重要性加权主动学习

Importance Weighted Active Learning

孟庆鑫

1 分类器空间

考虑对无标注样本集 $\mathcal{X}$ 二分类. $\mathcal{X}$ 配 $\sigma-$ 代数 $\mathcal{B}_{\mathcal{X}}$ 使 $(\mathcal{X},\mathcal{B}_{\mathcal{X}})$ 为 Borel 空间, 并配概率测度 $\mathcal{P}$(以下均考虑可测情况). 考虑 $\mathcal{X}$ 的子集 $X\in\mathcal{B}_{\mathcal{X}}$, 分类器集 $\mathcal{M}$ 足够大, 大到任何一种 $\mathcal{X}$ 的二分类标签情况 X, 都对应一簇分类器 $M\subset\mathcal{M}$, 使得 $\forall m\in M$, $\forall x\in X$, $m(x)=1$, $\forall s\notin X$, $m(s)=0$, 这样 $X\leftrightarrow M$ 是一一的. 记 $\mathcal{C}$ 为所有 M 构成的集合, $\mathcal{C}$ 即为 $\mathcal{M}$ 在给定 $(\mathcal{X},\mathcal{B}_{\mathcal{X}})$ 下的等价类集.

$\forall X,X'\in\mathcal{B}_{\mathcal{X}}$, 注意到 $(\mathcal{X},\mathcal{B}_{\mathcal{X}})$ 是带有测度 $\mathcal{P}$ 的, 于是可以定义伪度量

$$\rho(X,X')=\mathcal{P}(X\,\Delta\,X')=\mathcal{P}(X\cup X'\setminus X\cap X'), \tag{1}$$

如果 $\forall X(\neq\emptyset)\in\mathcal{B}_{\mathcal{X}}$, $\mathcal{P}(X)>0$, 则可以验证 (1) 式即为 $\mathcal{B}_{\mathcal{X}}$ 的度量. 又由于 $X\leftrightarrow M$ 是一一的, 可以构造 $\mathcal{C}$ 上的伪度量

$$d(M,M')=\rho(X,X')=\mathcal{P}(x\in\mathcal{X}:m(x)\neq m'(x),m\in M,m\in M'), \tag{2}$$

$(\mathcal{B}_{\mathcal{X}},\rho)$ 与 $(\mathcal{C},d)$ 同构. d 也可以看作是 $\mathcal{M}$ 上的伪度量.

记 $T:\mathcal{C}\to\mathcal{B}_{\mathcal{X}}$ 为 $X\leftrightarrow M$ 的一一映射, 注意到 (1) 式的对称差可以扩展到一组集合上, 因此可以将伪度量扩展到 $C\subset\mathcal{C}$ 上, 即把 (2) 式扩展为

$$d(C)=\mathcal{P}\left(\bigcup_{X\in TC}X\setminus\bigcap_{X\in TC}X\right)=\mathcal{P}(x\in\mathcal{X}:\exists\, m_1,m_2\in C\ \text{s.t.}\ m_1(x)\neq m_2(x)). \tag{3}$$

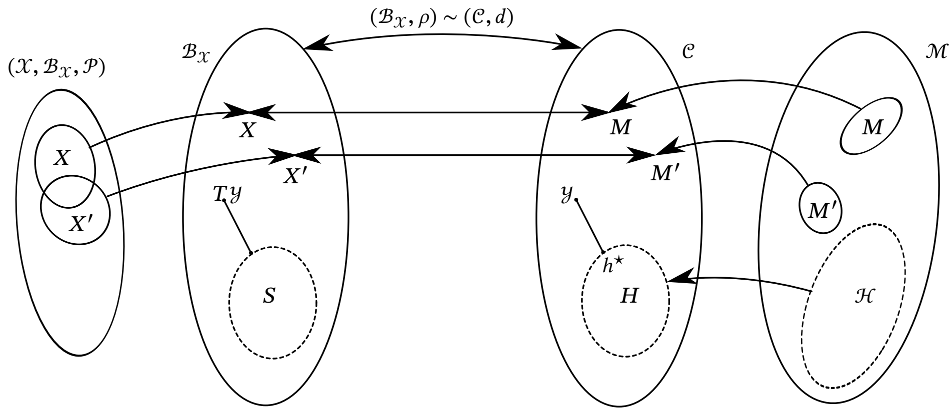
如果选取分类器空间 $\mathcal{H}\subset\mathcal{M}$, 对 $V\subset\mathcal{H}$, 改写 (3) 式有

$$d(V)=\mathcal{P}(x\in\mathcal{X}:\exists\, h_1,h_2\in V\ \text{s.t.}\ h_1(x)\neq h_2(x)). \tag{4}$$

