机器学习 课程笔记

酥雨
zusuyu@stu.pku.edu.cn

April 26, 2022

目录

1 Inequalities

定理 1.1 (Markov Inequality). 如果非负随机变量 X 期望存在, 则对于任意 $k>0$,

$$\mathbb{P}[X \geqslant k] \leqslant \frac{\mathbb{E}[X]}{k}$$

进一步地, 如果 r 阶矩 $\mathbb{E}[X^r]$ 存在, 则对于任意 $k>0$,

$$\mathbb{P}[X \geqslant k] \leqslant \min_{j \leqslant r} \frac{\mathbb{E}\left[X^j\right]}{k^j}$$

定理 1.2 (Chebyshev Inequality). 如果随机变量 X 方差存在, 则对于任意 $\varepsilon>0$,

$$\mathbb{P}[|X-\mathbb{E}[X]| \geqslant \varepsilon] \leqslant \frac{\operatorname{Var}[X]}{\varepsilon^2}$$

定义 1.1 (矩生成函数, Moment Generating Function, MGF). 如果随机变量 X 的任意 $n \in \mathbb{N}$ 阶矩存在, 则定义其矩生成函数为

$$M_X(t)=\mathbb{E}\left[\mathrm{e}^{tX}\right]=\sum_{i \geqslant 0} t^i \frac{\mathbb{E}\left[X^i\right]}{i!}$$

定理 1.3 (Chernoff Inequality).

$$\mathbb{P}[X \geqslant k] \leqslant \inf_{t>0} \mathrm{e}^{-tk} M_X(t)$$

定理 1.4. $X_1, X_2, \cdots, X_n \sim$ i.i.d. $\mathcal{B}(1,p)$, 对于任意 $\varepsilon>0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-nD_B(p+\varepsilon\|p)}$$

其中 $D_B(p\|q)$ 是两个 Bernoulli distribution $P=(p,1-p), Q=(q,1-q)$ 之间的相对熵.

证明.

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - p \geqslant \varepsilon\right] &= \mathbb{P}\left[\sum_{i=1}^n X_i \geqslant n(p+\varepsilon)\right] \\
&\leqslant \inf_{t>0} \mathrm{e}^{-tn(p+\varepsilon)} \mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^n X_i}\right] \\
&= \inf_{t>0} \mathrm{e}^{-tn(p+\varepsilon)} \prod_{i=1}^n \mathbb{E}\left[\mathrm{e}^{tX_i}\right] \\
&= \inf_{t>0} \mathrm{e}^{-tn(p+\varepsilon)} (p\mathrm{e}^t + 1 - p)^n \\
&= \inf_{t>0} \left(\frac{p\mathrm{e}^t + 1 - p}{\mathrm{e}^{t(p+\varepsilon)}}\right)^n
\end{aligned}$$

通过 “简单” 求导, 取 $t=\ln \frac{(1-p)(p+\varepsilon)}{p(1-p-\varepsilon)}$ 时上式右边取最小值, 从而有

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - p \geqslant \varepsilon\right] &\leqslant \left(\frac{\frac{(1-p)(p+\varepsilon)}{1-p-\varepsilon} + 1 - p}{\left(\frac{(1-p)(p+\varepsilon)}{p(1-p-\varepsilon)}\right)^{p+\varepsilon}}\right)^n = \left(\frac{\frac{1-p}{1-p-\varepsilon}}{\left(\frac{(1-p)(p+\varepsilon)}{p(1-p-\varepsilon)}\right)^{p+\varepsilon}}\right)^n \\
&= \left(\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right)^n = \mathrm{e}^{-nD_B(p+\varepsilon\|p)}
\end{aligned}$$

□

定理 1.5. $X_1, X_2, \cdots, X_n \in [0,1]$ 是 n 个期望相同的独立随机变量, $\mathbb{E}[X_i]=p$, 对于任意 $\varepsilon>0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-nD_B(p+\varepsilon\|p)}$$

证明. 注意到指数函数是下凸的, 根据 Jensen Inequality, 有

$$\mathbb{E}\left[\mathrm{e}^{tX}\right] \leqslant \mathbb{E}\left[X\mathrm{e}^{t} + (1-X)\mathrm{e}^{0}\right] = p\mathrm{e}^{t} + 1 - p$$

从而

$$\mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^{n} X_i}\right] \leqslant (p\mathrm{e}^{t} + 1 - p)$$

沿用定理 1.4 的证明即可. □

定理 1.6 (Chernoff Bound). $X_1, X_2, \cdots, X_n \in [0,1]$ 是 n 个独立随机变量, $\mathbb{E}[X_i] = p_i$, 记 $p = \frac{1}{n}\sum_{i=1}^{n} p_i$, 对于任意 $\varepsilon > 0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-nD_B(p+\varepsilon\|p)}$$

证明. 注意到对数函数是上凸的, 从而函数 $f(x) = \ln(x\mathrm{e}^{t} + 1 - x)$ 也是上凸的, 同样根据 Jensen Inequality, 有

$$\frac{1}{n}\sum_{i=1}^{n} \ln(p_i\mathrm{e}^{t} + 1 - p_i) \leqslant \ln(p\mathrm{e}^{t} + 1 - p)$$

从而

$$\mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^{n} X_i}\right] \leqslant \prod_{i=1}^{n}(p_i\mathrm{e}^{t} + 1 - p_i) \leqslant (p\mathrm{e}^{t} + 1 - p)^n$$

□

定理 1.7 (Additive Chernoff Bound). $X_1, X_2, \cdots, X_n \in [0,1]$ 是 n 个独立随机变量, $\mathbb{E}[X_i] = p_i$, 记 $p = \frac{1}{n}\sum_{i=1}^{n} p_i$, 对于任意 $\varepsilon > 0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-2n\varepsilon^2}$$

证明. 只需要证明 $D_B(p+\varepsilon\|p) \geqslant 2\varepsilon^2$ 即可. 听说可以暴力求导. □

定理 1.8 (Hoeffding Bound). $X_1, X_2, \cdots, X_n$ 是 n 个独立随机变量, $X_i \in [a_i, b_i]$, 记 $p = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i+b_i}{2}$, 对于任意 $\varepsilon > 0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-\frac{2n\varepsilon^2}{\left(\frac{1}{n}\sum_{i=1}^{n} b_i - a_i\right)^2}} \leqslant \mathrm{e}^{-\frac{2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

