_{\mathcal{H}}^{UC} : (0, 1)^2 \rightarrow \mathbb{N}$ ，使得对于所有 $\epsilon, \delta \in (0, 1)$ 和任意概率分布 $\mathcal{D}$ ，都有当 $m \geq m_{\mathcal{H}}^{UC}$ 时，至少在 $1 - \delta$ 的概率下，训练集 $\mathcal{S}$ 是 $\epsilon$ 代表性的，假设集合 $\mathcal{H}$ 具有一致收敛性。

#### 3.2 有限假设是不可知PAC可学习的

给定 $\epsilon$ 和 $\delta$ ，我们需要找一个样本大小 $m$ 保证：对于任何分布 $\mathcal{D}$ ，至少在 $1 - \delta$ 的概率下，从 $\mathcal{D}$ 中采样得到的独立同分布样本 $\mathcal{S} = (z_1, \dots, z_m)$ ，对于 $\forall h \in \mathcal{H}$ ，有 $|\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| \leq \epsilon$ 成立，即

$$\mathcal{D}^m(\{\mathcal{S} : \forall h \in \mathcal{H}, |\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1 - \delta$$

根据之前的推导逐步放缩

$$\begin{aligned} \{\mathcal{S} : \forall h \in \mathcal{H}, |\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| > \epsilon\} &= \bigcup_{h \in \mathcal{H}} \{\mathcal{S} : |\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| > \epsilon\} \\ \mathcal{D}^m(\{\mathcal{S} : \forall h \in \mathcal{H}, |\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| > \epsilon\}) &\leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{\mathcal{S} : \forall h \in \mathcal{H}, |\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| > \epsilon\}) \end{aligned}$$

根据大数定律，样本均值会随着样本数量增加，逐渐收敛至总体均值，但是大数定律并没有明确说明有多接近，只是说明了收敛的趋势，所以我们用Hoeffding Inequality来度量，令 $\theta_i$ 为随机变量 $\ell(h, z_i)$ ，假定 $\ell \in [0, 1]$ ，有

$$\mathcal{D}^m(\{\mathcal{S} : \forall h \in \mathcal{H}, |\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| > \epsilon\}) = \mathbb{P}\left[\left|\frac{1}{m} \sum_{i=1}^m \theta_i - \mu\right| > \epsilon\right] \leq 2 \exp(-2m\epsilon^2)$$

所以

$$\begin{aligned} \mathcal{D}^m(\{\mathcal{S} : \forall h \in \mathcal{H}, |\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| > \epsilon\}) &\leq \sum_{h \in \mathcal{H}} 2 \exp(-2m\epsilon^2) \\ &= 2|\mathcal{H}| 2 \exp(-2m\epsilon^2) \end{aligned}$$

当我们令

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

时就会有

$$\mathcal{D}^m(\{\mathcal{S} : \forall h \in \mathcal{H}, |\mathcal{L}_{\mathcal{S}}(h) - \mathcal{L}_{\mathcal{D}}(h)| > \epsilon\}) \leq \delta$$

**推论3.1**  $\mathcal{H}$ 具有一致收敛性，则其样本复杂度函数是

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$$

## 4 The Bias-Complexity Tradeoff

## 5 附录1: 不等式证明

### 5.1 Markov Inequality

马尔科夫不等式:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

其中 $X$ 是非负随机变量。

证:

$$\begin{aligned} \mathbb{E}(X) &= \int_0^{+\infty} xf(x)dx \\ &= \int_0^a xf(x)dx + \int_a^{+\infty} xf(x)dx \\ &\leq \int_0^a 0f(x)dx + \int_a^{+\infty} af(x)dx \\ &\leq 0 + \int_a^{+\infty} af(x)dx = a \int_a^{+\infty} f(x)dx = a\mathbb{P}(X > a) \end{aligned}$$

移动一下 $a$ 的位置, 不等式得证, 其中第二行到第三行是将积分中的 $x$ 换成积分下限。

### 5.2 引理1

设 $Z$ 是一个取值 $[0, 1]$ 的随机变量, 假定 $\mathbb{E}[Z] = \mu$ , 那么对于任意 $a \in (0, 1)$ , 都有

$$\mathbb{P}[Z > 1 - a] \geq \frac{\mu - (1 - a)}{a}$$

证: 令 $Y = 1 - Z$ , 则 $Y$ 是非负随机变量, 且 $\mathbb{E}[Y] = 1 - \mu$ , 根据马尔科夫不等式

$$\mathbb{P}[Z \leq 1 - a] = \mathbb{P}[1 - Z \geq a] = \mathbb{P}[Y \geq a] \leq \frac{\mathbb{E}[Y]}{a} = \frac{1 - \mu}{a}$$

