主动学习理论概要

Steve Hanneke's "Theory of Active Learning"

Chapter 2, 4, 5

孟庆鑫

1 基本定义和符号

样本空间 $\mathcal{X}$ 配 $\sigma-$ 代数 $\mathcal{B}_{\mathcal{X}}$ 使 $(\mathcal{X},\mathcal{B}_{\mathcal{X}})$ 为 Borel 空间, $\mathcal{Y}=\{-1,1\}$ 为标签空间, 对应在 $\mathcal{X}\times\mathcal{Y}$ 上配 $\sigma-$ 代数 $\mathcal{B}=\mathcal{B}_{\mathcal{X}}\otimes 2^{\mathcal{Y}}$, 并配概率测度 $\mathcal{P}_{XY}$, 记 $\mathcal{P}$ 为 $\mathcal{P}_{XY}$ 对 $\mathcal{X}$ 的边缘分布, 以下均考虑可测情况.

$\eta(x)\triangleq\mathbb{P}(Y=+1|X=x),\forall x\in\mathcal{X},(X,Y)\sim\mathcal{P}_{XY}$,

h 的错误率 $\operatorname{er}(h)\triangleq\mathcal{P}_{XY}((x,y):h(x)\neq y),\forall h:\mathcal{X}\to\mathcal{Y}$,

$\mathcal{Z}\triangleq\{(X_i,Y_i)\}_{i=1}^{\infty}\overset{iid}{\sim}\mathcal{P}_{XY}$, 前 m 截断 $\mathcal{Z}_m\triangleq\{(X_i,Y_i)\}_{i=1}^{m}$, 未标记数据集 $\mathcal{Z}_{\boldsymbol{X}}\triangleq\{X_i\}_{i=1}^{\infty}$, 并假设 $\mathcal{Z}_{\boldsymbol{X}}$ 可被完整访问, 在主动学习算法中, 给定 n, 请求观测值 Y_{i_1}, 请求观测值 Y_{i_2}, 依此 $\cdots$, 不超过 n 次请求, 算法停止并返回 $\hat{h}$. 我们感兴趣的是算法 $\mathcal{A}:\mathcal{Z}_n\to\hat{h}$ 作为 n 的函数, 其行为特征.

定义 1 *$\forall\mathcal{A}$ 为主动学习算法, 如果 $\forall\varepsilon>0,\forall\delta\in[0,1],\forall\mathcal{P}_{XY},\forall n\geqslant\Lambda(\varepsilon,\delta,\mathcal{P}_{XY})$, $n\in\mathbb{N},(\mathcal{A},n)$ 产生了 $\hat{h}$, $\operatorname{er}\left(\hat{h}\right)\leqslant\varepsilon$ 的概率至少为 $1-\delta$, 则算法 $\mathcal{A}$ 达到了标签复杂性 Λ.*

噪声率 $\nu\triangleq\inf_{h\in\mathbb{C}}\operatorname{er}(h)$, $\mathbb{C}$ 是确定的分类器族, 我们特别关注 $\Lambda(\nu+\varepsilon,\delta,\mathcal{P}_{XY})$. 记 $f^{\star}=\arg\inf_{h\in\mathbb{C}}\operatorname{er}(h)$, 有 $f^{\star}\in\bar{\mathbb{C}}$ ($\mathbb{C}$ 的闭包), $\nu=\operatorname{er}\left(f^{\star}\right)$,

$\forall$ 分类器族 $\mathcal{H},\forall\varepsilon\in[0,1],\varepsilon-$ 极小集 $\mathcal{H}(\varepsilon)\triangleq\left\{h\in\mathcal{H}:\operatorname{er}(h)-\inf_{g\in\mathcal{H}}\operatorname{er}(g)\leqslant\varepsilon\right\}$,

$\forall$ 中心 h 的 $\varepsilon-$ 球 $B_{\mathcal{H},\mathcal{P}}(h,\varepsilon)\triangleq\{g\in\mathcal{H}:\mathcal{P}(x:g(x)\neq h(x))\leqslant\varepsilon\}$,

$\mathcal{H}$ 的半径 $R(\mathcal{H})\triangleq\sup_{h\in\mathcal{H}}\mathcal{P}\left(x:h(x)\neq f^{\star}(x)\right)\overset{\text{显然}}{=\!=}\inf\left\{\varepsilon:\mathcal{H}=B_{\mathcal{H},\mathcal{P}}\left(f^{\star},\varepsilon\right)\right\}$,

分歧域 $\mathcal{D}(\mathcal{H})\triangleq\{x\in\mathcal{X}:\exists h,g\in\mathcal{H}\text{ s.t. }h(x)\neq g(x)\}$, 一种构造算法的策略是: 从 $\mathcal{D}(\mathcal{H})$ 里抽样本, 请求观测值, 再反过来更新 $\mathcal{H}$, 分歧率 $\Delta(\mathcal{H})\triangleq\mathcal{P}(\mathcal{D}(\mathcal{H}))$.