对于给定的 $V \subset \mathcal{H}$, (4) 式叫做 V 的分歧率. 若记 $d(V) = \mathcal{P}(\Delta V)$, 则

$$\Delta V = \{x \in \mathcal{X} : \exists\, h_1, h_2 \in V \text{ s.t. } h_1(x) \neq h_2(x)\}$$

为 V 的分歧域.



如上图, 对于确定的标签 $\mathcal{Y}$, 有 $\mathcal{Y} \in \mathcal{C}, T\mathcal{Y} \in \mathcal{B}_{\mathcal{X}}, \forall\, h \in \mathcal{H}$,

$$d(h, \mathcal{Y}) = \mathcal{P}(x \in \mathcal{X} : h(x) \neq y), \tag{5}$$

对分类器空间 $\mathcal{H}$, 其等价类集 $H, S = TH$,

$$d(H, \mathcal{Y}) = \inf_{h \in \mathcal{H}} d(h, \mathcal{Y}) = \inf_{h \in \mathcal{H}} \mathcal{P}(x \in \mathcal{X} : h(x) \neq y), \tag{6}$$

(5,6) 式中的 $y = \mathcal{Y}(x) \in \{-1, +1\}$ 是确定的. 如果考虑 y 也是随机变量, 那么需要在 $\mathcal{B}_{\mathcal{X}} \otimes 2^{\mathcal{Y}}$ 上配相容的概率测度 $\mathcal{D}$, (5) 式变为 h 的错误率

$$\mathrm{er}(h) = \mathcal{D}((x, y) : h(x) \neq y), \ \forall\, h \in \mathcal{H}, \tag{7}$$

(6) 式变为 $\mathcal{H}$ 的噪声率

$$\nu_{\mathcal{H}} = \inf_{h \in \mathcal{H}} \mathrm{er}(h) = \inf_{h \in \mathcal{H}} \mathcal{D}((x, y) : h(x) \neq y),$$

记 $h^\star = \arg\inf_{h \in \mathcal{H}} \mathrm{er}(h)$, 有 $h^\star \in \overline{H}, \nu_{\mathcal{H}} = \mathrm{er}(h^\star)$, 即 $h^\star$ 是 $\mathcal{Y}$ 对 $\overline{H}$ 的最佳投影. $\varepsilon-$ 极小集 $\mathcal{H}(\varepsilon) \triangleq \{h \in \mathcal{H} : \mathrm{er}(h) - \nu_{\mathcal{H}} < \varepsilon\}$.

有了 (2) 式的伪度量, 可以定义 $\varepsilon-$ 开球

$$B(h, \varepsilon) = \{g \in \mathcal{H} : d(h, g) < \varepsilon\},$$

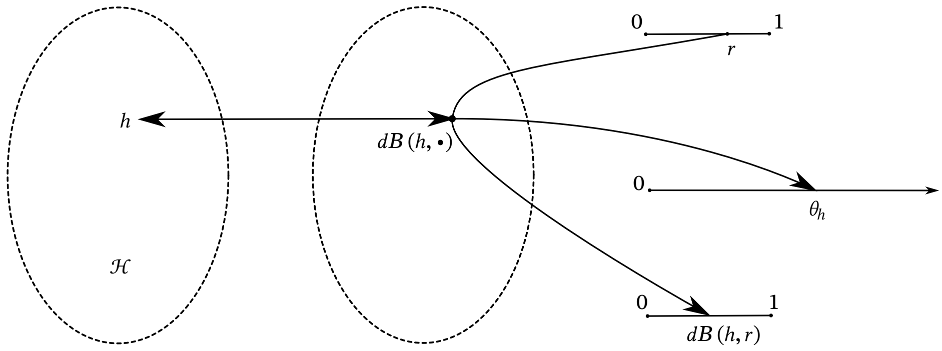
显然 $B(h, \varepsilon) \subset \mathcal{H}$, 并有分歧域 $\Delta B(h, \varepsilon)$, 分歧率 $dB(h, \varepsilon)$.

