第 *1* 章　概率近似正确

Siheng Zhang
zhangsiheng@cvte.com

2021 年 5 月 24 日

本章对应 UML 第 2-5 章，主要回答如下问题：

- 关于泛化误差，我们能知道什么？
- 当我们谈论归纳偏置的时候，我们在谈论什么？

目录

1 学习器形式化：输入、输出与评价

- **输入**：即训练集，为一个有限的序列 $S = ((x_1, y_1), \cdots, (x_m, y_m))$。训练集为实例集和标签集的笛卡尔积，其中实例 $x \in \mathcal{X}$，标签 $y \in \mathcal{Y}$。本章主要考虑二分类问题，即 $\mathcal{Y} = \{0, 1\}$。
- **输出**：假设（hypothesis），有时候称为分类器（classifier）、回归器（regressor）$h : \mathcal{X} \to \mathcal{Y}$。
- **数据生成模型**：假设实例依概率分布 $\mathcal{D}$ 产生，并且有一个“绝对正确”的标签生成函数（至少目前我们假设如此）$f : \mathcal{X} \to \mathcal{Y}$。

  独立同分布（i.i.d）假设：训练样本的产生独立地来自于同一个分布。
- **泛化误差**：也称为真实误差或者真实风险。以 $\mathcal{D}(A)$ 表示样例 $x \in A$ 的概率，则泛化误差为：

$$L_{\mathcal{D},f}(h) \overset{def}{=} \mathop{\mathbb{P}}_{x\sim\mathcal{D}}[h(x) \neq f(x)] \overset{def}{=} \mathcal{D}(\{x : h(x) \neq f(x)\}) \tag{1}$$

注 1. 数据生成模型对学习器不可见。
注 2. 通常我们说“训练集”，但是严格来说应该说“训练序列”，因为同样的样本可以多次出现，并且一些学习算法的输出因输入顺序不同而不同。
注 3. 严格来说，$\mathcal{D}$ 定义在 $\mathcal{X} \times \mathcal{Y}$ 上，但通常不严格区分。

2 从最小经验风险到概率近似正确

2.1 最小经验风险（Empirical Risk Minimization，ERM）准则可能造成过拟合

显然，泛化误差是不可知的。因此，我们转而最小化经验风险：

$$L_S(h) \overset{def}{=} \frac{|\{(x_i, y_i) \in S : h(x_i) \neq y_i\}|}{m} \tag{2}$$

考虑一个“懒惰”的学习器 h，它记住了所有的样本，当且仅当 $x = x_i$ 时候输出 $y = y_i$，其它时候都输出 0。显然，它的经验风险 $L_S(h) = 0$，但是对于未见样本，它有一半的概率预测失败，$L_{\mathcal{D},f}(h) = 1/2$。这种在训练集上表现特别好，但是泛化性能很差的现象，称为“过拟合”。这个例子蕴含着一个很重要的基本事实：

如果不对假设集加以限制，最小经验风险准则可能造成过拟合。

2.2 归纳偏置（inductive bias）下的最小经验风险准则

选取假设集即反映了人对任务的先验知识，即归纳偏置。记假设集为 $\mathcal{H}$，最小经验风险准则表述为：

$$h_S \in \arg\min_{h\in\mathcal{H}} L_S(h) \tag{3}$$

“理想”的情况下，假设集中包含有泛化误差为 0 的假设，即存在 $h^* \in \mathcal{H}$ 使得 $L_{\mathcal{D},f}(h^*) = 0$，我们称之为**可实现假设**。该条件蕴含着 $L_S(h^*) = 0, L_S(h_S) = 0$，而事实上，我们更感兴趣的是 $L_{\mathcal{D},f}(h_S)$。

2.3 概率近似正确（Probably Approximately Correct，PAC）可学习性

定义：称假设集是概率近似正确可学习的，当给定大小为 $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ 的样本集，假设集中存在一个假设以至少 $1 - \delta$ 的置信度达到不低于 $1 - \epsilon$ 的正确率，即 $P(L_{\mathcal{D},f}(h_S) < \epsilon) > 1 - \delta$。

Theorem 1 *有限假设集是 PAC 可学习的，且样本复杂度为* $m_{\mathcal{H}}(\epsilon, \delta) = \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$.

