机器学习理论研究导引
作业一

陈晟 MG21330006

2022 年 3 月 15 日

作业提交注意事项

(1) 本次作业提交截止时间为 **2022/03/15 23:59:59**，截止时间后不再接收作业，本次作业记零分；

(2) 作业提交方式：使用此 LaTex 模板书写解答，只需提交编译生成的 pdf 文件，将 pdf 文件以附件方式发送至邮箱：mlt2022spring@lamda.nju.edu.cn

(3) pdf 文件命名方式：学号-姓名-作业号-v 版本号, 例 MG1900000-张三-1-v1；如果需要更改已提交的解答，请在截止时间之前提交新版本的解答，并将版本号加一；

(4) 邮件标题命名方式同 pdf 文件命名方式；

(5) 未按照要求提交作业，或 pdf (邮件标题) 命名方式不正确，将会被扣除部分作业分数。

1 [25pts] Hölder's Inequality

Hölder 不等式是应用场景非常广泛的不等式之一。本题介绍最为常见的 Hölder 不等式的证明思路。

(1) **[10pts]** 请证明当 $a, b \geq 0$ 和 $0 \leq \theta \leq 1$ 时, 有

$$a^{\theta} b^{1-\theta} \leq \theta a + (1-\theta) b \tag{1.1}$$

(2) **[15pts]** 请利用第一问推导 Hölder 不等式: 对于 $p > 1, 1/p + 1/q = 1$ 和 $x, y \in \boldsymbol{R}^n$, 有

$$\sum_{i=1}^{n} x_i y_i \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q \right)^{1/q} \tag{1.2}$$

Solution.

(1) $\because ln\left(a^{\theta} b^{(1-\theta)}\right) \leq ln\left(\theta a + (1-\theta) b\right)$

$\because ln\left(a^{\theta} b^{(1-\theta)}\right) = \theta ln a + (1-\theta) ln b$

由于 lnx 是凹函数

$\therefore ln\left(\theta a + (1-\theta) b\right) \geq \theta ln a + (1-\theta) ln b$

即证 $a^{\theta} b^{1-\theta} \leq \theta a + (1-\theta) b$

(2) 将 $a^{\theta} b^{1-\theta} \leq \theta a + (1-\theta) b$ 改写为 $\frac{a_i}{p} + \frac{b_i}{q} \geq a_i^{\frac{1}{p}} b_i^{\frac{1}{q}}$

则令 $a_i = \frac{x_i^p}{\sum_{i=1}^n x_i^p}, b_i = \frac{y_i^q}{\sum_{i=1}^n y_i^q}$

由该写后的式子 $\because \sum_{i=1}^n \left(\frac{a^i}{p} + \frac{b^i}{q} \right) = 1 \geq \sum_{i=1}^n a_i^{\frac{1}{p}} b_i^{\frac{1}{q}}$

$\therefore \sum_{i=1}^n x_i y_i \leq \left(\sum_{i=1}^n x_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^n y_i^q\right)^{\frac{1}{q}}$

显然，上述证明了 xi 和 yi 大于等于 0 的情况，小于零时用绝对值替换也显然成立

2 [25pts] Hoeffding's Inequality

Hoeffding 不等式也是学习理论中常用的不等式之一。本题给出使用 Markov 不等式证明 Hoeffding 不等式的一种方法。

(1) **[5pts]** (Markov's inequality) 试证明：对随机变量 $X \geq 0$ 和任意的 $\epsilon > 0$，有

$$P(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon} .$$

(2) **[5pts]** (Chernoff bound) 试证明：对任意随机变量 X 和任意 $t > 0$,

$$P(X \geq \epsilon) \leq e^{-t\epsilon}\mathbb{E}\left[e^{tX}\right] .$$

(3) **[0pts]** (Hoeffding's lemma) 若随机变量 X 满足 $\mathbb{E}[X] = 0$ 且 $X \in [a, b]$，其中 $b > a$，试证明对任意 $t > 0$ 有