注意到每给一个 $h \in \mathcal{H}$, 都对应一个算子 (函数) $dB(h, \bullet) : [0, 1] \to [0, 1]$, 算子

$dB(h,\bullet)$ 的范数为

$$\|dB(h,\bullet)\| = \sup_{r\in[0,1]} \frac{dB(h,r)}{r}, \tag{8}$$

将 (8) 式定义为 h 的分歧系数 $\theta_h = \|dB(h,\bullet)\|$, 特殊的, 记 $\theta = \theta_{h^\star}$. 映射关系如下.



$\mathcal{H}$ 的半径 $R(\mathcal{H})$ 定义为

$$R(\mathcal{H}) = \sup_{h\in\mathcal{H}} d\left(h,h^\star\right) \overset{\text{显然}}{=\!=} \inf\left\{\varepsilon : \mathcal{H} = B\left(h^\star,\varepsilon\right)\right\}.$$

至此已构造了分类器空间 $\mathcal{H}$ 上的结构量, 下面将这些量一般化. 注意到 (7) 式

$$\mathcal{D}\left((x,y) : h(x) \neq y\right) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathbb{1}\left(h(x)\neq y\right),$$

而 $\mathbb{1}(h(x)\neq y)$ 是一种特殊的损失函数, 实际上可以取一般化的损失 $\ell(h(x),y)$, 进而将 (7) 式一般化为

$$L(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell\left(h(x),y\right),\ \forall h\in\mathcal{H}, \tag{9}$$

表示平均损失的算子记号 L 取代表示错误率的记号 er, ℓ 常取如下函数:

$$\begin{aligned}
&\ell(z,y) = (1-yz)_+, && \text{合页损失},\\
&\ell(z,y) = \ln\left(1+e^{-yz}\right), && \text{logistic 损失},\\
&\ell(z,y) = (y-z)^2 = (1-yz)^2, && \text{平方损失},\\
&\ell(z,y) = |y-z| = |1-yz|, && \text{绝对损失},
\end{aligned}$$

(9) 式可看作是对伪度量的加权, 也可看作是对原损失函数的连续松弛. 记 $h^* = \arg\inf_{h\in\mathcal{H}} L(h)$, 有 $h^*\in\overline{H}$. 主动学习的目标是找到满足

$$L(h) - L(h^*) < \epsilon \tag{10}$$

的 h, 其中 ϵ 足够小, 并尽可能少的查询标签. (10) 式对应于 L 下的 $\epsilon-$ 极小集.

从 (10) 式可看出, 可重新定义一个更广义的伪度量, 考虑 (5) 式 → (7) 式 → (9) 式

的逆过程, 不妨 $L(h_1) > L(h_2)$,

$$L(h_1) - L(h_2) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell\left(h_1(x), y\right) - \ell\left(h_2(x), y\right)\right],$$

将 y 改为变量, 保留 x 为随机变量, $\ell(h_1(x), y) - \ell(h_2(x), y)$ 为一随机过程, 将其视作 y 的函数, 其度量可取为

$$\max_y \left|\ell\left(h_1(x), y\right) - \ell\left(h_2(x), y\right)\right|,$$

由此可定义 $\mathcal{H}$ 在一般损失 ℓ 下的伪度量

$$\varrho(h_1, h_2) = \mathop{\mathbb{E}}_{x\sim\mathcal{P}} \max_y \left|\ell\left(h_1(x), y\right) - \ell\left(h_2(x), y\right)\right|, \tag{11}$$