所以

$$\mathbb{P}[Z > 1 - a] \geq 1 - \frac{1 - \mu}{a} = \frac{a + \mu - 1}{a}$$

### 5.3 Chebyshev Inequality

切比雪夫不等式:

$$\forall a > 0, \mathbb{P}[|Z - \mathbb{E}[Z]| \geq a] = \mathbb{P}[(Z - \mathbb{E}[Z])^2 \geq a^2] \leq \frac{\text{Var}[Z]}{a^2}$$



其中  $Var[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$  是  $Z$  的方差。

证明: 令  $Y = (Z - \mathbb{E}[Z])^2$  带入马尔科夫不等式即可得证。

#### 5.4 引理2

设  $Z_1, \dots, Z_m$  是独立同分布的随机变量, 假定  $\mathbb{E}[Z] = \mu$  且  $Var[Z] \leq 1$ , 那么对于任意的  $\delta \in (0, 1)$  有

$$|\frac{1}{m} \sum_{i=1}^m Z_i - \mu| \leq \sqrt{\frac{1}{\delta m}}$$

成立的概率大于  $1 - \delta$ 。

证: 由切比雪夫不等式, 对于  $a > 0$ , 我们有

$$\mathbb{P}[|\frac{1}{m} \sum_{i=1}^m Z_i - \mu| > a] \leq \frac{Var[Z_1]}{ma^2} \leq \frac{1}{ma^2}$$

令等式右边等于  $\delta$  即可得证。

#### 5.5 Chernoff Bound

切尔诺夫界: 假设  $Z_1, \dots, Z_m$  是独立的伯努利变量, 其中任意的  $i$  都有  $\mathbb{P}[Z_i = 1] = p_i$ 。令  $p = \sum_{i=1}^m p_i$  和  $Z = \sum_{i=1}^m Z_i$ , 对于  $t > 0$  有

$$\mathbb{P}[Z > (1 + \delta)p] = \mathbb{P}[e^{tZ} > e^{t(1+\delta)p}] \leq \frac{\mathbb{E}[e^{tZ}]}{e^{t(1+\delta)p}}$$

其中, 第一个等号成立是因为  $e^x$  单调递增, 之后的不等关系成立是因为马尔科夫不等式。

### 5.6 引理3

根据之前的切尔诺夫界, 对于 $\mathbb{E}[e^{tZ}]$ 有

$$\begin{aligned}
 \mathbb{E}[e^{tZ}] &= \mathbb{E}[e^{t \sum_i Z_i}] = \mathbb{E}[\prod_i e^{tZ_i}] \\
 &= \prod_i \mathbb{E}[e^{tZ_i}] \\
 &= \prod_i (p_i e^{(1 \times t)} + (1 - p_i) e^{0 \times t}) \\
 &= \prod_i (1 + p_i(e^t - 1)) \\
 &\leq \prod_i e^{p_i(e^t - 1)} \\
 &= e^{\sum_i p_i(e^t - 1)} \\
 &= e^{p(e^t - 1)}
 \end{aligned}$$

其中不等式成立因为 $e^x \approx 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}$ , 根据不同情况更改 $t$ 的值可以得到不同的概率上界。

取 $t = \log(1 + \delta)$ , 则

$$\mathbb{P}[Z > (1 + \delta)p] \leq e^{-h(\delta)p}$$

其中

$$h(\delta) = (1 + \delta) \log(1 + \delta) - \delta$$

根据 $h(\delta) \geq \frac{\delta^2}{(2+2\delta/3)}$ 有

$$\mathbb{P}[Z > (1 + \delta)p] \leq e^{-p \frac{\delta^2}{2+2\delta/3}}$$

取 $t = -\log(1 - \delta)$ 则有

$$\mathbb{P}[Z < (1 - \delta)p] \leq \frac{e^{-\delta p}}{e^{(1-\delta) \log(1-\delta)p}} = e^{-ph(-\delta)}$$