$$\mathbb{E}\left[e^{tX}\right] \leq exp\left(\frac{t^2(b-a)^2}{8}\right) .$$

(4) **[15pts]** (Hoeffding's inequality) 若 $X_1, X_2, ..., X_m$ 为 m 个独立随机变量，且对任意 $i \in [m]$ 有 $X_i \in [a_i, b_i]$，其中 $b_i > a_i$。定义随机变量 $S = \sum_{i=1}^{m} X_i$，试证明：对任意 $\epsilon > 0$ 有

$$P(|S - \mathbb{E}[S]| \geq \epsilon) \leq 2exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right) .$$

注意：第三小问不设分值，可以留空，其结论在第四问中可直接使用。有兴趣的同学可以尝试解决第三问。

Solution.

(1) $\because E[X] = \int X dP$,$P$ 为密度函数

$\int X dP \geq \int_{\varepsilon}^{+\infty} x dP(x) \geq t\int_{\varepsilon}^{+\infty} dP(x) = tP(X \geq \varepsilon)$

$\therefore P(X \geq \varepsilon) \leq \frac{E[X]}{\varepsilon}$

(2) 由 *(1)* $\because P\{X \geq \varepsilon\} = P\left\{e^{tX} \geq e^{t\varepsilon}\right\} \leq \frac{E\left[e^{tX}\right]}{e^{t\varepsilon}}$

$\therefore P(X \geq \varepsilon) \leq e^{-t\varepsilon}E\left[e^{tX}\right]$

(4) 取 $s \geq 0, \varepsilon \geq 0$, 由马尔可夫不等式得:$P(S - E(S) \geq \varepsilon) = P\left(e^{s(S-E(S))} \geq e^{s\varepsilon}\right) \leq e^{-s\varepsilon}E\left(e^{s(X_i - E(X_i))}\right) = e^{-s\varepsilon}\prod_{i=1}^{n} E\left(e^{s(X_i - E(X_i))}\right)$

再由 *(3)* 得：$P(S - E(S) \geq \varepsilon) \leq e^{-s\varepsilon}\prod_{i=1}^{n} e^{\frac{s^2(b_i-a_i)^2}{8}} = exp(-s\varepsilon + \frac{1}{8}s^2\sum_{i=1}^{n}(b_i - a_i)^2)$

到这一步，不等式中还多出了一个 s，因为 $s>0$，都有以上不等式成立，因此取右边关于 s 的二次函数的最小值。令 $g(s) = -s\varepsilon + \frac{1}{8}s^2\sum_{i=1}^{n}(b_i - a_i)^2$

令导数 $\dot{g}(s) = 0$

$s = \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}$

$\therefore P(S - E(S) \geq \varepsilon) \leq exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$

3 [25pts] PAC Learning for Finite Hypothesis Sets

课件中通过构造一个精巧的算法证明了布尔合取式概念类的可学习性。事实上，对于可分的有限概念类，简单的 ERM 算法也可以导出 PAC 可学习性。请证明：

令 $\mathcal{H}$ 为可分的有限概念类, D 为包含 m 个从 $\mathcal{D}$ 独立同分布采样所得的样本构成的训练集, 学习算法 $\mathfrak{L}$ 基于训练集 D 返回与训练集一致的假设 h_D, 对于任意 $c \in \mathcal{H}$, $0 < \epsilon, \delta < 1$, 如果有 $m \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$, 则

$$P\left(E(h_D) \leq \epsilon\right) \geq 1 - \delta, \tag{3.1}$$

即 $E(h) \leq \epsilon$ 以至少 $1-\delta$ 的概率成立.

提示：注意到 h_D 必然满足 $\widehat{E}_D(h_D) = 0$.

Proof.