对 $V \subset \mathcal{H}$, 分歧率

$$\varrho(V) = \mathop{\mathbb{E}}_{x\sim\mathcal{P}} \sup_{g,\, h\in V} \sup_y \left|\ell\left(g(x), y\right) - \ell\left(h(x), y\right)\right|,$$

$\varepsilon-$ 开球

$$B_\varrho(h, \varepsilon) = \{g \in \mathcal{H} : \varrho(h, g) < \varepsilon\},$$

以及分歧系数

$$\begin{aligned} \vartheta_h &= \left\|\varrho B_\varrho(h, \bullet)\right\| = \sup_r \frac{\varrho B_\varrho(h, r)}{r} \\ &= \sup_r \frac{\mathbb{E}_{x\sim\mathcal{P}} \sup_{g_1,\, g_2\in B_\varrho(h,r)} \sup_y \left|\ell\left(g_1(x), y\right) - \ell\left(g_2(x), y\right)\right|}{r}, \end{aligned} \tag{12}$$

中心分歧系数

$$\begin{aligned} \dot{\vartheta}_h &= \left\|\varrho\left(B_\varrho(h, \bullet), h\right)\right\| = \sup_r \frac{\varrho\left(B_\varrho(h, r), h\right)}{r} \\ &= \sup_r \frac{\mathbb{E}_{x\sim\mathcal{P}} \sup_{g\in B_\varrho(h,r)} \sup_y \left|\ell\left(g(x), y\right) - \ell\left(h(x), y\right)\right|}{r}, \end{aligned} \tag{13}$$

显然有 $\vartheta_h \leqslant 2\dot{\vartheta}_h$. 特殊的, 记 $\vartheta = \vartheta_{h^*}$, $\dot{\vartheta} = \dot{\vartheta}_{h^*}$.

2 IWAL

算法 1 描述了重要性加权主动学习 (IWAL) 的基本框架. 当见到 x_t 后, 算法调用子程序 rejection-threshold[算法 2], 查询历史并返回是否请求标签 y_t 的概率 p_t. 如果 y_t 最终被查询, 则把它的权值设为 $1/p_t$.

Algorithm 1 : Importance Weighted Active Learning

$S_0 \longleftarrow \varnothing$,
for $t = 1 : T$ **do**
  receive x_t,
  $p_t \longleftarrow \text{rejection-threshold}\,(x_t, \{x_i, y_i, p_i, Q_i : 1 \leqslant i < t\})$,
  $Q_t \longleftarrow \mathbb{1}\,(\text{rand}\,(0,1) < p_t)$,
  if $Q_t = 1$, **then**
    request y_t, **and** $S_t \longleftarrow S_{t-1} \cup \left\{\left(x_t, y_t, \frac{1}{p_t}\right)\right\}$,
  else
    $S_t \longleftarrow S_{t-1}$,
  $h_t \longleftarrow \arg\min_{h\in\mathcal{H}} \sum_t \frac{1}{p_t}\,\ell\,(h\,(x_t)\,, y_t)$.

T 步时损失 L 的重要性加权估计为

$$L_T\,(h) = \frac{1}{T}\sum_{t=1}^{T}\frac{Q_t}{p_t}\,\ell\,(h\,(x_t)\,, y_t)\,,$$

由

$$\mathbb{E}\,L_T\,(h) = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{E}\,Q_t}{p_t}\;\mathbb{E}\,\ell\,(h\,(x_t)\,, y_t) = \mathbb{E}\,\ell\,(h\,(x)\,, y) = \mathbb{E}\,L\,(h)$$

可知, $L_T\,(h)$ 是 $L\,(h)$ 的无偏估计.

一个理想的学习算法具有一致性, 即给定无限的未标记和已标记样本, 算法将收敛到最优估计. 下面的定理 1 指出 IWAL 是一致的.