Proof *在假设集中存在着一些糟糕的假设，它们的泛化误差高于一定阈值。形式化来说，记糟糕的假设集为* $\mathcal{H}_B$，*它是* $\mathcal{H}$ *的子集，并且* $\forall h \in \mathcal{H}_B, L_{\mathcal{D},f}(h) > \epsilon$。*而对于某些数据集，这样的假设仍然能有良好的训练性*

能，记为 $M=\{S:\exists h\in\mathcal{H}_B, L_S(h)=0\}$。注意到，$M=\bigcup\limits_{h\in\mathcal{H}_B}\{S:L_S(h)=0\}$。而我们的证明目标是事件 $L_{\mathcal{D},f}(h_S)>\epsilon$ 的概率上界，即

$$\begin{aligned}&\mathcal{D}^m(\{S:L_{\mathcal{D},f}(h_S)>\epsilon\})\le\mathcal{D}^m(M)=\mathcal{D}^m(\bigcup_{h\in\mathcal{H}_B}\{S:L_S(h)=0\})\\&=\sum_{h\in\mathcal{H}_B}\prod_{i=1}^{m}\mathcal{D}(\{x_i:f(x_i)=h(x_i)\})\overset{i.i.d.}{=}\sum_{h\in\mathcal{H}_B}(1-L_{\mathcal{D},f}(h))^m\\&\le\sum_{h\in\mathcal{H}_B}(1-\epsilon)^m\le\sum_{h\in\mathcal{H}_B}\exp(-\epsilon m)\le|\mathcal{H}|\exp(-\epsilon m)\end{aligned}$$

令 $|\mathcal{H}|\exp(-\epsilon m)\le\delta$，可以得到 $m\ge\log(|\mathcal{H}|/\delta)/\epsilon$。

2.4 “没有免费午餐”（No-Free-Lunch，NFL）定理

Theorem 2 令 A 表示 *0-1* 损失下二分类任务的任一学习算法，假设训练集规模为 $m<|\mathcal{X}|/2$。则存在一个定义在 $X\times\{0,1\}$ 上的分布 $\mathcal{D}$ 使得：随机采样训练集 $S\sim\mathcal{D}^m$，$P(L_{\mathcal{D}}(A(S))\ge1/8)\ge1/7$。

Proof 令 $C\subseteq\mathcal{X}$ 含有 $2m$ 个样本，则标签函数存在 2^{2m} 种可能，记为 $f_1,\cdots,f_T$。对任一标签函数，定义在 $\mathcal{C}\times\{0,1\}$ 上的分布 $\mathcal{D}_i$:

$$\mathcal{D}_i(\{(x,y)\})=\begin{cases}1/|C|, & \text{if } y=f_i(x)\\0, & \text{otherwise}\end{cases}$$

从 C 中获取 m 个样本构成数据集，共有 $k=(2m)^m$ 个可能的数据集。记由函数 f_i 打标签的数据集为 $\mathcal{S}_j^i, j=1,2,\cdots,k$，因此

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))]=\frac{1}{k}\sum_{j=1}^{k}L_{\mathcal{D}_i}(A(S_j^i))$$

注意到

$$\max_{i\in\{1,\cdots,T\}}\frac{1}{k}\sum_{j=1}^{k}L_{\mathcal{D}_i}(A(S_j^i))\ge\frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k}L_{\mathcal{D}_i}(A(S_j^i))\ge\min_{j\in\{1,\cdots,k\}}\frac{1}{T}\sum_{i=1}^{T}L_{\mathcal{D}_i}(A(S_j^i))$$

