Hypothesis Class and ERM

Apply ERM over a **restricted set** of hypotheses $\mathcal{H}$ = *hypothesis class* → $\mathcal{H} \in$ { "linear models", "SVM", "NNs", ... }

- each $h \in \mathcal{H}$ is a function $h : \mathcal{X} \to \mathcal{Y}$

$\mathrm{ERM}_{\mathcal{H}}$ learner:

$$\mathrm{ERM}_{\mathcal{H}} \in \arg\min_{h \in \mathcal{H}} L_S(h)$$

→ we could find multiple best hypotesis

Which hypothesis classes $\mathcal{H}$ do not lead to overfitting?

↳ which one are good

Finite Hypothesis Classes

Assume $\mathcal{H}$ is a finite class: $|\mathcal{H}| < \infty$

Let h_S be the output of $\mathrm{ERM}_{\mathcal{H}}(S)$, i.e. $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$

Assumptions

- **Realizability:** there exists $h^* \in \mathcal{H}$ such that $L_{\mathcal{D},f}(h^*) = 0$
- **i.i.d.:** examples in the training set are independently and identically distributed (i.i.d) according to $\mathcal{D}$, that is $S \sim \mathcal{D}^m$

Observation: realizability assumption implies that $L_S(h^*) = 0$

Can we *learn* (i.e., find using ERM) h^*? → no deterministic guarantee

(Simplified) PAC learning

Probably Approximately Correct (PAC) learning

Since the training data comes from $\mathcal{D}$:

- we can only be **approximately** correct → close to be correct
- we can only be **probably** correct → high chance

Parameters:

→ how good the model is

- *accuracy parameter* ε: we are satisfied with a *good* h_S: $L_{\mathcal{D},f}(h_S) \leq \varepsilon$ → I don't look for 0 cause I'm not sure it even exists
- *confidence parameter* δ: want h_S to be a *good* hypothesis with probability $\geq 1 - \delta$ → how sure I am that the model is good

We want both ε and δ small (~ 0)

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$, $\varepsilon \in (0,1)$, and $m \in \mathbb{N}$ such that

if $\Rightarrow$ $$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}.$$ size of the training set

[Then for any f and any $\mathcal{D}$] for which the realizability assumption holds, with probability $\geq 1-\delta$ we have that for every ERM hypothesis h_S it holds that

$$L_{\mathcal{D},f}(h_S) \leq \varepsilon.$$

I can apply this to every situation

Note: $\log$ = natural logarithm

With finite hypotesis classes ($\mathcal{H}$) I can almost always find a good hypotesis if I have enough data

$L_{\mathcal{D},f}(h_S) \leq \varepsilon$

$1-\delta$

$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$

Quantifies how large the dataset should be

The smaller the ε and δ, the more data I'll need

Proof (see book as well, Corollary 2.3)

let $S|_x = \{x_1, x_2, \ldots, x_m\}$ be the instances in the training set S.
We want to bound to:

$$D^m(\{S|_x : L_{D,f}(h_S) > \varepsilon\})$$

We call $H_B = \{h \in H : L_{D,f}(h_S) > \varepsilon\}$ (bad hypotesis) and

$$M = \{S|_x : \exists h \in H_B,\ L_S(h) = 0\} \quad \text{(misleading samples)}$$

Since the realisability assumption holds: $L_S(h_S) = 0$

$$L_{D,f}(h_S) > \varepsilon \text{ only if some } h \in H_B \text{ has } L_S(h) = 0$$

That is, our training data must be in the set M:

$$\{S|_x : L_{D,f}(h_S) > \varepsilon\} \subseteq M$$

Note that: $M = \bigcup_{h \in H_B} \{S|_x : L_S(h) = 0\}$

Therefore $D^m(\{S|x : L_{D,f}(h_S) > \varepsilon\}) \leq D^m(M) = D^m\left(\bigcup_{h \in H_B} \{S|x \mid L_S(h) = 0\}\right)$

$$D^m\left(\bigcup_{h\in H_B} \{S|x : L_S(h) = 0\}\right) \overset{\text{union bound}}{\leq} \sum_{h \in H_B} D^m(\{S|x : L_S(h) = 0\})$$

Now let's fix $h \in H_B$: $L_S(h) = 0 \iff \forall i = 1,\dots,m : h(x_i) = f(x_i)$

Therefore : $D^m(\{S|x : L_S(h) = 0\}) = D^m(\{S|x : \forall i = 1,\dots,m ; h(x_i) = f(x_i)\})$

$$= \prod_{i=1}^{m} D(\{x_i : h(x_i) = f(x_i)\})$$

identically distributed from D

Consider some i, $1 \leq i \leq m$: $D(\{x_i : h(x_i) = f(x_i)\})$

$$= 1 - \underbrace{D(\{x_i : h(x_i) \neq f(x_i)\})}_{L_{D,f}(h) = P_{x \sim D}[h(x) \neq f(x)]}$$

$$= 1 - L_{D,f}(h) \leq 1 - \varepsilon \leq e^{-\varepsilon}$$

since $h \in H_B$

Taylor expansion

Combining this result with the product of the probabilities:

$$D^m(\{S|_x : L_S(h) = 0\}) \le \prod_{i=1}^{m} e^{-\varepsilon} = e^{-m\varepsilon}$$

Combining the above with the sum of the probabilities:

$$D^m(\{S|_x : L_{D,f}(h_S) > \varepsilon\}) \le \sum_{h \in H_B} e^{-m\varepsilon} = |H_B| e^{-m\varepsilon} \le |H| e^{-m\varepsilon}$$

Now, given the choice of m, we have

$$\le \cancel{|H|}\, \cancel{e}^{-\cancel{\varepsilon}\left(\frac{\cancel{\log} \frac{\cancel{|H|}}{\delta}}{\cancel{\varepsilon}}\right)} = \delta$$

PAC Learning

Definition (PAC learnability)

A hypothesis class $\mathcal{H}$ is *PAC learnable* if there exist a function $m_{\mathcal{H}}$: $(0,1)^2 \to \mathbb{N}$ and a learning algorithm such that for every $\delta, \varepsilon \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, if the realizability assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples generate by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability $\geq 1 - \delta$ (over the choice of examples): $L_{\mathcal{D},f}(h) \leq \varepsilon$.

$m_{\mathcal{H}}$: $(0,1)^2 \to \mathbb{N}$: *sample complexity* of learning $\mathcal{H}$.

- $m_{\mathcal{H}}$ is the minimal integer that satisfies the requirements.

Corollary

Every finite hypothesis class is PAC learnable with sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \right\rceil$.

Algorithm to find a good hypothesis
$\Downarrow$
ERM