定理 1 $\forall\,\mathcal{D},\ \forall\,\mathcal{H} : |\mathcal{H}| < \infty,\ \forall\,\delta > 0,\ \exists\, p_{\min} > 0$ s.t. $1 \leqslant \forall\, t \leqslant T,\ p_t \geqslant p_{\min}$, 则

$$\mathbb{P}\left(\max_{h\in\mathcal{H}}|L_T\,(h) - L\,(h)| > \frac{\sqrt{2}}{p_{\min}}\sqrt{\frac{1}{T}\left(\ln|\mathcal{H}| + \ln\frac{2}{\delta}\right)}\right) < \delta.$$

证明: 任选 $h \in \mathcal{H}$, 注意到 $h\,(x)$ 有界, 于是总可以选取 $\ell\,(h\,(x)\,, y)$ 使其映射到 $[0,1]$. 考察随机序列 $U_1, \ldots, U_T$, 其中

$$U_t = \frac{Q_t}{p_t}\,\ell\,(h\,(x_t)\,, y_t) - L\,(h)\,,$$

且有

$$|U_t| = \frac{1}{p_t}\left|\,Q_t\,\ell\,(h\,(x_t)\,, y_t) - p_t L\,(h)\right| \leqslant \frac{1}{p_t} \leqslant \frac{1}{p_{\min}},$$

记 $Z_t = \sum_{i=1}^{t} U_i$, $Z_0 = 0,\ 1 \leqslant \forall t \leqslant T$,

$$\begin{aligned}\mathbb{E}\left[Z_t \mid Z_{t-1}, \cdots, Z_0\right] &= \mathop{\mathbb{E}}_{Q_t, x_t, y_t, p_t}\left[U_t + Z_{t-1} \mid Z_{t-1}, \cdots, Z_0\right] \\ &= Z_{t-1} + \mathop{\mathbb{E}}_{Q_t, x_t, y_t, p_t}\left[\frac{Q_t}{p_t}\ell\left(h\left(x_t\right), y_t\right) - L\left(h\right) \middle| Z_{t-1}, \cdots, Z_0\right] \\ &= Z_{t-1},\end{aligned}$$

Z_t 是鞅, $|Z_t - Z_{t-1}| = |U_t| \leqslant 1/p_{\min}$. 注意到 $Z_T = T\left(L_T\left(h\right) - L\left(h\right)\right)$, 应用 Azuma-Hoeffding 不等式 注1, 对 $\lambda > 0$ 有

$$\mathbb{P}\left(\left|L_T\left(h\right) - L\left(h\right)\right| > \frac{\lambda}{p_{\min}\sqrt{T}}\right) = \mathbb{P}\left(\left|Z_T - Z_0\right| > \frac{\lambda\sqrt{T}}{p_{\min}}\right) \leqslant 2\mathrm{e}^{-\frac{\lambda^2}{2}},$$

令 $\lambda = \sqrt{2\left(\ln|\mathcal{H}| + \ln\frac{2}{\delta}\right)}$, 有

$$\mathbb{P}\left(\left|L_T\left(h\right) - L\left(h\right)\right| > \frac{\sqrt{2}}{p_{\min}}\sqrt{\frac{1}{T}\left(\ln|\mathcal{H}| + \ln\frac{2}{\delta}\right)}\right) \leqslant \frac{\delta}{|\mathcal{H}|},$$

由 h 的任意性即得到结论. ∎

注 1 **(Azuma-Hoeffding)** 鞅 $\{Z_i\}$ 满足 $|Z_i - Z_{i-1}| \leqslant \gamma_i$ 几乎必然成立, $\forall n \in \mathbb{N}$, $\forall \lambda > 0$,

$$\mathbb{P}\left(Z_n - Z_0 \begin{array}{c}\geqslant \lambda \\ \leqslant -\lambda\end{array}\right) \leqslant \exp\left(-\frac{\lambda^2}{2\sum_{i=1}^{n}\gamma_i^2}\right).$$

3 损失加权

算法 2 给出了 IWAL算法 1 中子程序 rejection-threshold 的一个实例.