记 $S_j=(x_1,\cdots,x_m)$，同时记 $v_1,\cdots,v_p$ 为 C 中未出现在 S_j 的样本。显然，$p\ge m$（当 S_j 中存在重复样本的时候，不等号严格成立）。因此，对任意 $h:C\to\{0,1\}$，

$$L_{\mathcal{D}_i}(h)=\frac{1}{2m}\sum_{x\in C}\mathbb{1}_{[h(x)\neq f_i(x)]}\ge\frac{1}{2p}\sum_{r=1}^{p}\mathbb{1}_{[h(v_r)\neq f_i(v_r)]}$$

$$\frac{1}{T}\sum_{i=1}^{T}L_{\mathcal{D}_i}(A(S_j^i))\ge\frac{1}{2}\min_{r\in\{1,\cdots,p\}}\frac{1}{T}\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]}$$

其中 $r\in[p]$。

将标签函数 $f_1,\cdots,f_T$ 切分成 $T/2$ 对不相交的函数对，对于其中的某对函数 $(f_i,f_{i'})$，有 $\forall c\in C, f_i(c)\neq f_{i'}(c)$ 当且仅当 $c=v_r$。则它们对数据集 S_j 给出了同样的标签，因而 $S_j^i=S_j^{i'}$。从而有，$\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]}+\mathbb{1}_{[A(S_j^{i'})(v_r)\neq f_{i'}(v_r)]}=1$。这意味着

$$\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]}=\frac{1}{2}$$

至此，

$$\max_{i\in\{1,\cdots,T\}} \mathop{\mathbb{E}}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] \geq 1/4$$

这说明了，存在某些标签函数和分布 $f, \mathcal{D}$ 使得 $L_{\mathcal{D}}(f) = 0$，且 $\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq 1/4$。而后者可以进一步推出违背 *PAC* 可学习性的结论。注意到，对一个随机变量 $\theta \in [0,1]$，如果满足 $\mathbb{E}(\theta) \geq 1/4$，则有：

$$p\left(\theta \geq \frac{1}{8}\right) = \int_{\frac{1}{8}}^{1} p(\theta)\mathrm{d}\theta \geq \int_{\frac{1}{8}}^{1} \theta p(\theta)\mathrm{d}\theta = \mathbb{E}(\theta) - \int_{0}^{\frac{1}{8}} \theta p(\theta)\mathrm{d}\theta \geq \frac{1}{4} - \frac{1}{8}\int_{0}^{\frac{1}{8}} p(\theta)\mathrm{d}\theta \Rightarrow p\left(\theta \geq \frac{1}{8}\right) \geq \frac{1}{7}$$

NFL 定理从数学形式上确认了归纳偏置的必要性。

3 不可知情况下的 PAC（Agnostic PAC, A-PAC）可学习性

在实际情况中，也许不存在所谓正确的标签函数，标签也不一定可以由手头的特征决定。因此可实现假设是不成立的，称之为不可知情况。此时，PAC 可学习性需要重新定义。称假设集是 A-PAC 可学习的，当在规模为 $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ 的训练集上训练时，存在一个算法，以至少 $1-\delta$ 的置信度成立如下条件：

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

其中 $L_{\mathcal{D}}(h) \overset{def}{=} \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \overset{def}{=} \mathcal{D}(\{x : h(x) \neq y\})$.