定理 1.9 (McDiarmid Inequality). $X_1, X_2, \cdots, X_n \in \mathcal{X}$ 是 n 个独立随机变量, 如果对于 $f : \mathcal{X}^n \to \mathbb{R}$ 存在常数 $c_1, c_2, \cdots, c_n$ 使得

$$|f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \leqslant c_i$$

对于任意 $i \in [n], x_1, \cdots, x_n, x_i'$ 成立, 则对于任意 $\varepsilon > 0$, 有

$$\mathbb{P}\left[f(x_1, \cdots, x_n) - \mathbb{E}\left[f(x_1, \cdots, x_n)\right] \geqslant \varepsilon\right] \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right)$$

定理 1.10 (Draw with/without Replacement). 有 m 个数 $a_1, \cdots, a_m \in \{0,1\}$, 记 $p = \frac{1}{m}\sum_{i=1}^{m} a_i$. $X_1, \cdots, X_n$ 为从 $\{a_1, \cdots, a_m\}$ 中的随机放回抽样, $Y_1, \cdots, Y_n$ 为从 $\{a_1, \cdots, a_m\}$ 中的随机不放回抽样, 则对于任意 $\varepsilon > 0$ 有

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-2n\varepsilon^2}, \qquad \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-2n\varepsilon^2}$$

证明. 对于随机放回抽样, 显然每次抽样是独立的, 从而结论是 Chernoff Bound 的平凡推论.

对于随机不放回抽样, 注意到 $\mathbb{E}\left[\prod_{i\in I} Y_i\right] \leqslant \mathbb{E}\left[\prod_{i\in I} X_i\right]$ 对任意指标集 $I \subseteq \{1,\cdots,n\}$ 成立, 从而可以证明 $\mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^{n} Y_i}\right] \leqslant \mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^{n} X_i}\right]$. □

2 VC Theory

对一个分类器 f, 通常有两种评价指标: training error $err_S(f) = \mathbb{P}_{(x,y)\in S}[y \neq f(x)]$ 与 generalization error $err_D(f) = \mathbb{P}_{(x,y)\sim D}[y \neq f(x)]$. 接下来可能会不加声明地用 S 表示从数据集 D 中 sample 出来的训练集.

称 $err_D(f) - err_S(f)$ 为分类器 f 的 generalization gap. 我们提出**一致收敛 (uniformly converge)** 的概念, 它表示随着训练集 S 的增大, hypothesis space $\mathcal{F}$ 中的所有分类器 f 的 generalization gap 都会 "一致" 地被 bound 住.

定理 2.1 (Uniform Convergence when $|\mathcal{F}| < \infty$). S 是从数据集 D 中随机采样的训练集, $|S| = n$, 有

$$\mathbb{P}\left[\forall f \in \mathcal{F}, err_D(f) - err_S(f) \geqslant \varepsilon\right] \leqslant |\mathcal{F}|\mathrm{e}^{-2n\varepsilon^2}$$

证明. 对于某个确定的 $f \in \mathcal{F}$, 注意到 $err_S(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i \neq f(x_i)]$, $\mathbb{E}\left[y_i \neq f(x_i)\right] = err_D(f)$, 故根据 Chernoff Bound 有 $\mathbb{P}\left[err_D(f) - err_S(f) \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-2n\varepsilon^2}$. 再结合 Union Bound 即得结论. □

定理 2.2 (VC Theorem). 对于 VC-dimension (会在接下来定义) 为 d 的 hypothesis space $\mathcal{F}$, 从数据集 D 中随机采样大小为 n 的训练集 S, 则

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}|err_D(f) - err_S(f)| \geqslant \varepsilon\right] \leqslant 2\left(\frac{2\mathrm{e}n}{d}\right)^d \mathrm{e}^{-cn\varepsilon^2}$$

其中 c 是常数. 或者等价地, 有至少 $1-\delta$ 的概率, 对所有 $f \in \mathcal{F}$ 有

$$err_D(f) \leqslant err_S(f) + O\left(\sqrt{\frac{d\ln n - \ln\delta}{n}}\right)$$

为了接下来的叙述方便, 我们引入一些记号:

- 对于分类器 $f \in \mathcal{F}$ 以及数据点 $z = (x, y) \sim D$, 定义 $\phi_f(z) = \mathbb{1}[y \neq f(x)]$, 即每个 ϕ_f 是一个 "长度为 $|D|$" 的 01 串, 1 表示 f 会在这一位对应的数据点上出错.
- 定义 $\Phi_{\mathcal{F}} = \{\phi_f | f \in \mathcal{F}\}$. 由于以下不会超过一个 hypothesis space, 故省略角标简记为 Φ.

如此一来, 对于 $S = \{z_1 = (x_1, y_1), \cdots, z_n = (x_n, y_n)\}$, 两种错误率 $err_S(f)$ 和 $err_D(f)$ 就分别等价于 $\frac{1}{n}\sum_{i=1}^{n}\phi_f(z_i)$ 和 $\mathbb{E}_{z\sim D}\phi_f(z)$, 而我们需要限制的概率也变成了

$$\mathbb{P}_{S\sim D^n}\left[\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i) - \mathbb{E}_{z\sim D}[\phi(z)]\right| \geqslant \varepsilon\right]$$

引理 2.1 (Double Sampling). 取 $n \geqslant \frac{\ln 2}{\varepsilon^2}$, 有

$$\mathbb{P}_{S\sim D^n}\left[\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i) - \mathbb{E}_{z\sim D}[\phi(z)]\right| \geqslant \varepsilon\right] \leqslant 2\mathbb{P}_{S\sim D^{2n}}\left[\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_i)\right| \geqslant \frac{\varepsilon}{2}\right]$$

通过 Double Sampling, 我们只需要限制 $\frac{1}{n}\sum_{i=1}^{n}\phi(z_i)$ 与 $\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_i)$ 的差. 考虑一种新的抽样方式, 先随机抽取 $\{z_1, \cdots, z_{2n}\}$, 再对其随机排列, 这样显然是与原先等价的, 即

$$\mathbb{P}_{S\sim D^{2n}}\left[\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_i)\right| \geqslant \varepsilon\right] = \mathbb{E}_{S\sim D^n}\left[\mathbb{P}_\sigma\left[\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)}) - \frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\right| \geqslant \varepsilon\right]\right]$$

这么做的意义是什么? 意义是可以先只考虑内层的 $\mathbb{P}_\sigma$ 而不管 $S \sim D^n$ 的选取. 看似强行取的随机排列 σ 是为了内层可以被 bound, 不然 $\mathbb{1}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_i)\right| \geqslant \varepsilon\right]$ 还不太方便处理.