定义 2 $\forall r_0>0, \forall h$ 对 $\mathbb{C}$ 在 $\mathcal{P}$ 下的分歧率为

$$\theta_h(r_0)=\sup_{r>r_0}\frac{\Delta(B(h,r))}{r}.$$

特殊的, 当 $h=f^\star$ 时记为 $\theta(r_0)$, 称为分类器族 $\mathbb{C}$ 在 $\mathcal{P}_{XY}$ 下的分歧率.

情况 1 $\exists a\in[1,\infty),\ \alpha\in[0,1],\ \forall h\in\mathbb{C}$,

$$\mathcal{P}\left(x: h(x)\neq f^\star(x)\right)\leq a\left(\mathrm{er}(h)-\mathrm{er}\left(f^\star\right)\right)^\alpha.$$

例 1 $\mathcal{X}=[0,1], \mathbb{C}=\left\{\mathbb{1}^{\pm}_{[z,1]}: z\in(0,1)\right\}$. 注意到 $\mathbb{C}$ 同构于 $z\in(0,1)$, 问题等价于从 $(0,1)$ 上查找分段点, 可采用类似二分查找策略. 相较于被动学习, 可以带来指数级的提升.

2 标签复杂度下界

给定分布 $\mathcal{P}$, $\mathbb{C}$ 的 $\varepsilon-$ 覆盖数 $\mathcal{N}(\varepsilon,\mathcal{P})\triangleq\min\left\{|\mathcal{H}|: \bigcup_{h\in\mathcal{H}}B(h,\varepsilon)=\mathbb{C}\right\}$.

定理 1 $\forall\mathcal{P}, \forall\Lambda$ (主动学习算法所达到的), $\forall\varepsilon>0, \exists\mathcal{P}_{XY}$ 与 $\mathcal{P}$ 相容, 且

$$\Lambda(\varepsilon,\delta,\mathcal{P}_{XY})\geqslant\left\lceil\log_2\left((1-\delta)\mathcal{N}(2\varepsilon,\mathcal{P})\right)\right\rceil.$$

考虑噪声情况的下界, 有下面的定理.

定理 2 任给 $\gamma\in(0,1), \delta\in\left(1,\frac{1}{4}\right), n\in\mathbb{N}$, 令 $p_0=\frac{1}{2}-\frac{\gamma}{2}, p_1=\frac{1}{2}+\frac{\gamma}{2}$. 给定 $\hat{t}:\{0,1\}^n\to\{0,1\}$(可能是随机的). 如果

$$n<2\left\lfloor\frac{1-\gamma^2}{2\gamma^2}\ln\left(\frac{1}{8\delta(1-2\delta)}\right)\right\rfloor,$$

则对 $t\sim\mathrm{Bernoulli}\left(\frac{1}{2}\right), B_1|t,\cdots,B_n|t\overset{iid}{\sim}\mathrm{Bernoulli}(p_t), B_1,\cdots,B_n$ 与 $\hat{t}$ 独立, 则 $\hat{t}(B_1,\cdots,B_n)\neq t$ 的概率大于 δ.

定理 3 存在全局常量 $q\in\mathbb{R}^+$ 使得: 如果 $|\mathbb{C}|\geqslant 3$, 则 $\forall\Lambda, \forall\nu\in\left(1,\frac{1}{2}\right)$, 以及充分小的 $\varepsilon,\delta>0, \exists\mathcal{P}_{XY}, \mathrm{er}\left(f^\star\right)=\nu$, 使得

$$\Lambda(\nu+\varepsilon,\delta,\mathcal{P}_{XY})\geqslant q\left(\frac{\nu^2}{\varepsilon^2}\right)\left(d+\mathrm{Log}\left(\frac{1}{\delta}\right)\right),$$

其中 $\mathrm{Log}(x)\triangleq\max\{\ln(x),1\}, x\geqslant 0$. 进一步, $\forall a\in[4,\infty), \alpha\in(0,1]$, 以及充分小的 $\varepsilon,\delta>0, \exists\mathcal{P}_{XY}$ 满足情况 *1*, 则有

$$\Lambda(\nu+\varepsilon,\delta,\mathcal{P}_{XY})\geqslant qa^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d+\mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