此外，损失函数也可以拓展到一般情况，仅需将二分类下单个样本的损失函数拓展到一般度量，$L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[l((x), y)]$。

定义 (ϵ-典型性)：在定义域 Z，假设集 $\mathcal{H}$，损失函数 l 和分布 $\mathcal{D}$ 给定的前提下，称训练集 S 具有 ϵ-典型性，如果其满足，$\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$。

Lemma 1 *假设训练集 S 具有 $\epsilon/2$-典型性，则任一 $h_S \in \arg\min_{h\in\mathcal{H}} L_S(h)$ 都满足 $L_{\mathcal{D}}(h_S) \leq L_{\mathcal{D}}(h) + \epsilon$。*

Proof *对任一 $h \in \mathcal{H}$，$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \epsilon$。*

该引理说明了，ERM 规则在 $\epsilon/2$-典型的数据集上能够成功返回好的假设。于是，要使得 ERM 规则是否适用于 A-PAC，就只要保证随机选取一个训练集，其为 ϵ-典型的概率至少为 $1-\delta$。可想而知，决定随机选取的数据集是否足够典型的一个先决条件是其大小。下面的定义对其所需的大小进行了刻画。

定义 (一致收敛性)：称假设集 $\mathcal{H}$ 具有一致收敛性，如果存在函数 $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$，使得对任意 $\epsilon, \delta \in (0,1)$ 和定义在 Z 上的分布 $\mathcal{D}$，数据集 S 的样本独立同分布地采样自 $\mathcal{D}$，S 的大小满足 $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$，则 S 为 ϵ-典型的概率至少为 $1-\delta$。

Corollary 1 *如果一个假设集 $\mathcal{H}$ 在函数 $m_{\mathcal{H}}^{UC}$ 下具有一致收敛性，则其在训练集大小满足 $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$ 条件下 A-PAC 可学习，且 $ERM_{\mathcal{H}}$ 策略是 $\mathcal{H}$ 成功的 A-PAC 学习器.*

Lemma 2 *(Hoeffding 不等式) 令 $\theta_1, \cdots, \theta_m$ 为一序列独立同分布的随机变量，假设 $\forall i, \mathbb{E}[\theta_i] = \mu$ 且 $P[a \leq \theta_i \leq b] = 1$，则对任一 $\epsilon > 0$，*

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

Theorem 3 ***A-PAC 样本复杂度*** *假设损失函数的取值范围为 $[a,b]$，则满足有限假设集 $\mathcal{H}$ 是 A-PAC 可学习的样本复杂度为*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2} \right\rceil \tag{4}$$

Proof 固定 ϵ, δ，目的是找到样本数量 m 来保证：对任一分布 $\mathcal{D}$，样本集 $S \overset{i.i.d.}{\sim} \mathcal{D}^m$，使得

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta$$

式子左侧可以放缩为：

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$$

由于样本独立地采样自 $\mathcal{D}$，由期望地线性性，$\mathbb{E}(L_S(h)) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}(l(h(x_i), y_i)) = L_{\mathcal{D}}(h)$。因此，$|L_{\mathcal{D}}(h) - L_S(h)|$ 可以视为随机变量 $L_S(h)$ 和其期望之间的距离。应用 *Hoeffding* 不等式可以得到，

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq 2 \exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

结合上述方程可以得到

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} 2 \exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right) = 2|\mathcal{H}| \exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right)$$

选取 $m \geq \frac{2\log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2}$，则方程左侧的上界为 δ。

4 误差分解

$$L_{\mathcal{D}}(h_S) = \underbrace{\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)}_{\epsilon_{\text{app}}} + \underbrace{L_{\mathcal{D}}(h_S) - \epsilon_{\text{app}}}_{\epsilon_{\text{est}}} \tag{5}$$

- **逼近误差**衡量了选取当前假设集（即归纳偏置）的风险，注意其与数据集规模无关。扩大假设集的规模可以减少逼近误差。
- **拟合误差**衡量了经验风险（即训练误差），是真实风险的一个经验估计。拟合误差取决于训练集的大小（随其增大而减小）与假设集的规模（随其增大而对数上升）。

5 总结

注. 在本系列笔记的整理过程中，这一小节的出现意味着，这章枯燥且乏味，但这章的偏理论的内容是不需要记住的，因为这些推导所指向的下述结论是非常浅显的。小结最后还会写到，这章的内容将如何与其它章节前后关联，以使读者明白其在整个机器学习理论中所处的地位。如果某些章没有这一小节，那么说明该章节的内容浅显易懂，不需要这一小节。）