记 $N^{\Phi}(z_1,\cdots,z_n)$ 表示 $\#\{(\phi(z_1),\cdots,\phi(z_n))|\phi\in\Phi\}$, 即 Φ 中的所有 01 串在数据点 $z_1,\cdots,z_n$ 上有多少种不同的. 从这个角度想, 其实 $\sup_{\phi\in\Phi}$ 只是在对有限项求 max, 故根据 Union Bound 可以得到

$$\mathbb{P}_\sigma\left[\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\right|\geqslant\varepsilon\right]\leqslant N^{\Phi}(z_1,\cdots,z_{2n})\mathbb{P}_\sigma\left[\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\right|\geqslant\varepsilon\right]$$

其实这里写得不太严谨, 右边应该是对 $N^{\Phi}(z_1,\cdots,z_{2n})$ 个不同的 ϕ 分别求概率再相加, 但我们接下来会对任意 ϕ 限制 $\mathbb{P}_\sigma\left[\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\right|\geqslant\varepsilon\right]$, 所以应该也无伤大雅.

对于一个特定的 $\phi\in\Phi$, 考虑 $\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})$ 其实就是在 $\{\phi(z_1),\cdots,\phi(z_{2n})\}$ 这 $2n$ 个数中做不放回抽样, 故根据定理 1.10, 有

$$\begin{aligned}
\mathbb{P}_\sigma\left[\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\right|\geqslant\varepsilon\right]&=2\mathbb{P}_\sigma\left[\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\geqslant\varepsilon\right]\\
&=2\mathbb{P}_\sigma\left[\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{2n}\sum_{i=1}^{2n}\phi(z_{\sigma(i)})\geqslant\frac{\varepsilon}{2}\right]\\
&=2\mathbb{P}_\sigma\left[\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-p\geqslant\frac{\varepsilon}{2}\right]\\
&\leqslant 2\mathrm{e}^{-\frac{n\varepsilon^2}{2}}
\end{aligned}$$

从而我们得到了

$$\mathbb{P}_{S\sim D^{2n}}\left[\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i)-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_i)\right|\geqslant\varepsilon\right]\leqslant 2\mathrm{e}^{-\frac{n\varepsilon^2}{2}}\mathbb{E}_{S\sim D^{2n}}\left[N^{\Phi}(z_1,\cdots,z_{2n})\right]$$

3 Lagrange Duality

4 Game Theory

Game theory is the study of mathematical models of strategic interactions among rational agents, cited from Wikipedia.

我们引入“双人矩阵博弈”作为对博弈论最基础的介绍. 注意, 接下来我们考虑的所有问题都是零和的.

定义 4.1 (Two-player Matrix Game). 有一个 $M \in \mathbb{R}^{m\times n}$ 的矩阵. 两名玩家 Alice 和 Bob 参加了这场博弈. Alice, **the row player** 选择一行 $i \in [m]$, 相应的, Bob, **the column player** 选择一列 $j \in [n]$, 此时 Alice 获得收益 $-M_{ij}$, Bob 获得收益 M_{ij}.

我们首先探讨**纯策略 (pure strategy)** 的情景, 指的是 Alice 和 Bob 必须分别选择某个确定的行或列.

当 Alice 先做出选择时, 当她选出第 i 行后, 她会认为 Bob 会选择第 $j_i = \arg\max_j M_{ij}$ 列, 因此她会选择第 $\arg\min_i \max_j M_{ij}$ 行, 导致最终的博弈结果为 $\min_i \max_j M_{ij}$.

同理, 当 Bob 先做选择时, 他会选择第 $\arg\max_j \min_i M_{ij}$ 列, 导致最终的博弈结果为 $\max_j \min_i M_{ij}$.

我们指出在纯策略的情境下, 后手是有优势的, 即

定理 4.1. $\min_i \max_j M_{ij} \geqslant \max_j \min_i M_{ij}$ 对于任意 $M \in \mathbb{R}^{m\times n}$ 都成立, 同时存在 M', 使 $\min_i \max_j M'_{ij} > \max_j \min_i M'_{ij}$.

证明. 记 $i_0 = \arg\min_i \max_j M_{ij}, j_0 = \arg\max_j \min_i M_{ij}$, 有

$$\min_i \max_j M_{ij} = \max_j M_{i_0 j} \geqslant M_{i_0 j_0} \geqslant \min_i M_{i j_0} = \max_j \min_i M_{ij}$$

考虑 $M' = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$, 有 $\min_i \max_j M'_{ij} = -1, \max_j \min_i M'_{ij} = 1$. □

接着我们研究**混合策略 (mixed strategy)**, 其意味着玩家做出的决策可以不是确定的行列选择, 而是一个概率分布. 相应地, 得到的收益也就变成了期望收益.

形式化地, Alice 选择给出概率分布 $p = (p_1, \cdots, p_m) \in [0,1]^m$, Bob 给出概率分布 $q = (q_1, \cdots, q_n) \in [0,1]^n$. 合法的概率分布需要满足 $\|p\|_1 = \|q\| = 1$, 而此时两人的收益也分别是 $-p^{\mathrm{T}}Mq$ 与 $p^{\mathrm{T}}Mq$.

与纯策略的情境同理, 当 Alice 先手时, 博弈结果为 $\min_{p\in[0,1]^m, \|p\|_1=1} \max_{q\in[0,1]^n, \|q\|_1=1} p^{\mathrm{T}}Mq$, 当 Bob 先手时, 博弈结果为 $\max_{q\in[0,1]^n, \|q\|_1=1} \min_{p\in[0,1]^m, \|p\|_1=1} p^{\mathrm{T}}Mq$. 在接下来的叙述中, 我们默认 p, q 应取合法的概率分布, 而忽略在 $\min, \max$ 记号下的明确限制.

我们想要知道混合策略下后手还有没有优势. John von Neuman 告诉我们, 没有.

