Discussion 7

No free lunch theory

田鹏超

tianpch@shanghaitech.edu.cn

PAC/SLT models for Supervised Learning

- X – feature or instance space; distribution D over X

  e.g., $X = R^d$ or $X = \{0,1\}^d$

- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots, (x_m,c^*(x_m))$, x_i i.i.d. from D
  - labeled examples - assumed to be drawn i.i.d. from some distr. D over X and labeled by some target concept c^*
  - labels $\in \{-1,1\}$ - binary classification

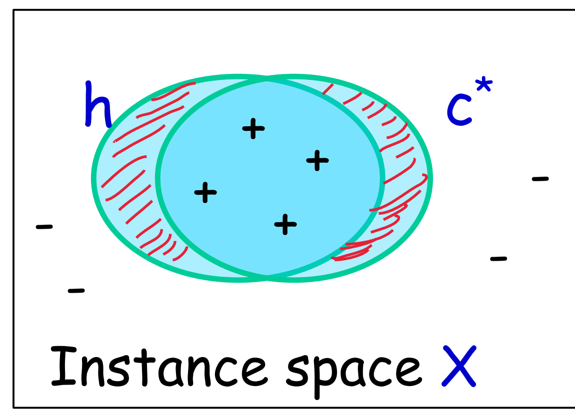
$c^*: X \to \{0,1\}$

- Algo does optimization over S, find hypothesis h.

- Goal: h has small error over D.

$= E_D(h(x) \neq c^*(x))$

$$err_D(h) = \Pr_{x \sim D}(h(x) \neq c^*(x))$$



$$L(h(x), c^*(x)) = \begin{cases} 1, & h(x) \neq c^*(x) \\ 0, & h(x) = c^*(x) \end{cases}$$

Need a bias: no free lunch.

$c^* \in H$

NFL (no free lunch) Theorem:

(无万金油算法定理)

- $P(h \mid X, \mathscr{A})$: the probability of finding hypothesis h when applying learning algorithm $\mathscr{A}$ on training set $X \subset \mathscr{X}$
- f is the target function, $X \subset \mathscr{X}$ is noise-free
- Consider the 0-1 error
- Purpose: compare the performance of different $\mathscr{A}$

The out-of-sample error of $\mathscr{A}$ is given by:

$$E(\mathscr{A} \mid X, f) = \sum_{h} \sum_{\boldsymbol{x} \in \mathscr{X} - X} P(\boldsymbol{x}) \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) P(h \mid X, \mathscr{A})$$

1. D. H. Wolpert, "The Lack of A Priori Distinctions Between Learning Algorithms," in *Neural Computation*, vol. 8, no. 7, pp. 1341-1390, Oct. 1996.
2. David H. Wolpert and William G. Macready, “No Free Lunch Theorems for Optimization”, *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, VOL. 1, NO. 1, APRIL 1997
3. https://www.kdnuggets.com/2019/09/no-free-lunch-data-science.html

无万金油算法定理

$$E(\mathscr{A} \mid X, f) = \sum_h \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) P(h \mid X, \mathscr{A})$$

If you apply $\mathscr{A}$ to any problem (any f), the overall out-of-sample error of $\mathscr{A}$ is given by:

$$\begin{aligned}
\sum_f E(\mathscr{A} \mid X, f) &= \sum_f \sum_h \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) P(h \mid X, \mathscr{A}) \\
&= \sum_h \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) P(h \mid X, \mathscr{A}) \sum_f \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathscr{A}) \sum_f \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathscr{A}) \frac{1}{2} 2^{|\mathcal{X}|} \\
&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X} - X} P(\boldsymbol{x}) \cdot 1
\end{aligned}$$