Algorithm 2 : loss-weighting $(x_t, \{x_i, y_i, p_i, Q_i : i < t\})$

1: $\mathcal{H}_0 \longleftarrow \mathcal{H}$,

2: $\mathcal{H}_t \longleftarrow \left\{h \in \mathcal{H}_{t-1} : L_{t-1}\left(h\right) \leqslant \min_{h \in \mathcal{H}_{t-1}} L_{t-1}\left(h\right) + \Delta_{t-1}\right\}$,

3: **return** $p_t \longleftarrow \max_{h_1, h_2 \in \mathcal{H}_t} \max_{y} \ell\left(h_1\left(x_t\right), y\right) - \ell\left(h_2\left(x_t\right), y\right)$.

算法的核心思想是缩小模型空间 $\mathcal{H}$ 使其所有点都落在当前最小损失估计的 Δ 邻域内. 注意到 $\mathbb{E}_{x\sim\mathcal{P}}\, p_t\left(x\right) = \varrho\left(\mathcal{H}_t\right)$, 若 x 导致 p_t 越大, 则对应的 $\mathcal{H}_t$ 的分歧率就越大, 就越有大概率查询 x 的标签. 由于总可以使 ℓ 的值被限定在 $[0, 1]$ 区间, 因此 p_t 可看作概率.

我们曾构造了 (11) 式的伪度量. 下面的定理 2 与这个伪度量存在关联. 为了说明定理 2, 先从引理 1 开始.

引理 1 $\forall \mathcal{D},\ \forall \mathcal{H},\ \forall \delta > 0,\ \forall f, g \in \mathcal{H}_T$, 选择合适的 Δ_T, 有

$$\mathbb{P}\left(\left|L_T(f) - L_T(g) - L(f) + L(g)\right| > \Delta_T\right) < \delta.$$

证明: 任取 T, 任选 $f, g \in \mathcal{H}_T \subset \mathcal{H}_{T-1} \subset \cdots \subset \mathcal{H}_1 = \mathcal{H}$, 参考定理 1 的证明, 令

$$U_t = \left(\frac{Q_t}{p_t}\ell\left(f(x_t), y_t\right) - L(f)\right) - \left(\frac{Q_t}{p_t}\ell\left(g(x_t), y_t\right) - L(g)\right),$$

由于 $p_t = \max\limits_{h_1, h_2 \in \mathcal{H}_t} \max\limits_{y} \ell\left(h_1(x_t), y\right) - \ell\left(h_2(x_t), y\right)$, 有 $p_t \geqslant \left|\ell\left(f(x_t), y_t\right) - \ell\left(g(x_t), y_t\right)\right|$,

$$|U_t| \leqslant \frac{Q_t}{p_t}\left|\ell\left(f(x_t), y_t\right) - \ell\left(g(x_t), y_t\right)\right| + \left|L(f) - L(g)\right| \leqslant 2,$$

记 $Z_t = \sum\limits_{i=1}^{t} U_i,\ Z_0 = 0,\ 1 \leqslant \forall t \leqslant T$,

$$\begin{aligned}
&\mathbb{E}\left[Z_t \mid Z_{t-1}, \cdots, Z_0\right] \\
=\ &\mathop{\mathbb{E}}_{Q_t, x_t, y_t, p_t}\left[U_t + Z_{t-1} \mid Z_{t-1}, \cdots, Z_0\right] \\
=\ &Z_{t-1} + \mathop{\mathbb{E}}_{Q_t, x_t, y_t, p_t}\left[\left(\frac{Q_t}{p_t}\ell\left(f(x_t), y_t\right) - L(f)\right) - \left(\frac{Q_t}{p_t}\ell\left(g(x_t), y_t\right) - L(g)\right)\middle| Z_{t-1}, \cdots, Z_0\right] \\
=\ &Z_{t-1},
\end{aligned}$$

Z_t 是鞅, $|Z_t - Z_{t-1}| = |U_t| \leqslant 2$. 应用 Azuma-Hoeffding 不等式 [注1], 有

$$\mathbb{P}\left(\left|L_T(f) - L_T(g) - L(f) + L(g)\right| > \Delta_T\right) = \mathbb{P}\left(\left|Z_T - Z_0\right| > T\Delta_T\right) \leqslant 2\mathrm{e}^{-\frac{T\Delta_T^2}{8}},$$