定理 4.2 (von Neuman Minimax Theorem).

$$\min_p \max_q p^{\mathrm{T}}Mq = \max_q \min_p p^{\mathrm{T}}Mq$$

5 Boosting

Algorithm 1 AdaBoost

Require: training set $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, weak learning algorithm $\mathcal{A}$

$D_1(i) \leftarrow \frac{1}{n}, \forall i \in [n]$.
for $t = 1 \rightarrow T$ **do**
  Use $\mathcal{A}$ to learn a classifier h_t based on D_t.
  $\varepsilon_t \leftarrow \sum_{i=1}^{n} D_t(i)\mathbb{1}[y_i \neq h_t(x_i)]$
  $\gamma_t \leftarrow 1 - 2\varepsilon_t$
  $\alpha_t \leftarrow \frac{1}{2}\ln\frac{1+\gamma_t}{1-\gamma_t}$
  $Z_t \leftarrow 2(\varepsilon_t(1-\varepsilon_t))^2$
  $D_{t+1}(i) \leftarrow \frac{1}{Z_t} D_t(i)\exp(-y_i\alpha_t h_t(x_i))$
end for
return a classifier F, $F(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

6 PAC-Bayesian Theory

定理 6.1 (PAC-Bayesian Theorem). 对于给定的 prior distribution of classifiers $\mathcal{P}$, 从数据集 D 中随机抽取大小为 n 的训练集 S, 有至少 $1-\delta$ 的概率, 对于任意 distribution of classifiers $\mathcal{Q}$ 有如下不等式成立

$$\mathbb{E}_{h\sim\mathcal{Q}}[err_D(h)] \leqslant \mathbb{E}_{h\sim\mathcal{Q}}[err_S(h)] + \sqrt{\frac{D_{KL}(\mathcal{Q}\|\mathcal{P}) + \log(3/\delta)}{n}}$$

其中 $err_X(f)$ 表示 classifier f 在数据集 X 上的错误率, 即 $\mathbb{P}_{(x,y)\in X}[y \neq f(x)]$, $D_{KL}(\mathcal{Q}\|\mathcal{P}) = \mathbb{E}_{h\sim\mathcal{Q}}\left[\ln \frac{\mathcal{Q}_h}{\mathcal{P}_h}\right]$ 为概率分布 $\mathcal{Q}$ 与 $\mathcal{P}$ 的 KL 散度.

引理 6.1. 对于任意在 hypothesis space $\mathcal{F}$ 上的概率分布 $\mathcal{P}, \mathcal{Q}$, 以及任意函数 $f : \mathcal{F} \to \mathbb{R}$, 都有

$$\mathbb{E}_{h\sim Q}[f(h)] \leqslant \ln \mathbb{E}_{h'\sim P}[\exp(f(h'))] + D_{KL}(\mathcal{Q}\|\mathcal{P})$$

证明.

$$\begin{aligned}
\text{RHS} - \text{LHS} &= \ln \mathbb{E}_{h'\sim P}[\exp(f(h'))] + D_{KL}(\mathcal{Q}\|\mathcal{P}) - \mathbb{E}_{h\sim Q}[f(h)] \\
&= \ln \mathbb{E}_{h'\sim P}[\exp(f(h'))] + \mathbb{E}_{h\sim\mathcal{Q}}\left[\ln \frac{\mathcal{Q}_h}{\mathcal{P}_h}\right] - \mathbb{E}_{h\sim Q}[f(h)] \\
&= \mathbb{E}_{h\sim Q}\left[\ln \frac{\mathcal{Q}_h}{\frac{\mathcal{P}_h \exp(f(h))}{\mathbb{E}_{h'\sim\mathcal{P}}[\exp(f(h'))]}}\right] \\
&= \mathbb{E}_{h\sim Q}\left[\ln \frac{\mathcal{Q}_h}{\mathcal{R}_h}\right] \\
&= D_{KL}(\mathcal{Q}\|\mathcal{R}) \\
&\geqslant 0
\end{aligned}$$

其中 $\mathcal{R}$ 也是一个 $\mathcal{F}$ 上的概率分布, $\mathcal{R}_h = \frac{\mathcal{P}_h \exp(f(h))}{\mathbb{E}_{h'\sim\mathcal{P}}[\exp(f(h'))]}$. □

引理 6.2. 对于任意 $\delta > 0$, 有

$$\mathbb{P}_{S\sim D^n}\left(\mathbb{E}_{h\sim\mathcal{P}}[e^{n(err_D(h)-err_S(h))^2}] \geqslant 3/\delta\right) \leqslant \delta$$

证明. 先证明对于某个固定的 $h \sim \mathcal{P}$, 有

$$\mathbb{E}_{S\sim D^n}[e^{n(err_D(h)-err_S(h))^2}] \leqslant 3$$

记 $\Delta = |err_D(h) - err_S(h)|$, 根据 Chernoff bound, 有

$$\mathbb{P}_{S\sim D^n}(\Delta \geqslant \varepsilon) \leqslant 2\exp(-2n\varepsilon^2)$$

于是

$$\begin{aligned}
\mathbb{E}_{S\sim D^n}\left[e^{n\Delta^2}\right] &= \int_0^{+\infty} \mathbb{P}_{S\sim D^n}\left(e^{n\Delta^2} \geqslant t\right) \mathrm{d}t \\
&= \int_1^{+\infty} \mathbb{P}_{S\sim D^n}\left(\Delta \geqslant \sqrt{\frac{\ln t}{n}}\right) \mathrm{d}t + 1 \\
&\leqslant \int_1^{+\infty} 2e^{-2\ln t} \mathrm{d}t + 1 \\
&= 3
\end{aligned}$$

随后, 使用 Markov Inequality 得到

$$\mathbb{P}_{S\sim D^n}\left(\mathbb{E}_{h\sim\mathcal{P}}[\mathrm{e}^{n\Delta^2}]\geqslant 3/\delta\right)\leqslant\frac{\mathbb{E}_{S\sim D^n}\left(\mathbb{E}_{h\sim\mathcal{P}}[\mathrm{e}^{n\Delta^2}]\right)}{3/\delta}=\frac{\mathbb{E}_{h\sim\mathcal{P}}\left(\mathbb{E}_{S\sim D^n}[\mathrm{e}^{n\Delta^2}]\right)}{3/\delta}\leqslant\frac{\mathbb{E}_{h\sim\mathcal{P}}(3)}{3/\delta}=\delta$$

□

我们利用上述两个引理证明定理 6.1. 有至少 $1-\delta$ 的概率,

$$\begin{aligned}\left(\mathbb{E}_{h\sim\mathcal{Q}}[err_D(h)-err_S(h)]\right)^2&\leqslant\mathbb{E}_{h\sim\mathcal{Q}}[\Delta^2]\\&=\frac{1}{n}\mathbb{E}_{h\sim\mathcal{Q}}[n\Delta^2]\\&\leqslant\frac{1}{n}\left(\ln\mathbb{E}_{h\sim P}\left[\mathrm{e}^{n\Delta^2}\right]+D_{KL}(\mathcal{Q}\|\mathcal{P})\right)\\&\leqslant\frac{1}{n}\left(\ln(3/\delta)+D_{KL}(\mathcal{Q}\|\mathcal{P})\right)\end{aligned}$$

其中第一行等号使用了 Cauchy Inequality, 第三行使用了引理 6.1 代入 $f(h)=n\Delta^2$, 第四行使用了引理 6.2, with probability at least $1-\delta$.