至此，在 PAC 学习理论框架下，我们有了如下重要结论：

1. 不存在对一切问题通用的学习器；
2. 归纳偏置对于防止过拟合是必要的;
3. 样本复杂度函数跟假设集规模、置信度和误差水平有关，但有趣的是，它和样本空间并无关联;
4. 归纳偏置控制着拟合误差和逼近误差的折衷。

现在，我们触及了学习理论的核心问题：**在什么样的假设集下，ERM 规则不会过拟合（即 PAC 可学习）？**当前，我们仅保证了在有限假设集上的性质。下一章谈及 VC 维时候，这个问题才会有一个更加精确的答案。

6 练习与答案

Ex1 (UML Ex2.2) Let $\mathcal{H}$ be a class of binary classifiers over a domain $\mathcal{X}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$, and let f be the target hypothesis in $\mathcal{H}$. Fix some $h \in \mathcal{H}$, show that the expected value of $L_S(h)$ over the choice of S equals $L_{\mathcal{D},f}(h)$, namely,

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_S(h)] = L_{\mathcal{D},f}(h)$$

Solution: By the linearity of expectation,

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_S(h)] &= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{[h(x_i)\neq f(x_i)]}] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{x_i\sim\mathcal{D}^m}[\mathbb{1}_{[h(x_i)\neq f(x_i)]}] \\
&= \frac{1}{m}\cdot m \cdot L_{\mathcal{D},f}(h) = L_{\mathcal{D},f}(h)
\end{aligned}$$

Ex2 (UML Ex2.3) Axis Aligned rectangles: An axis aligned rectangle classifier in the plane is a classifier that assigns the value 1 to a point if and only if it is inside a certain rectangle. Formally, given real numbers $a_1 \leq b_1, a_2 \leq b_2$, define the classifier $h(a_1, b_1, a_2, b_2)$ by:

$$h(a_1,b_1,a_2,b_2)(x_1,x_2) = \begin{cases} 1, & \text{if } a_1 \leq x_1 \leq b1 \text{ and } a_2 \leq x_2 \leq b_2 \\ 0, & \text{otherwise} \end{cases}$$

The class of all axis aligned rectangles in the plane is defined as:

$$\mathcal{H}^2_{\text{rec}} = \{h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1, \text{ and } a_2 \leq b_2\}$$

Note that this is an infinite size hypothesis class. Throughout this exercise we rely on the realizability assumption.

2.1 Let A be the algorithm that returns the smallest rectangle enclosing all positive examples in the training set. Show that A is an ERM.

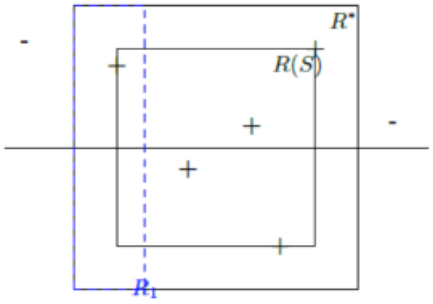
2.2 Show that if A receives a training set of size $\geq 4\frac{\log(4/\delta)}{\epsilon}$, then, with probability of at least $1-\delta$ it returns a hypothesis with error of at most ϵ.

Hint: Fix some distribution $\mathcal{D}$ over $\mathcal{X}$, let $R^* = R(a_1^*, b_1^*, a_2^*, b_2^*)$ be the rectangle that generates the labels, and let f be the corresponding hypothesis. Let $a_1 \geq a_1^*$ be a number such that the probability mass (with respect to $\mathcal{D}$) of the rectangle $R_1 = R(a_1^*, a_1, a_2^*, b_2^*)$ is exactly $\epsilon/4$. Similarly, let b_1, a_2, b_2 be numbers such that the probability masses of the rectangles $R_2 = R(b_1, b_1^*, a_2^*, b_2^*)$, $R_3 = R(a_1^*, b_1^*, a_2^*, a_2)$, $R_4 = R(a_1^*, b_1^*, b_2, b_2^*)$ are all $\epsilon/4$. Let $R(S)$ be the rectangle returned by A. See illustration in the following figure.