根据 $h(-\delta) \geq h(\delta)$ 有

$$\mathbb{P}[Z < (1 - \delta)p] \leq e^{-ph(\delta)} \leq e^{ph(\delta)} \leq e^{-p \frac{\delta^2}{2+2\delta/3}}$$

### 5.7 Hoeffding Inequality

霍夫丁不定式: 假设  $Z_1, \dots, Z_m$  是独立同分布的随机变量, 令  $\bar{Z} = \frac{1}{m} \sum_{i=1}^m Z_i$ , 假定  $\mathbb{E}[\bar{Z}] = \mu$  且  $\mathbb{P}[a \leq Z_i \leq b] = 1$  对于所有  $i$  成立, 那么对于任意  $\epsilon$  有

$$\mathbb{P}\left[\left|\frac{1}{m} \sum_{i=1}^m Z_i - \mu\right| > \epsilon\right] \leq 2e^{(-\frac{2m\epsilon^2}{(b-a)^2})}$$

证: 记  $X_i = Z_i - \mathbb{E}[Z_i]$  且  $\bar{X}$ , 且  $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i$ , 对于任意的  $\lambda, \epsilon > 0$  有

$$\mathbb{P}[\bar{X} - \epsilon] = \mathbb{P}[e^{-\lambda\bar{X}} - e^{-\lambda\epsilon}] \leq e^{-\lambda\epsilon} \mathbb{E}[e^{-\lambda\bar{X}}]$$

其中

$$\mathbb{E}[e^{-\lambda\bar{X}}] = \mathbb{E}\left[\prod_i e^{-\lambda X_i/m}\right] = \prod_i \mathbb{E}[e^{-\lambda X_i/m}]$$

对于凸函数  $f(x) = e^x$ , 对于任意  $\alpha \in [0, 1]$  和  $x \in [a, b]$  都有满足

$$f(x) \leq \alpha f(a) + (1 - \alpha)f(b)$$

令  $\alpha = \frac{b-x}{b-a} \in [0, 1]$ , 则

$$e^{\lambda x} \leq \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b}$$

当  $\mathbb{E}[X] = 0$  时, 对两边同时取期望

$$\mathbb{E}[e^{\lambda x}] \leq \frac{b - \mathbb{E}[X]}{b-a} e^{\lambda a} + \frac{\mathbb{E}[X] - a}{b-a} e^{\lambda b} = \frac{b}{b-1} e^{\lambda a} - \frac{b}{b-1} e^{\lambda b}$$

记  $h = \lambda(b-a)$  和  $p = \frac{-a}{b-a}$ , 可以将上式右边写作  $e^{-hp + \log(1-p+pe^h)}$ , 记为  $e^{L(h)}$ 。

我们将  $L(h)$  作泰勒展开

$$L(h) = \frac{L(0)}{0!}(x-0)^0 + \frac{L'(0)}{1!}(x-0)^1 + \frac{L''(0)}{2!}(x-0)^2 + o(x^2)$$

易得  $L(0) = L'(0) = 0$ , 而

$$L''(h) = \frac{(1-p)pe^h}{(1-p+pe^h)^2}$$

得  $L''(0) = (1-p)p \leq 1/4$ , 将其放缩后得到

$$L(h) = \frac{L''(0)}{2!}(x-0)^2 + o(x^2) \leq \frac{1/4}{2!}(x-0)^2 = \frac{h^2}{8}$$

将其回带, 对于任意 $i$ 有

$$\mathbb{E}[e^{\lambda X_i/m}] \leq e^{\frac{\lambda^2(b-a)^2}{8m^2}}$$

因此

$$\mathbb{E}[\bar{X} \geq \epsilon] \leq e^{-\lambda\epsilon} \prod_i e^{\frac{\lambda^2(b-a)^2}{8m^2}} = e^{-\lambda\epsilon + \frac{\lambda^2(b-a)^2}{8m}}$$

此时我们令 $\lambda = 4m\epsilon/(b-a)^2$

$$\mathbb{E}[\bar{X} \geq \epsilon] \leq e^{-4m\epsilon/(b-a)^2\epsilon + \frac{(4m\epsilon/(b-a)^2)^2(b-a)^2}{8m}}$$

$$\mathbb{E}[\bar{X} \geq \epsilon] \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

将 $\bar{X}_i = Z_i - \mathbb{E}[Z_i]$ 带回, 不等式得证。