6.1 PAC-Bayesian Bound for SVM

命题 6.1. 对于任意的 distribution of classifiers $\mathcal{Q}$, 令 $g_{\mathcal{Q}}$ 为一个确定性二分类器, $g_{\mathcal{Q}}(x)=\mathrm{sgn}(\mathbb{E}_{h\sim\mathcal{Q}}h(x))$, 则

$$err_D(g_{\mathcal{Q}})\leqslant 2\mathbb{E}_{h\sim\mathcal{Q}}[err_D(h)]$$

证明. 如果 $g_{\mathcal{Q}}$ 在一个数据点 x 上出错, 则说明 $\mathcal{Q}$ 中至少一半的 classifier 都在 x 上出错. □

考虑两个 distribution of classifiers $\mathcal{P}=\mathcal{N}(\mathbf{0},I_d),\mathcal{Q}=\mathcal{N}(\mu\mathbf{w},I_d)$, 其中 $\|\mathbf{w}\|_2=1$, μ 是缩放系数. 此时 $g_{\mathcal{Q}}$ 就是传统理解下的 linear classifier $\mathbf{w}$ (这里不考虑常数 b).

根据定理 6.1 的结论, 我们有

$$err_D(g_{\mathcal{Q}})\leqslant 2\left[\mathbb{E}_{h\sim\mathcal{Q}}err_S(h)+\sqrt{\frac{D_{KL}(\mathcal{Q}\|\mathcal{P})+\log(3/\delta)}{n}}\right]$$

$$\begin{aligned}D_{KL}(\mathcal{Q}\|\mathcal{P})&=\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}\|\mathbf{x}-\mu\mathbf{w}\|^2\right]\frac{1}{2}\left(\|\mathbf{x}\|^2-\|\mathbf{x}-\mu\mathbf{w}\|^2\right)\mathrm{d}\mathbf{x}\\&=\int_{\lambda}\int_{\mathbf{y}\in\mathbb{R}^{d-1},\mathbf{y}\perp\mathbf{w}}\frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}\|\lambda\mathbf{w}+\mathbf{y}-\mu\mathbf{w}\|^2\right]\frac{1}{2}\left(\|\lambda\mathbf{w}+\mathbf{y}\|^2-\|\lambda\mathbf{w}+\mathbf{y}-\mu\mathbf{w}\|^2\right)\mathrm{d}\lambda\mathrm{d}\mathbf{y}\\&=\int_{\lambda}\int_{\mathbf{y}\in\mathbb{R}^{d-1},\mathbf{y}\perp\mathbf{w}}\frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}(\lambda-\mu)^2-\frac{1}{2}\|\mathbf{y}\|^2\right]\frac{1}{2}\left(\lambda^2+\|\mathbf{y}\|^2-(\lambda-\mu)^2-\|\mathbf{y}\|^2\right)\mathrm{d}\lambda\mathrm{d}\mathbf{y}\\&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\lambda-\mu)^2\right]\frac{1}{2}(2\lambda\mu-\mu^2)\mathrm{d}\lambda\left[\int_{\mathbf{y}\in\mathbb{R}^{d-1},\mathbf{y}\perp\mathbf{w}}\frac{1}{(2\pi)^{(d-1)/2}}\exp\left(-\frac{1}{2}\|\mathbf{y}\|^2\right)\mathrm{d}\mathbf{y}\right]\\&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\lambda-\mu)^2\right](\lambda\mu-\mu^2)\mathrm{d}\lambda+\frac{\mu^2}{2}\\&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\lambda-\mu)^2\right]\mu\mathrm{d}\frac{(\lambda-\mu)^2}{2}+\frac{\mu^2}{2}\\&=\frac{\mu^2}{2}\end{aligned}$$

7 Algorithmic Stability

定义 7.1 (一致稳定, Uniform Stability). $\mathcal{A}$ 是输入训练集 $S=(z_1,\cdots,z_n)$, 输出一个分类器 $\mathcal{A}(S)$ 的学习算法. 记 $S^i=(z_1,\cdots,z_{i-1},z_i',z_{i+1},\cdots,z_n)$ 是与 S 只相差第 i 个数据点的相邻训练集, $\ell(\cdot,\cdot)$ 是损失函数, 即 $\ell(f,z)$ 是在分类器 f 下, 数据点 z 产生的损失.

称学习算法 $\mathcal{A}$ 关于 $\ell(\cdot,\cdot)$ 满足 $\beta(n)$-一致稳定性, 如果对于任意大小为 n 的训练集 S 及其相邻训练集 S^i, 以及任意数据点 z, 都有

$$|\ell(\mathcal{A}(S),z)-\ell(\mathcal{A}(S^i),z)|\leqslant\beta(n)$$

定义 7.2 (Risk & Empirical Risk). 分别类似于 test error 与 training error, 定义 risk 与 empirical risk 为

$$R(\mathcal{A}(S))=\mathbb{E}_{z\sim D}[\ell(\mathcal{A}(S),z)]$$

$$R_{emp}(\mathcal{A}(S))=\frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}(S),z_i)$$

以下讨论中不会出现超过一个学习算法, 故简记 $\Phi(S)=R(\mathcal{A}(S))-R_{emp}(\mathcal{A}(S))$.