- Show that $R(S) \subseteq R^*$.
- Show that if S contains (positive) examples in all of the rectangles R_1, R_2, R_3, R_4, then the hypothesis returned by A has error of at most ϵ.
- For each $i \in \{1, \cdots, 4\}$, upper bound the probability that S does not contain an example from R_i.
- Use the union bound to conclude the argument.

2.3 Repeat the previous question for the class of axis aligned rectangles in $\mathbb{R}^d$.

2.4 Show that the runtime of applying the algorithm A mentioned earlier is polynomial in $d, 1/\epsilon$, and in $\log(1/\delta)$.



Solution

2.1 In realizable setup, since the tightest rectangle enclosing all positive example is returned, all positive and negative instances are correctly classified.

2.2 By definition, algorithm A returns the tightest rectangle, so $R(S) \subseteq R^*$.

Ex3 (UML Ex3.2) Let $\mathcal{X}$ be a discrete domain, and let $\mathcal{H}_{\text{Singleton}} = \{h_z : z \in \mathcal{X}\} \cup \{h^-\}$, where for each $z \in \mathcal{X}$, h_z is the function defined by $h_z(x) = 1$ if $x = z$ and $h_z(x) = 0$ if $x \neq z$. h^- is simply the all-negative hypothesis, namely, $\forall x \in \mathcal{X}, h^-(x) = 0$. The realizability assumption here implies that the true hypothesis f labels negatively all examples in the domain, perhaps except one.

3.1 Describe an algorithm that implements the ERM rule for learning $\mathcal{H}_{\text{Singleton}}$ in the realizable setup.

3.2 Show that $\mathcal{H}_{\text{Singleton}}$ is PAC learnable. Provide an upper bound on the sample complexity.

Solution:

3.1 Traverse $z \in \mathcal{X}$ then output h_z or h^-.

3.2 If for any $i \in [1, \cdots, m]$, h_{x_i} is the true hypothesis, the algorithm can find it in the realizable setup. Otherwise, the algorithm outputs h^-, which can be either true or false (i.e., the target z^* is not in the training set). Note that in the second case, the algorithm only makes a single error when generalize to all cases, and hence $p(z^*) \geq \epsilon$ (otherwise, it is meaningless),

$$\mathbb{P}(L_{\mathcal{D},f}(h_S) > \epsilon) \leq (1 - p(z^*))^m \leq (1-\epsilon)^m \leq \exp(-\epsilon m) \leq \delta$$

which leads to

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$$

Ex4 (UML Ex3.3) Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}$ be the class of concentric circles in the plane, that is, $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$, where $h_r(x) = \mathbb{1}_{[\|x\| \leq r]}$. Prove that $\mathcal{H}$ is PAC learnable (assume realizability), and its sample complexity is bounded by:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$$

Solution Denote the probability of x such that $r \leq \|x\|_2 \leq r^*$ is ϵ, then:

$$P(L_{\mathcal{D}}(h_r(S)) \geq \epsilon) \leq (1-\epsilon)^m \leq e^(- m\epsilon)$$

Bound it by confidence δ leads to the conclusion.