只需要选取 $\Delta_T > \sqrt{\frac{8}{T}\ln\frac{2}{\delta}}$ 即可. ∎

在算法 2 中,

$$\Delta_t = \sqrt{\frac{8}{t}\ln\frac{2t(t+1)\left|\mathcal{H}\right|^2}{\delta}}.$$

定理 2 $\forall \mathcal{D},\ \forall \mathcal{H},\ h^* \in \mathcal{H},\ \forall \delta > 0,\ \forall f, g \in \mathcal{H}_T,\ \forall T \geqslant 1$, 都有 $h^* \in \mathcal{H}_T$, 且

$$\mathbb{P}\left(L(f) - L(g) > 2\Delta_{T-1}\right) < \delta.$$

特殊的, 若 $h_T = \arg\min\limits_{h \in \mathcal{H}_T} L_T(h)$, 则 $L(h_T) - L(h^*) \leqslant 2\Delta_{T-1}$ 依概率不小于 $1 - \delta$ 成立.

证明: 保持引理 1 的概率事件不变. $h^* \in \mathcal{H} = \mathcal{H}_0 = \mathcal{H}_1$, 若 $h^* \in \mathcal{H}_T$, 由引理 1, $L_T(h^*) - L_T(h_T) \leqslant L(h^*) - L(h_T) + \Delta_T \leqslant \Delta_T$, 即 $L_T(h^*) \leqslant L_T(h_T) + \Delta_T$, 由算法 2 的第 2 行可知 $h^* \in \mathcal{H}_{T+1}$. 归纳可知 $\forall T \geqslant 1$, 都有 $h^* \in \mathcal{H}_T$.
由于 $f, g \in \mathcal{H}_T \subset \mathcal{H}_{T-1}$, 再由引理 1, $L(f) + L(g) \leqslant L_{T-1}(f) - L_{T-1}(g) + \Delta_{T-1}$, 联合 $L_{T-1}(f) \leqslant L_{T-1}(h_{T-1}) + \Delta_{T-1}$, $L_{T-1}(h_{T-1}) \leqslant L_{T-1}(g) + \Delta_{T-1}$ 即得证. ∎

4 标签复杂度

定义损失函数 ℓ 的非对称斜率

$$K_\ell = \sup_{z,z'} \frac{\max_y |\ell(z,y) - \ell(z',y)|}{\min_y |\ell(z,y) - \ell(z',y)|}, \tag{14}$$

$K_{\mathbb{1}} = 1$, $K_{(1-yz)_+} = \infty$. 对 (14) 式放缩, 有

$$K_\ell \geqslant \frac{\max_y |\ell(z,y) - \ell(z',y)|}{\min_y |\ell(z,y) - \ell(z',y)|} \geqslant \frac{\max_y |\ell(z,y) - \ell(z',y)|}{|\ell(z,y) - \ell(z',y)|},$$

即 $K_\ell |\ell(z,y) - \ell(z',y)| \geqslant \max_y |\ell(z,y) - \ell(z',y)|$, 两侧同时用 $\mathbb{E}_{(x,y)\sim\mathcal{D}}$ 作用,

$$\begin{aligned}\varrho(z,z') = \mathop{\mathbb{E}}_{x\sim\mathcal{P}} \max_y |\ell(z,y) - \ell(z',y)| &\leqslant K_\ell \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} |\ell(z,y) - \ell(z',y)| \\ &\leqslant K_\ell \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \big(\ell(z,y) + \ell(z',y)\big) = K_\ell\big(L(z) + L(z')\big),\end{aligned}$$

上式用到了 (11) 式, 做 $z \mapsto h$, $z' \mapsto h^*$ 替换, 便有下面的引理 2:

引理 2 $\forall \mathcal{D}$, $\forall h \in \mathcal{H}$, $\forall \ell$, $\varrho(h,h^*) \leqslant K_\ell\big(L(h) + L(h^*)\big)$.

