choosing hypothesis class $\hat{=}$ making use of prior knowledge

No-Free-Lunch Theorem

Let A be any learning alg. (for binary classification), the loss function the 0-1 loss over domain $\mathcal{X}$. Let S be some training set of size $m < \frac{|\mathcal{X}|}{2}$.
Then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, s.t.

1. $\exists f: \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$
2. With prob. at least $\frac{1}{7}$, $L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}$

<u>Proof</u>: Let $C \subseteq \mathcal{X}$, $|C| = 2m$. Number of possible functions from C to $\{0,1\}$: $T = 2^{2m}$. Let these be denoted by $f_1, \dots, f_T$.
For each f_i, define

$$\mathcal{D}_i(\{(x,y)\}) = \begin{cases} \frac{1}{|C|}, & f_i(x) = y \\ 0, & \text{else} \end{cases}$$

$\Rightarrow L_{\mathcal{D}_i}(f_i) = 0$

For all alg. A receiving m examples and returning $A(S)$, we have

$$\max_{i \in [T]} \mathbb{E}[L_{\mathcal{D}_i}(A(S))] \geq \frac{1}{4}$$

\# possible sequences $k = (2m)^m$, $S_1, \dots, S_k$. $S_j^i \hat{=} S_j$ labeled by f_i

For fixed $\mathcal{D}_i$, we can only receive $S_1^i, \dots, S_k^i$.

$$\Rightarrow \max_{i \in [T]} \mathbb{E}[L_{\mathcal{D}_i}(A(S))] = \max_{i \in [T]} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}[A(S_j^i)]$$

$$\geq \min_{j \in [k]} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}[A(S_j^i)]$$

Fix $j \in [k]$, $S_j = (x_1, \dots, x_m)$, $v_1, \dots, v_p$ not in S. $\Rightarrow p \geq m$.

$$L_{\mathcal{D}_i}[A(S_j^i)] = \frac{1}{2m} \sum_{x \in C} \mathbb{1}_{[A(S_j^i)(x) \neq f_i(x)]}$$

$$\geq \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]}$$

000
001
010
011
100
101
110
111

$$\Rightarrow \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}[A(S_j^i)] \geq \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[\dots]}$$

$$\geq \frac{1}{2} \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[\dots]}$$

Fix $r \in [p]$. Partition $f_1, \dots, f_T$ into $T/2$ disjoint pairs $(f_i, f_{i'})$ s.t. f_i and $f_{i'}$ only "disagree" on v_r $\Rightarrow S_j^i = S_j^{i'}$

$$\Rightarrow \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} + \mathbb{1}_{[A(S_j^{i'})(v_r) \neq f_{i'}(v_r)]} = 1$$

$$\Rightarrow \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[\dots]} = \frac{1}{2}$$

Bias-Complexity Tradeoff

$$L_{\mathcal{D}}(h_S) = \varepsilon_{app} + \varepsilon_{est}$$

$$\varepsilon_{app} := \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$$

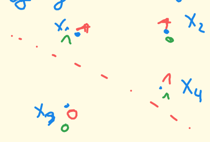
$$\varepsilon_{est} := L_{\mathcal{D}}(h_S) - \varepsilon_{app}$$

VC-Dimension

- finite classes are PAC-learnable
- some infinite classes are as well

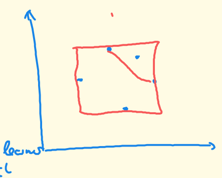
$\Rightarrow$ size of $\mathcal{H}$ is not the right criterion

Restriction of $\mathcal{H}$ to C:
$\mathcal{H}$ class of functions from $\mathcal{X}$ to $\{0,1\}$, $C \subseteq \mathcal{X}$, $|C| = m$.
The Restriction of $\mathcal{H}$ to C $\mathcal{H}_C$ is the set of functions from C to $\{0,1\}$ that can be derived from $\mathcal{H}$.
If $|\mathcal{H}_C| = 2^{|C|}$, then $\mathcal{H}$ <u>shatters</u> C.



The VC-Dimension $VCdim(\mathcal{H})$ is the largest size of a set $C \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.

Fundamental Theorem of Statistical Learning



The following are equivalent:

1. $\mathcal{H}$ has the uniform convergence prop.
2. any ERM rule is an (agnostic) PAC-learner for $\mathcal{H}$
3. $\mathcal{H}$ is (agnostic) PAC-learnable
4. $\mathcal{H}$ has finite VC-Dimension

If $VCdim(\mathcal{H}) = d < \infty$, then $\exists C_1, C_2$, s.t.

$\mathcal{H}$ is agnostic PAC-learnable with

$$C_1 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2}$$