定理 7.1 (一致稳定能说明泛化). 对于一个关于 $\ell(\cdot,\cdot)$ 满足 $\beta(n)$-一致稳定性的学习算法 $\mathcal{A}$, 其中 $|\ell(\cdot,\cdot)|\leqslant M$ 有上界, 有

$$\mathbb{P}\left[\Phi(S)\leqslant\varepsilon+\beta(n)\right]\leqslant\exp\left(-\frac{n\varepsilon^2}{2(n\beta(n)+M)^2}\right)$$

或者等价的, 有至少 $1-\delta$ 的概率下式成立

$$R(\mathcal{A}(s))\leqslant R_{emp}(\mathcal{A}(s))+\beta(n)+(n\beta(n)+M)\sqrt{\frac{2\ln(1/\delta)}{n}}$$

证明. 先证明两个引理.

引理 7.1. 假设 $\mathcal{A}$ 是对称的, 即对于任意 n 元置换 σ, 有 $\mathcal{A}(\{z_1,\cdots,z_n\})=\mathcal{A}(\{z_{\sigma_1},\cdots,z_{\sigma_n}\})$, 则

$$\mathbb{E}_S[\Phi(S)]\leqslant\beta(n)$$

证明.

$$\mathbb{E}_S[\Phi(S)]=\mathbb{E}_{S,z}[\ell(\mathcal{A}(S),z)]-\mathbb{E}_S[\ell(\mathcal{A}(S),z_1)]=\mathbb{E}_{S,S^1}[\ell(\mathcal{A}(S^1),z_1)-\ell(\mathcal{A}(S),z_1)]\leqslant\beta(n)$$

□

引理 7.2. 如果 $|\ell(\cdot,\cdot)|\leqslant M$ 有上界, 则对于任意 S,S^i, 有

$$|\Phi(S)-\Phi(S^i)|\leqslant 2\left(\beta(n)+\frac{M}{n}\right)$$

证明. 除了 $\ell(\mathcal{A}(S),z_i)-\ell(\mathcal{A}(S^i),z_i')$ 一项外, 其余所有项都可以被 $\beta(n)$-稳定性限制住.

$$\begin{aligned}
|\Phi(S)-\Phi(S^i)|&=|R(\mathcal{A}(S))-R_{emp}(\mathcal{A}(S))-R(\mathcal{A}(S^i))+R_{emp}(\mathcal{A}(S^i))|\\
&\leqslant|R_{emp}(\mathcal{A}(S))-R_{emp}(\mathcal{A}(S^i))|+|R(\mathcal{A}(S))-R(\mathcal{A}(S^i))|\\
&=\frac{1}{n}|\ell(\mathcal{A}(S),z_i)-\ell(\mathcal{A}(S^i),z_i')|+\frac{1}{n}\sum_{j\neq i}|\ell(\mathcal{A}(S),z_j)-\ell(\mathcal{A}(S^i),z_j)|+\left|\mathbb{E}_{z\sim D}[\ell(\mathcal{A}(S),z)-\ell(\mathcal{A}(S^i),z)]\right|\\
&\leqslant\frac{2M}{n}+\frac{n-1}{n}\beta(n)+\beta(n)\\
&\leqslant 2\left(\beta(n)+\frac{M}{n}\right)
\end{aligned}$$

□

考虑 McDiarmid Inequality (定理 1.9), 把 Φ 视作一个关于 $z_1, \cdots, z_n$ 的多元函数, 则引理 7.1 与引理 7.2 分别给出了 Φ 的期望以及在相邻输入上的差的上界. 于是

$$\mathbb{P}\left[\Phi(S) \geqslant \beta(n) + \varepsilon\right] \leqslant \mathbb{P}\left[\Phi(S) - \mathbb{E}\left[\Phi(S)\right] \geqslant \varepsilon\right] \leqslant \exp\left(-\frac{2n\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right) = \exp\left(-\frac{n\varepsilon^2}{2(n\beta(n) + M)^2}\right)$$

□

8 Unsupervised Learning

前面讨论的都是监督学习. 现在我们讨论一下无监督学习.
无监督学习其实主要在做两件事情: Clustering, 以及 Dimensionality Reduction.

8.1 Clustering

对于一组 $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$, 需要把这些数据点划分成 k 个 cluster $S_1, \cdots, S_k$.
可以如下定义一种划分的损失函数: 记 $\mu_i = \frac{1}{|S_i|} \sum_{j \in S_i} \mathbf{x}_j$ 为第 i 个 cluster 的中心, 损失函数为

$$L(\{S_1, \cdots, S_k\}) = \sum_{i=1}^{k} \sum_{j \in S_i} \|\mathbf{x}_j - \mu_i\|^2$$

8.1.1 K-means

Algorithm 2 K-means

1: Choose k points as cluster centers $\mu_1, \cdots, \mu_k$ uniformly at random.
2: **repeat**
3: $\quad S_i \leftarrow \{j : \|\mathbf{x}_j - \mu_i\|^2 \leqslant \|\mathbf{x}_j - \mu_k\|^2, \forall k \in [m]\}$
4: $\quad \mu_i \leftarrow \frac{1}{|S_i|} \sum_{j \in S_i} \mathbf{x}_j$
5: **until** k cluster centers do not change
6: **return** $\{\mu_1, \cdots, \mu_k\}$

8.1.2 K-means++

Algorithm 3 K-means++

1: Choose a point as the cluster center μ_1 uniformly at random.
2: **for** $i : 2 \to n$ **do**
3: $\quad$ Choose a point as the cluster center μ_i, with probability proportional to $\min_{1 \leqslant k < i} \|\mathbf{x}_j - \mu_k\|^2$.
4: **end for**
5: **return** $\{\mu_1, \cdots, \mu_k\}$

定理 8.1. K-means++ 算法给出的损失 L 与最优解 L_{opt} 满足

$$\mathbb{E}[L] \leqslant 8(\ln k + 2) L_{opt}$$

8.2 Dimensionality Reduction

wlw 不讲了.

9 Online Learning

在在线学习的设定下, 数据是以流的形式给出的, 在每次得到一个数据点之后, 都需要以恰当的方式更新预测器, 以优化将来的预测.

相比监督学习, 在线学习主要区别在于: **(1)** 不再区分 training 与 test, **(2)** 没有对数据的分布假设, 因而不存在 generalization 的概念. 相应的, mistake model 以及 regret 的概念会被用于衡量在线学习算法的表现效果.