由于 $m \geq \frac{ln|\mathcal{H}| + ln\frac{1}{\sigma}}{\epsilon}$

$\therefore m \geq \frac{\left(ln(\mathcal{H}) + ln\left(\frac{1}{\sigma}\right)\right)}{-ln(e^{-\epsilon})}$

$\therefore m \geq \frac{(ln(\mathcal{H}) - ln(\sigma))}{-ln(1-\epsilon)}$

$\therefore mln\left(1-\epsilon\right) \leq ln\left(\sigma\right) - ln\left(\mathcal{H}\right)$

因为我们需要的是给出一个一致收敛界，即包括 h_D 在内的所有一致的构成的集合均成立的界。所以我们可以把上述公式表述为所有一致但是误差大于 ϵ 的 h 概率要小于 σ

即 $Pr[\exists h \in H\mathcal{H} : \widehat{E}_D(h_D) = 0 \wedge E\left(h\right) > \epsilon] < \sigma$

由上述推导可知实际情况下 h 的概率质量小于等于 $1-\epsilon$(即分布 D 随机产生的样本 (即真正正确的样本) 落在 h 所划定的范围内的概率必须小于 $1-\epsilon$，因为如果连落入 h 的区域的概率都大于 $1-\epsilon$ 的话，则 h 的误差根本超不过 ϵ)。又因为有最多 $\mathcal{H}$ 个这样的概率相加。

$\therefore E\left(h\right) > \epsilon$

重写为概率的形式：$\sum_{h\in\mathcal{H}} Pr\left[\widehat{E}_D(h_D) = 0 | E\left(h\right) > \epsilon\right] \geq \sum_{h\in\mathcal{H}} Pr\left[\widehat{E}_D(h_D) = 0 \wedge E\left(h\right) > \epsilon\right] \geq Pr\left[\left(h_1 \in \mathcal{H} : \widehat{E}_1(h_1) = 0 \wedge E\left(h_1\right) > \epsilon\right) \vee \left(h_2 \in \mathcal{H} : \widehat{E}_2(h_2) = 0 \wedge E\left(h_2\right) > \epsilon\right) \vee ...\right]$

即 $Pr[\exists h \in H\mathcal{H} : \widehat{E}_D(h_D) = 0 \wedge E\left(h\right) > \epsilon] < \sigma$

即证：$P\left(E(h_D) \leq \epsilon\right) \geq 1 - \delta$

proof2.

用泛化边界的思路叙述为：对于任何 $\epsilon, \sigma > 0$ ，有至少 $1-\sigma$ 的概率下式成立 $E\left(h_D\right) \leq \frac{1}{m}\left(ln|\mathcal{H}| + ln\frac{1}{\sigma}\right)$

固定 $\epsilon > 0$ 。我们将有限假说集合 $\mathcal{H}$ 中泛化误差大于 ϵ 的假说再集合起来，记作 $\mathcal{H}_\epsilon = \{h \in \mathcal{H} : E\left(h\right) > \varepsilon\}$ 。我们来研究这个集合 $\mathcal{H}_\epsilon$ 。其中，可能存在这样的假说，它们的经验误差是 0，概率是：

$P\left[\exists h \in \mathcal{H}_\epsilon : \widehat{E}_D(h_D) = 0\right] = P[\widehat{E}_D(h_1) = 0 \cap ... \cap \widehat{E}_D(H_{|\mathcal{H}_\epsilon|}) = 0] \leq \sum_{h\in\mathcal{H}_\epsilon} P\left[\widehat{E}_D(h) = 0\right]$

因为 $\mathcal{H}_\epsilon$ 中的假说泛化误差都大于 ϵ ，泛化误差的含义是对随机一个样本，预测错误的概率。因此，任一 $\mathcal{H}_\epsilon$ 中的假说，其经验误差为 0 的概率为：

$P\left[\widehat{E}_D(h) = 0, h \in \mathcal{H}_\epsilon\right] = (1-\epsilon)^m$

$\therefore P\left[\exists h \in \mathcal{H}_\epsilon : \widehat{E}_D(h) = 0\right] \leq |\mathcal{H}| (1-\epsilon)^m \leq |\mathcal{H}| e^{-m\epsilon}$