Ex5 (UML Ex3.4) In this question, we study the hypothesis class of Boolean conjunctions defined as follows. The instance space is $\mathcal{X} = \{0,1\}^d$ and the label set is $\mathcal{Y} = \{0,1\}$. A literal over the variables $x_1, \cdots, x_d$ is a simple Boolean function that takes the form $f(\mathbf{x}) = x_i$, for some $i \in [d]$, or $f(\mathbf{x}) = 1 - x_i$ for some $i \in [d]$. We use the notation $\overline{x}_i$ as a shorthand for $1 - x_i$. A conjunction is any product of literals. In Boolean logic, the product is denoted using the $\wedge$ sign. For example, the function $h(\mathbf{x}) = x_1 \cdot (1 - x_2)$ is written as $x_1 \wedge \overline{x}_2$. We consider the hypothesis class of all conjunctions of literals over the d variables. The empty conjunction is interpreted as the all-positive hypothesis (namely, the function that returns $h(\mathbf{x}) = 1$ for all $\mathbf{x}$). The conjunction $x_1 \wedge \tilde{x}_1$ (and similarly any conjunction involving a literal and its negation) is allowed and interpreted as the all-negative hypothesis (namely, the conjunction that returns $h(\mathbf{x}) = 0$ for all $\mathbf{x}$). We assume realizability: Namely, we assume that there exists a Boolean conjunction that generates the labels. Thus, each example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ consists of an assignment to the d Boolean variables $x_1, \cdots, x_d$, and its truth value (0 for false and 1 for true). For instance, let $d = 3$ and suppose that the true conjunction is $x_1 \wedge \overline{x}_2$. Then, the training set S might contain the following instances:

$$((1,1,1),0), ((1,0,1),1), ((0,1,0),0), ((1,0,0),1)$$

Prove that the hypothesis class of all conjunctions over d variables is PAC learnable and bound its sample complexity. Propose an algorithm that implements the ERM rule, whose runtime is polynomial in $d \cdot m$.

Solution: $\mathcal{H}$ is finite, and hence is PAC learnable. Besides the all-negative conjunction, each hypothesis is determined by deciding for each variable x_i with 3 possible choices (x_i, $\overline{x}_i$ or none). Thus, $\mathcal{H} = 3^d + 1$, and

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{d \log 3 + \log(1/\delta)}{\epsilon} \right\rceil$$

Below is the learning algorithm. Start from $h = x_1 \wedge \overline{x}_1 \wedge \cdots \wedge x_d \wedge \overline{x}_d$, which is always negative. The algorithm does nothing when feed a negative example, and remove x_1 or $\overline{x}_1$ when feed a positive example. With realizability assumption, it can labels all training examples correctly and hence is an ERM. Since the algorithm takes linear time (in terms of the dimension d) to process each example, the running time is bounded by $O(m \times d)$.

Ex6 (UML Ex3.7) The Bayes optimal predictor: Show that for every probability distribution $\mathcal{D}$, the Bayes optimal predictor $f_{\mathcal{D}}$ is optimal, in the sense that for every classifier g from X to $\{0,1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Solution: The Bayes predictor labels a sample x according to

$$f_{\mathcal{D}}(x) = \begin{cases} 0, & \text{if } \mathcal{D}((x,0)) \geq \mathcal{D}((x,1)) \\ 1, & \text{otherwise} \end{cases}$$

When it labels a sample to be class 0, it holds that $\mathcal{D}((x,0)) \geq \mathcal{D}((x,1))$. If the true label function also makes such a decision, then Bayes predictor makes no error. Otherwise, $f(x) = 1$, but its probability is no more than 1/2. Any other classifier that labels x to be class 1 will suffer a risk no less than 1/2. Hence, in total, Bayes predictor is the optimal.

Ex7 (UML Ex3.9) Consider a variant of the PAC model in which there are two example oracles: one that generates positive examples and one that generates negative examples, both according to the underlying distribution $\mathcal{D}$ on $\mathcal{X}$. Formally, given a target function $f : \mathcal{X} \to \{0,1\}$, let $\mathcal{D}^+$ be the distribution over $\mathcal{X}^+ = \{x \in \mathcal{X} : f(x) = 1\}$ defined by $\mathcal{D}^+(A) = \mathcal{D}(A)/\mathcal{D}(\mathcal{X}^+)$, for every $A \in \mathcal{X}^+$. Similarly, $\mathcal{D}^-$ is the distribution over $\mathcal{X}^-$ induced by $\mathcal{D}$.