9.1 Online Learning with Expert Advice

有 n 位专家. 预测会持续 T 轮, 每轮中每位专家都会给出各自的预测 $y_{\tilde{t},i} \in \{0,1\}$, 学习者需要根据此前得到的所有信息给出预测 $\tilde{y}_t \in \{0,1\}$, 同时也会获得正确结果 $y_t \in \{0,1\}$. 学习者的目标是让自己的预测结果与最好的专家尽量接近, 即最小化 $\sum_{t=1}^{T} \mathbb{1}[\tilde{y}_t \neq y_t]$ 与 $\min_{i\in[n]} \sum_{t=1}^{T} \mathbb{1}[y_{\tilde{t},i} \neq y_t]$ 的差 (这就是 regret).

9.1.1 Weighted Majority Vote

Algorithm 4 Weighted Majority Vote

Initialize $w_{1,i} \leftarrow 1, \forall i \in [n]$
Choose parameter $\beta \in (0,1)$
for $t = 1 \to T$ **do**
    Make the Weighted Majority Vote $\tilde{y}_t = \begin{cases} 0, & \sum_{y_{\tilde{t},i}=0} > \sum_{y_{\tilde{t},i}=1} \\ 1, & \text{otherwise} \end{cases}$
    if $\tilde{y}_t = y_t$ **then**
        $w_{t+1,i} \leftarrow w_{t,i}, \forall i \in [n]$
    else
        $w_{t+1,i} \leftarrow \begin{cases} \beta \cdot w_{t,i}, & y_{\tilde{t},i} \neq y_t \\ w_{t,i}, & y_{\tilde{t},i} = y_t \end{cases}, \forall i \in [n]$
    end if
end for

即每轮选择 $\tilde{y}_t$ 为 n 位专家预测的加权 marjority, 如果出错了, 就把所有导致自己出错的专家的权值乘上 β 作为惩罚.

定理 9.1. 记 $L_T = \sum_{t=1}^{T} \mathbb{1}[\tilde{y}_t \neq y_t]$ 为学习者的 loss, $m_T^* = \min_{i\in[n]} \sum_{t=1}^{T} \mathbb{1}[y_{\tilde{t},i} \neq y_t]$ 为最好的专家的 loss, 则在 Weighted Majority Vote 算法下, 有

$$L_T \leqslant \frac{m_T^* \log(1/\beta) + \log n}{\log(2/(1+\beta))}$$

证明. 注意到 **(1)** T 轮结束后, 所有专家剩余的总权值至少还有 $\beta^{m_T^*}$, **(2)** 每次学习者出错都会导致总权值乘上不大于 $\frac{1+\beta}{2}$ 的系数, 故

$$\beta^{m_T^*} \leqslant n\left(\frac{1+\beta}{2}\right)^{L_T} \Rightarrow L_T \leqslant \frac{m_T^* \log(1/\beta) + \log n}{\log(2/(1+\beta))}$$

□

注 9.1. 考虑 $\beta \to 1$, 由 L'Hospital Rule 可知 $\frac{\log(1/\beta)}{\log(2/(1+\beta))} \to 2$, 即 Weighted Majority Vote 算法给出的最好的界中, m_T^* 前的系数至少是 2. 接下来的 Randomized Weighted Updating 算法会给出更好的界.

Algorithm 5 Randomized Weighted Updating

1: Initialize $w_{1,i} \leftarrow 1, \forall i \in [n]$
2: Choose parameter $\beta \in [\frac{1}{2}, 1)$
3: **for** $t = 1 \to T$ **do**
4: Chooes $\tilde{y_t} = y_{\tilde{t},i}$ with probability proportional to $w_{t,i}$
5: $w_{t+1,i} \leftarrow \begin{cases} \beta \cdot w_{t,i}, & y_{\tilde{t},i} \neq y_t \\ w_{t,i}, & y_{\tilde{t},i} = y_t \end{cases}, \forall i \in [n]$
6: **end for**

9.1.2 Randomized Weighted Updating

定理 9.2. 在 Randomized Weighted Updating 算法下, 有

$$\mathbb{E}[L_T] \leqslant (2-\beta)m_T^* + \frac{\ln n}{1-\beta}$$

证明. 注意到权值的更新无关于每轮有没有答错, 因此 $\mathbb{1}[\tilde{y_t} \neq y_t]$ 是独立随机变量.

第 i 轮结束后, 总权值的变化一定是 $W \to W(1-(1-\beta)\mathbb{P}[\tilde{y_t} \neq y_t])$, 由于 $\mathbb{E}[L_T] = \sum_{t=1}^{T} \mathbb{P}[\tilde{y_t} \neq y_t]$, 因此

$$\beta^{m_T^*} \leqslant n\prod_{t=1}^{T}(1-(1-\beta)\mathbb{P}[\tilde{y_t} \neq y_t]) \leqslant n\prod_{t=1}^{T} \mathrm{e}^{-(1-\beta)\mathbb{P}[\tilde{y_t} \neq y_t]} = n\mathrm{e}^{-(1-\beta)\mathbb{E}[L_T]}$$

从而得到了

$$\mathbb{E}[L_T] \leqslant \frac{\ln(1/\beta)m_T^* + \ln n}{1-\beta}$$

只需要进一步证明 $\frac{\ln(1/\beta)}{1-\beta} \leqslant 2-\beta$. 考虑函数 $f(\beta) = \ln\beta + (1-\beta)(2-\beta)$, $f'(\beta) = \frac{(1-\beta)(1-2\beta)}{\beta}$, 当 $\beta \in [\frac{1}{2}, 1)$ 时恒有 $f'(\beta) \leqslant 0$, 从而 $f(\beta) \geqslant f(1) = 0$, 说明了 $\ln(1/\beta) \leqslant (1-\beta)(2-\beta)$, $\frac{\ln(1/\beta)}{1-\beta} \leqslant 2-\beta$. □

9.1.3 Hedge Algorithm

我们再提出一种叫做 Hedge Algorithm 的算法, 它其实只是 Randomized Weighted Updating 的推广, 但这个结果可以为后续证明定理 4.2 的工作做准备.

在 Hedge Algorithm 的设定下, loss 不再是“答错了几次”, 而是每一轮每一位专家的回答都有一个 loss $g_t(i) \in [0,1]$, 记学习者在第 t 轮的 loss 为 l_t, 则 l_t 的期望就是 n 位专家的加权平均:

$$\mathbb{E}[l_t] = \left(\sum_{i=1}^{n} w_{t,i}g_t(i)\right) \Big/ \left(\sum_{i=1}^{n} w_{t,i}\right)$$

.