我们并不知道算法将会输出哪一个一致性的假说，但我们知道 (事件的包含关系，左边事情发生，则右边事情一定发生):

$\therefore P[E(h_D) > \epsilon] \leq |\mathcal{H}| e^{-m\epsilon} = \sigma$

故至少有 $1-\sigma$ 的概率使得下式成立: $E(h_D) \leq \frac{1}{m}\left(ln|\mathcal{H}| + ln\frac{1}{\sigma}\right)$

即证: $P(E(h_D) \leq \epsilon) \geq 1-\delta$

4 [25pts] PAC Learning for Infinite Hypothesis Sets

课件中已经证明了轴平行矩形的概念类是可学习的。这启发我们，无限概念类也可能是可学习的。本题给出另一个可学习的无限概念类的简单的例子。

令 $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}\}$ 表示实轴上所有闭区间上的指示函数 $h_{a,b}(\boldsymbol{x}) = \mathbb{I}(\boldsymbol{x} \in [a, b])$ 构成的概念类. 假设目标概念 $c \in \mathcal{H}$. 试证明这个无限概念类 $\mathcal{H}$ 是 PAC 可学的.

也许有同学会注意到：定义在实轴上的分布，其累积分布函数可能是不连续的，这可能会为证明过程带来困难。在本题中我们假设：要考虑的所有分布，其累积分布函数处处连续可微。

Proof.

说一个问题是 *PAC* 可学习的，需要定义 m 个 *sample* 组成 S 空间，其中每个 *sample* 服从 D 分布，并且互相独立；如果存在一个算法 A，在 m（*sample* 个数）有限的情况下，找到假设 h；使得对于任意两个数 x，y，概率 $P(h$ 对 S 中 *sample* 预测错误次数大于 $x) < y$, 即：

$Pr_{S\sim D^m}\left[R\left(h_S\right) \leq \epsilon\right] \geq 1 - \sigma$

为了证明上面的问题是 *PAC* 可学习的，我们需要找到一个算法 A，并且证明只需要 m 个实例，就可以是的概率等式成立。这个算法就是找到实轴上的点，如果点在实轴表示的范围内，则是正例，否则是负例。接下来证明那么这样的算法是否存在

我们选定一条实轴，假设该实轴上能够包含所有正例的点，排除所有负例的点。然后让该实轴向内扩展，画出两条新轴 *l1,l2*，每条轴的概率 *x/4*。中间的轴表示为 L

如果 L 的错误个数大于 x，那么 L 必然与 *l1,l2* 中的至少一个有交集。(否则错误个数必定小于 x)

$\therefore Pr_{S\sim D^m}\left[R\left(h_S\right) > \epsilon\right] \leq \left[\bigcup_{i=1}^{2}\left\{R_S \bigcap l_i = \phi\right\}\right]$

由于并集的概率小于各自概率的和：

$\therefore Pr_{S\sim D^m}\left[R\left(h_S\right) > \epsilon\right] \leq \left[\bigcup_{i=1}^{2}\left\{R_S \bigcap l_i = \phi\right\}\right] \leq Pr_{S\sim D^m}\left[\left\{R_S \bigcap l_i = \phi\right\}\right] \leq \sum_{i=1}^{2} Pr_{S\sim D^m}\left[\left\{R_S \bigcap l_i = \phi\right\}\right.$

由于 S 中的每个 *sample* 的独立分布的，并且落在 *r1* 中的概率为 *x/4*，所以

$\leq \sum_{i=1}^{2} PR_{S\sim D^m}\left[\left\{R_S \bigcap l_i = \phi\right\}\right] \leq 2\left(1 - \epsilon/2\right)_m \leq 2exp\left(-m\epsilon/2\right)$

由于我们要求错误个数大于 x 的概率小于 y，所以可以定义如下的不等式。

$2exp\left(-m\epsilon/2\right) \leq \delta \Leftrightarrow m \geq \frac{2}{\epsilon} log\frac{2}{\delta}$

推导出了 m 的下限，这就说明了，上面的问题是 *PAC* 可学习的。