3 基于分歧的主动学习

3.1 CAL

Cohn, Atlas, Ladner 给出了通用主动学习算法可实现的例子:

Algorithm 1 : **CAL** (n) :

1: $m \longleftarrow 0$, $Q \longleftarrow \emptyset$
2: **while** $|Q| < n \,\&\, m < 2^n$, **do**
3: $\quad m \longleftarrow m + 1$
4: $\quad$ **if** $\forall y \in \mathcal{Y}$, $\exists h \in \mathbb{C}$ s.t. $\mathrm{er}_{Q \cup \{(X_m, y)\}}(h) = 0$, **then**
5: $\quad\quad$ request label Y_m ; $Q \longleftarrow Q \cup \{(X_m, Y_m)\}$
6: **return** any $\hat{h} \in \mathbb{C}$ s.t. $\mathrm{er}_Q\left(\hat{h}\right) = 0$

注意第 4 行的逻辑含义是: 不论对 X_m 打上何种标签, 分类器族 $\mathbb{C}$ 对 Q 形成的分布都有分类器可接受, 这意味着 X_m 的标签无法断言, 需要请求 Y_m. 算法 1 可做如下等价:

Algorithm 2 : **CAL** (n) :

1: $m \longleftarrow 0$, $t \longleftarrow 0$, $V \longleftarrow \mathbb{C}$
2: **while** $t < n \,\&\, m < 2^n$, **do**
3: $\quad m \longleftarrow m + 1$
4: $\quad$ **if** $X_m \in \mathcal{D}(V)$, **then**
5: $\quad\quad$ request label Y_m ; $V \longleftarrow \{h \in V : h(X_m) = Y_m\}$; $t \longleftarrow t + 1$
6: **return** any $\hat{h} \in V$

定理 4 *CAL 达到了标签复杂性 Λ, 使得对可实现的 $\mathcal{P}_{XY}$, $\forall \varepsilon, \delta > 0$, 有*

$$\Lambda(\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \theta(\varepsilon)\left(d\,\mathrm{Log}\left(\theta(\varepsilon)\right) + \mathrm{Log}\left(\frac{\mathrm{Log}\left(1/\varepsilon\right)}{\delta}\right)\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

符号 $\lesssim$ 在 $u(\varepsilon, \delta) \lesssim v(\varepsilon, \delta)$ 里指 $\exists c \in \mathbb{R}^+$ 且与 $\mathbb{C}, \mathcal{P}_{XY}$ 或其他问题特定变量都无关, 并满足 $u(\varepsilon, \delta) \leqslant c v(\varepsilon, \delta)$, $\forall \varepsilon, \delta \in (0, 1)$. 定理 4 与 ε 渐近相关的界为

$$\mathcal{O}\left(\theta(\varepsilon)\,\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\mathrm{Log}\left(\theta(\varepsilon)\,\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right)\right).$$

定理 5 $\forall m \in \mathbb{N} \cup \{0\}$, $\forall r \in (0, 1)$, $\mathbb{E}\left[\Delta\left(V_m^\star\right)\right] \geqslant (1 - r)^m \Delta\left(B\left(f^\star, r\right)\right)$. *进一步, 这意味着 $\forall \varepsilon \in (0, 1)$,*

$$\mathbb{E}\left[N\left(\left\lceil \frac{1}{\varepsilon} \right\rceil\right)\right] \geqslant \frac{\theta(\varepsilon)}{2}.$$

定理 6 $\forall n \in \mathbb{N}, \forall r \in (0,1), \mathbb{E}\left[\Delta\left(V^{\star}_{M(n)}\right)\right] \geqslant \Delta\left(B\left(f^{\star}, r\right)\right) - nr$. 进一步, 这意味着 $\forall n \in \mathbb{N}, \forall \varepsilon \in (0,1)$,

$$n \leqslant \frac{\theta(\varepsilon)}{2} \Longrightarrow \mathbb{E}\left[\Delta\left(V^{\star}_{M(n)}\right)\right] \geqslant \frac{\Delta\left(B\left(f^{\star}, r\right)\right)}{2}.$$