Algorithm 6 Hedge Algorithm

1: Initialize $w_{1,i} \leftarrow 1, \forall i \in [n]$
2: Choose parameter $\beta \in (0,1)$
3: **for** $t = 1 \to T$ **do**
4: Chooes $i_t \in [n]$ with probability proportional to $w_{t,i}$, and obtain the loss $l_t = g_t(i_t)$
5: $w_{t+1,i} \leftarrow w_{t,i} \cdot \beta^{g_t(i)}, \forall i \in [n]$
6: **end for**

定理 9.3. 重新定义 $L_T = \sum_{t=1}^{T} l_t$, 在 Hedge Algorithm 下, 有

$$\mathbb{E}[L_T] - \min_{i\in[n]} \sum_{t=1}^{T} g_t(i) = O(\sqrt{T\log n})$$

证明. 仍然注意到 l_t 是独立随机变量.

第 i 轮结束后, 总权值的变化是 $W \to W \cdot \mathbb{E}\left[\beta^{l_t}\right]$, 从而有

$$\begin{aligned}
\mathrm{e}^{-\ln(1/\beta)m_T^*} = \beta^{m_T^*} &\leqslant n\prod_{t=1}^{T}\mathbb{E}\left[\beta^{l_t}\right] = n\prod_{t=1}^{T}\mathbb{E}\left[\mathrm{e}^{-\ln(1/\beta)l_t}\right] \\
&\leqslant n\prod_{t=1}^{T}\mathbb{E}\left[1-\ln(1/\beta)l_t+\ln^2(1/\beta)l_t^2\right] \\
&\leqslant n\prod_{t=1}^{T}\left(1-\ln(1/\beta)\mathbb{E}[l_t]+\ln^2(1/\beta)\right) \\
&\leqslant n\prod_{t=1}^{T}\mathrm{e}^{-\ln(1/\beta)\mathbb{E}[l_t]+\ln^2(1/\beta)} \\
&= n\mathrm{e}^{-\ln(1/\beta)\mathbb{E}[L_T]+T\ln^2(1/\beta)}
\end{aligned}$$

其中 $m_T^* = \min_{i\in[n]}\sum_{t=1}^{T} g_t(i)$. 两边取对数得到

$$\mathbb{E}[L_T] - \min_{i\in[n]}\sum_{t=1}^{T} g_t(i) \leqslant \frac{\ln n}{\ln(1/\beta)} + T\ln(1/\beta) \leqslant 2\sqrt{T\ln n} = O(\sqrt{T\log n})$$

□

9.2 Proof of Minimax Theorem via Online Learning

在 Game Theory 一章中, 我们陈述了 Minimax Theorem (定理 4.2), 其表明在混合策略的双人零和博弈下, 先后手并不会影响博弈的最终结果. 接下来我们利用在线学习的技术来证明这个结论.

$$\min_p \max_q p^{\mathrm{T}} M q = \max_q \min_p p^{\mathrm{T}} M q$$

9.2.1 The $\geqslant$ Direction

这个方向的结论应该是平凡的, 直观上来说就是“后手总不劣于先手”.

形式化地, 记 $p^* = \arg\min_p \max_q p^{\mathrm{T}} M q$ 为 row player 后手时选择的最优的 p, $q^* = \arg\max_q \min_p p^{\mathrm{T}} M q$ 为 column player 后手时选择的最优的 q, 则

$$\min_p \max_q p^{\mathrm{T}} M q = \max_q p^{*\mathrm{T}} M q \geqslant p^{*\mathrm{T}} M q^* \geqslant \min_p p^{\mathrm{T}} M q^* = \max_q \min_p p^{\mathrm{T}} M q$$

9.2.2 The $\leqslant$ Direction

把 row player 当作在线学习中的学习者, column player 则对应 adversary, 收益矩阵 M 的 m 行分别是一位专家.

在第 t 轮中, 学习者选择列向量 p_t 满足 $(p_t)_i = \frac{w_{t,i}}{\sum_{i=1}^{m} w_{t,i}}$, 其中 $w_{t,i}$ 表示第 t 轮时第 i 位专家的权值. 给出了 p_t 后, adversary 可以很容易地给出 $q_t = \max_q p_t^{\mathrm{T}} M q$. 第 i 位专家建议选第 i 行, 他这样的方案对应的 loss

是 $g_t(i) = (Mq_t)_i$. 显然学习者此时的 loss 的期望恰好等于 m 为专家各自损失的加权平均, 即

$$\mathbb{E}[l_t] = \left(\sum_{i=1}^{n} w_{t,i} g_t(i)\right) \Big/ \left(\sum_{i=1}^{n} w_{t,i}\right) = p_t^{\mathrm{T}} M q_t$$

由 Hedge Algorithm 以及定理 9.3, 我们知道了

$$\mathbb{E}[L_T] - \min_{i\in[n]} \sum_{t=1}^{T} g_t(i) = \sum_{t=1}^{T} p_t^{\mathrm{T}} M q_t - \min_{i\in[n]} \left(M \sum_{t=1}^{T} q_t\right)_i \leqslant O(\sqrt{T \log m})$$

由此得到

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} p_t^{\mathrm{T}} M q_t &\leqslant \min_{i\in[n]} \left(M \sum_{t=1}^{T} q_t\right)_i + O\left(\sqrt{\frac{\log m}{T}}\right) \\
&= \min_{p} \left(p^{\mathrm{T}} M \left(\frac{1}{T}\sum_{t=1}^{T} q_t\right)\right) + o(1) \\
&\leqslant \max_{q} \min_{p} p^{\mathrm{T}} M q + o(1)
\end{aligned}$$

(其中 $O\left(\sqrt{\frac{\log m}{T}}\right) = o(1)$ 因为我们视 m 为常数) 而又注意到

$$\min_{p} \max_{q} p^{\mathrm{T}} M q \leqslant \max_{q} \left(\frac{1}{T}\sum_{t=1}^{T} p_t^{\mathrm{T}}\right) M q \leqslant \frac{1}{T}\sum_{t=1}^{T} \max_{q} p_t^{\mathrm{T}} M q = \frac{1}{T}\sum_{t=1}^{T} p_t^{\mathrm{T}} M q_t$$

因此 $\min_p \max_q p^{\mathrm{T}} M q \leqslant \max_q \min_p p^{\mathrm{T}} M q + o(1)$, 即 $\min_p \max_q p^{\mathrm{T}} M q \leqslant \max_q \min_p p^{\mathrm{T}} M q$.