根据算法 2 下的解释内容, $p_t(x)$ 是第 t 轮查询一个样本的概率, 对样本 x, 其出现的概率为 $\mathcal{P}(x)$, 查询次数为 $\mathcal{P}(x)\, p_t(x)$, 因此对样本集 $\mathcal{X}$, T 轮内查询次数期望为

$$\sum_{t=1}^{T} \sum_x \mathcal{P}(x)\, p_t(x) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{x\sim\mathcal{P}} p_t(x) = \sum_{t=1}^{T} \varrho(\mathcal{H}_t), \tag{15}$$

$\forall h \in \mathcal{H}_t$, 根据定理 2, $L(h) \leqslant L(h^*) + 2\Delta_{t-1}$, 根据引理 2, $\varrho(h,h^*) \leqslant K_\ell\big(L(h) + L(h^*)\big)$, 于是有 $\varrho(h,h^*) \leqslant 2K_\ell\big(L(h^*) + \Delta_{t-1}\big)$, 注意到 h 取遍 $\mathcal{H}_t$ 中的点, 这说明 $\mathcal{H}_t \subset \overline{B}_\varrho(h^*, r)$, 其中 $r = 2K_\ell\big(L(h^*) + \Delta_{t-1}\big)$, 因此有下面的不等式:

$$\varrho(\mathcal{H}_t) = \mathop{\mathbb{E}}_{x\sim\mathcal{P}} \sup_{g,\, h\in\mathcal{H}_t} \sup_y |\ell(g,y) - \ell(h,y)| \leqslant \mathop{\mathbb{E}}_{x\sim\mathcal{P}} \sup_{g,\, h\in\overline{B}_\varrho(h^*, r)} \sup_y |\ell(g,y) - \ell(h,y)| \leqslant \vartheta r,$$

放缩过程使用了 (12) 式的定义, 因为比 (13) 式的结果更紧. 代入 (15) 式,

$$\sum_{t=1}^{T} \varrho(\mathcal{H}_t) \leqslant 2\vartheta K_\ell \sum_{t=1}^{T} \left(L(h^*) + \Delta_{t-1}\right) = 2\vartheta K_\ell \left(L(h^*)\,T + \sum_{t=1}^{T-1} \sqrt{\frac{8}{t} \ln \frac{2t(t+1)|\mathcal{H}|^2}{\delta}}\right),$$

我们更关注的是渐近上界, 并注意到 Δ_t 作为 t 的函数, 在 t 很大的时候是单调递减的, 但是级数和却随着 T 的增加而发散, 因此不必在乎级数的前有限项,

$$\begin{aligned}
O\left(\sum_{t=1}^{T-1} \sqrt{\frac{8}{t} \ln \frac{2t(t+1)|\mathcal{H}|^2}{\delta}}\right) &= O\left(\sum_{t=1}^{T-1} \sqrt{\frac{1}{t} \ln \frac{2t^2|\mathcal{H}|^2}{\delta}}\right) \\
&\leqslant O\left(\int_1^T \sqrt{\frac{1}{t} \ln \frac{2t^2|\mathcal{H}|^2}{\delta}}\,\mathrm{d}t\right) \qquad t \mapsto \frac{\sqrt{\delta}}{\sqrt{2}\,|\mathcal{H}|} s \\
&= O\left(\int_1^S \sqrt{\frac{\ln s}{s}}\,\mathrm{d}s\right) \\
&= O\left(2\sqrt{S \ln S} + \sqrt{2\pi}\, i\, \mathrm{erf}\left(i\sqrt{\frac{\ln S}{2}}\right)\right) \\
&= O\left(\sqrt{T \ln \frac{|\mathcal{H}|\,T}{\delta}}\right),
\end{aligned}$$

至此, 得到如下的标签复杂度上界定理:

定理 3 $\forall \mathcal{D}, \forall \mathcal{H}, \forall \ell$ 且 $\exists K_\ell, \exists \vartheta$, 则 $\forall \delta > 0$, 依概率不小于 $1 - \delta$, IWAL 算法 T 轮内查询次数期望不超过

$$2\vartheta K_\ell \left(L(h^*)\,T + O\left(\sqrt{T \ln \frac{|\mathcal{H}|\,T}{\delta}}\right)\right),$$

其中 $h^* = \arg\inf_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(h(x), y)$, 且期望大于随机抽样的期望.