考虑

$$\left\{q_m \in \{0,1\}\right\}_{m=1}^{\infty} : q_m \perp\!\!\!\perp \mathcal{Z} \,\middle|\, \left\{(X_i, q_i Y_i)\right\}_{i=1}^{m-1}, X_m,$$
$$\left\{\hat{h}_m\right\}_{m=0}^{\infty} : \hat{h}_m \perp\!\!\!\perp \mathcal{Z} \,\middle|\, \left\{(X_i, q_i Y_i)\right\}_{i=1}^{m},$$

随机变量 q_m 的取值表示是否请求 Y_m, CAL 作为一种选择性抽样算法, 由

$$\left(\left\{q_m\right\}_{m=1}^{\infty}, \left\{\hat{h}_m\right\}_{m=0}^{\infty}\right) : \hat{h}_m \in V_m^{\star},\ q_m = \mathbb{1}_{\mathcal{D}\left(V^{\star}_{m-1}\right)}(X_m)$$

确定.

3.2 噪声情况

注意到取得最佳分类器 $f^{\star}$, 仍然有 $\nu = \mathrm{er}\left(f^{\star}\right)$, 下面考虑 CAL 带噪声的情况. 一个具体的算法 (Balcan, Beygelzimer, Langford) 是 A^2 策略的变体, 算法如下. 记 $\delta_m \triangleq \dfrac{\delta}{\log_2^2(2m)}$, $\delta \in (0,1)$, $m \in \mathbb{N}$.

Algorithm 3 : **RobustCAL**$_{\delta}(n)$:

$m \longleftarrow 0$, $i \longleftarrow 1$, $Q_i \longleftarrow \emptyset$
while $|Q_i| < n \,\&\, m < 2^n$, **do**
  $m \longleftarrow m + 1$
  if $\forall y \in \mathcal{Y}, \exists h \in \mathbb{C}$ s.t. $h(X_m) = y, \forall j < i, \dfrac{\left(\mathrm{er}_{Q_j}(h) - \mathrm{er}_j^{*}\right)|Q_j|}{U\left(2^j, \delta_{(2^j)}\right)} \leqslant 2^j$, **then**
    request label Y_m ; $Q_i \longleftarrow Q_i \cup \{(X_m, Y_m)\}$
  if $\log_2(m) \in \mathbb{N}$, **then**
    $\mathrm{er}_j^{*} \longleftarrow \min\left\{\mathrm{er}_{Q_j}(h) : h \in \mathbb{C}, \forall j < i, \left(\mathrm{er}_{Q_j}(h) - \mathrm{er}_j^{*}\right)|Q_j| \leqslant U\left(2^j, \delta_{(2^j)}\right) 2^j\right\}$;
    $i \longleftarrow i + 1$; $Q_i \longleftarrow Q_{i-1}$
return any $\hat{h} \in \mathbb{C}$ s.t. $\forall j < i, \left(\mathrm{er}_{Q_j}(h) - \mathrm{er}_j^{*}\right)|Q_j| \leqslant U\left(2^j, \delta_{(2^j)}\right) 2^j$

同 CAL 一样, 算法 3 也有等价形式 (算法 4).

定理 7 $\forall \delta \in (0,1)$, RobustCAL$_{\delta}$ 达到了标签复杂度 Λ, 使得 $\forall \mathcal{P}_{XY}, a, \alpha$ 满足情况 *1*, $\forall \varepsilon \in (0,1)$,

$$\Lambda(\nu + \varepsilon, \delta, \mathcal{P}_{XY}) \lesssim a^2 \theta\left(a\varepsilon^{\alpha}\right)\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d\,\mathrm{Log}\left(\theta\left(a\varepsilon^{\alpha}\right)\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(a/\varepsilon)}{\delta}\right)\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

Algorithm 4 : **RobustCAL**$_\delta(n)$:

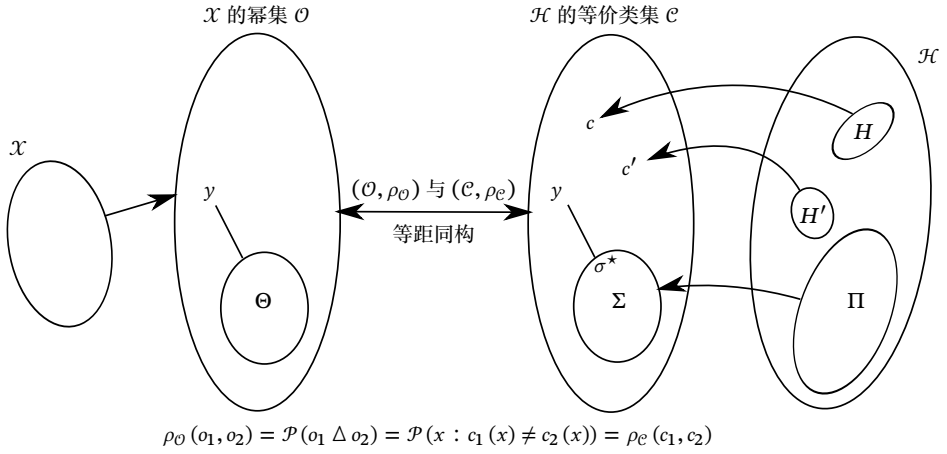
$m \longleftarrow 0$, $Q \longleftarrow \emptyset$, $V \longleftarrow \mathbb{C}$
while $|Q| < n \,\&\, m < 2^n$, **do**
  $m \longleftarrow m+1$
  if $X_m \in \mathcal{D}(V)$, **then**
    request label Y_m ; $Q \longleftarrow Q \cup \{(X_m, Y_m)\}$
  if $\log_2(m) \in \mathbb{N}$, **then**
    $V \longleftarrow \left\{h \in V : \left(\mathrm{er}_Q(h) - \min_{g \in V} \mathrm{er}_Q(g)\right)|Q| \leqslant U(m, \delta_m)\, m\right\}$
return any $\hat{h} \in V$

进一步,

$$\Lambda(\nu+\varepsilon, \delta, \mathcal{P}_{XY}) \lesssim \theta(\nu+\varepsilon)\left(\frac{\nu^2}{\varepsilon^2} + \mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right)\left(d\,\mathrm{Log}\left(\theta(\nu+\varepsilon)\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right).$$

4 注

本质上, 基本定义是在指出分类器集合 $\mathcal{H}$ 的拓扑性质. 理想情况, 假设 $\mathcal{H}$ 足够大, 大到任何一种 $\mathcal{X}$ 的二分类标签情况, 都对应一簇分类器 $H \subset \mathcal{H}$, 这样, 一方面, 将 H 视作等价类集 $\mathcal{C}_{\mathcal{X},\mathcal{H}}$ 的点, 一方面记 $\mathcal{X}$ 的幂集 $\mathcal{O}_\mathcal{X}$, 并假设 $\mathcal{X}$ 上的测度函数为 $\mathcal{P}$, 且不存在非空集零测, 这样 $\mathcal{O}_\mathcal{X}$ 与 $\mathcal{C}_{\mathcal{X},\mathcal{H}}$ 可以构造等距同构. 略去脚标, 如下图.



若 $h \in H$, 则 $H = \{\forall h' \in \mathcal{H} : h'(x) = h(x), \forall x \in \mathcal{X}\}, \mathcal{H} \supset H \mapsto c \in \mathcal{C}$. 因为是等距同构, 与距离相关的定义都可以对等写出, 如 $\varepsilon-$ 闭球.

现在考虑分类器族并不足够大, 不足以打散 $\mathcal{X}$, 可以用 Π 来表示实际的分类器族, $\Pi \subset \mathcal{H}$, 显然有如图对应的 Σ, Θ, 图中 y 是标签, 表示 Π 没有一个元素可以做到

完全正确的分类的情况, 于是问题等价于求 y 对 Σ 的投影, $\sigma^\star = \arg\inf_{\sigma\in\Sigma} \rho_{\mathcal{C}}(\sigma, y) = \arg\inf_{h\in\Pi} \mathcal{P}(x : h(x) \neq y)$, 这与噪声率的定义有出入, 噪声率用 $\mathcal{P}_{XY}$ 定义的.

再次审视度量空间, 必须要求不存在非空集零测, 否则 ρ 是伪度量, 好在按照度量的理解可以弱化到伪度量上. 注意 $\rho_{\mathcal{O}}(o_1, o_2) = \mathcal{P}(o_1 \Delta o_2) = \mathcal{P}(x : c_1(x) \neq c_2(x)) = \rho_{\mathcal{C}}(c_1, c_2)$ 里的对称差 $o_1 \Delta o_2 = o_1 \cup o_2 \setminus o_1 \cap o_2$, 将其扩展到子集 O 上, 有

$$\rho_{\mathcal{O}}(O) = \mathcal{P}\left(\bigcup_{o\in O} o \setminus \bigcap_{o\in O} o\right) = \mathcal{P}(x : \exists c_1, c_1 \in C \text{ s.t. } c_1(x) \neq c_2(x)) = \rho_{\mathcal{C}}(C).$$

这就得到了分歧域和分歧率的定义.