The definition of PAC learnability in the two-oracle model is the same as the standard definition of PAC learnability except that here the learner has access to $m^+_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples from $\mathcal{D}^+$ and $m^-(\epsilon, \delta)$ i.i.d.

examples from $\mathcal{D}^-$. The learner's goal is to output h s.t. with probability at least $1-\delta$ (over the choice of the two training sets, and possibly over the nondeterministic decisions made by the learning algorithm), both $L(\mathcal{D}^+, f)(h) \leq \epsilon$ and $L(\mathcal{D}^-, f)(h) \leq \epsilon$.

7.1 Show that if $\mathcal{H}$ is PAC learnable in the standard one-oracle model, then H is PAC learnable in the two-oracle model.

7.2 Define h^+ to be the always-plus hypothesis and h^- to be the always minus hypothesis. Assume that $h^+, h^- \in \mathcal{H}$. Show that if $\mathcal{H}$ is PAC learnable in the two-oracle model, then $\mathcal{H}$ is PAC learnable in the standard one-oracle model.

Solution:

7.1 Drawing points from the negative and positive oracles with equal provability is equivalent to obtaining i.i.d. examples from a distribution $\mathcal{D}'$ which gives equal probability to positive and negative examples. If we let an algorithm to access to a training set which is drawn i.i.d. according to the $\mathcal{D}'$ with size $m_{\mathcal{H}}(\epsilon/2, \delta)$, then with probability at least $1-\delta$, it returns h with:

$$\begin{aligned}
\epsilon/2 \geq L_{(\mathcal{D}',f)}(h) &= \mathop{\mathbb{P}}_{x\sim\mathcal{D}'}[h(x) \neq f(x)] \\
&= \mathop{\mathbb{P}}_{x\sim\mathcal{D}'}[f(x)=1, h(x)=0] + \mathop{\mathbb{P}}_{x\sim\mathcal{D}'}[f(x)=0, h(x)=1] \\
&= \mathop{\mathbb{P}}_{x\sim\mathcal{D}'}[f(x)=1] \cdot \mathop{\mathbb{P}}_{x\sim\mathcal{D}^+}[h(x)=0] + \mathop{\mathbb{P}}_{x\sim\mathcal{D}'}[f(x)=0] \cdot \mathop{\mathbb{P}}_{x\sim\mathcal{D}^-}[h(x)=1] \\
&= \frac{1}{2} L_{(\mathcal{D}^+,f)}(h) + \frac{1}{2} L_{(\mathcal{D}^-,f)}(h)
\end{aligned}$$

which implies $L_{(\mathcal{D}^+,f)}(h) \leq \epsilon$ and $L_{(\mathcal{D}^-,f)}(h) \leq \epsilon$.

7.2

Ex8 (UML Ex5.3) Prove that if $|\mathcal{X}| \geq km$ for a positive integer $k \geq 2$, then we can replace the lower bound in the No-Free-Lunch theorem. Namely, for the task of binary classification, there exists a distribution $\mathcal{D} \sim \mathcal{X} \times \{0,1\}$ such that:

8.1 There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.

8.2 $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \frac{1}{2} - \frac{1}{2k}$.

Solution: Only the second proposition should be proved. Similar with the proof in above,

$$L_{\mathcal{D}_i}(h) = \frac{1}{km} \sum_{x\in C} \mathbb{1}_{[h(x)\neq f_i(x)]} \geq \frac{1}{km} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]} \geq \frac{k-1}{pk} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]}$$

And similarity,

$$\frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{k-1}{k} \min_{r\in\{1,\cdots,p\}} \frac{1}{T} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]}$$

So the final bound is $1/2 - 1/2k$.