PAC learning - lite

Setup

- $\mathcal{X}$: Domain set (usually $= \mathbb{R}^n$)
- $\mathcal{Y}$: Label set
- $S = ((x_1, y_1) \ldots (x_m, y_m))$ = training set
- Learner's output: $h: \mathcal{X} \to \mathcal{Y}$ (prediction rule, hypothesis, classifier)
- Data generation model

D is a distribution over $\mathcal{X}$, $f: \mathcal{X} \to \mathcal{Y}$

$x \sim D$, $y = f(x)$

D over $\mathcal{X} \times \mathcal{Y}$

Later: $(x, y) \sim D$

- Measure of Success: Error of a classifier

$$L_{D,f}(h) := \mathbb{P}_{x \sim D}(h(x) \neq f(x)) = D(\{x : h(x) \neq f(x)\})$$

Terminology: generalization error, risk, true error

Note: Learner does not know D. only interacts w/ the env. by observing S.

ERM

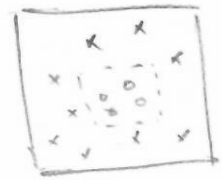
Training error / Empirical risk:

$$L_S(h) := \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m} = \frac{1}{m} \sum_{(x_i, y_i) \in S} 1_{\{h(x_i) \neq y_i\}}$$

<u>overfitting:</u>

label = 1   label = 0

D: o inside, x outside, uniformly distributed



Area of inner box = 1
outer box = 2

Given some S,
Let
$$h_S(x) = \begin{cases} y_i & \text{if } \exists\, i \in [m] \text{ s.t. } x_i = x \\ 0 & \text{else} \end{cases}$$

This assigns a correct label for all points in the train set and a label of zero otherwise. for any S; It has perfect empirical risk: $L_S(h_S) = 0$, but

$$L_{D,f}(h_S) = \frac{1}{2} \quad \forall S$$

This is an ERM hypothesis ($h_S \in \operatorname{argmin}_h L_S(h)$) that fails miserably.

Solution: Inductive bias. Let $\mathcal{H}$ be some restricted set of hypotheses (functions $h: \mathcal{X} \to \mathcal{Y}$). E.g., in above example, could let $\mathcal{H}$ = indicator functions over axis aligned rectangles. ~~We will solve~~ An ERM learner chooses a hypothesis ~~that minimizes~~ in $\operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$, i.e.,

$$ERM_{\mathcal{H}}(S) \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h)$$

Finite hypothesis classes

Suppose $\mathcal{H}$ is some finite hypothesis class. Let h_S be an ERM hypothesis given S

Def (Realizability assumption). $\exists\, h^* \in \mathcal{H}$ s.t. $L_{D,f}(h^*) = 0$.

Note: Assumption $\Rightarrow$ that w.p. 1, for every S sampled iid according to D, have $L_S(h^*) = 0$. Why? b/c $D(\{x : h^*(x) \neq f(x)\}) = 0$, so probability of drawing a sample & getting a mismatch is, by def, zero.

Note: There's always a chance a sample S will not be a good representation of D. We want our ERM hypothesis to be "good" according to D with high probability. It's also not reasonable to need ~~the hypo~~ a correct hypothesis exactly. Instead, want a hypothesis s.t. $L_D(h_S) < \varepsilon$, for some small ε.

Formally, given $\varepsilon, \delta > 0$, want a sample such that

$$\mathbb{P}(\, L_D(h_S) \leq \varepsilon\,) \geq 1-\delta.$$

Let D^m denote the product measure of D. Want $D^m(\{S|_x : L_{D,f}(h_S) > \varepsilon\}) \leq \delta$

S_x = just x's $= \{x_1, \dots, x_m\}$

Let $\mathcal{H}_B$ be set of "bad" hypotheses

$$\mathcal{H}_B = \{ h \in \mathcal{H} : L_D(h) > \varepsilon \}$$

Let

$$M = \{ S|_x : \exists\, h \in \mathcal{H}_B,\ L_S(h) = 0 \}$$

be the set of misleading samples

Recall that the realizability assumption $\Rightarrow$ that $L_S(h_S) = 0$ w.p. 1

So, w.p. 1, it can only happen that $L_D(h_S) > \varepsilon$ if for

some $h \in \mathcal{H}_B$ we have $L_S(h) = 0$. i.e., $S \in M$. Hence,

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

(More accurately, I think its $\mathbb{P}(\{S:\dots\} \setminus M) = 0$, but whatever)

write

$$M = \bigcup_{h \in \mathcal{H}_B} \{ S|_x : L_S(h) = 0 \}$$

Have

$$D^m(\{S|_x : L_{D,f}(h_S) > \varepsilon\}) \leq D^m\Big(\bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}\Big)$$

$$\leq \sum_{h \in \mathcal{H}_B} D^m(\{S|_x : L_S(h) = 0\})$$

Union bound

By the iid assumption,

$$D^m(\{S|_x : L_S(h) = 0\}) = D^m(\{S|_x : \forall i,\ h(x_i) = f(x_i)\})$$

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

For each i we have $D(\{x_i : h(x_i) = f(x_i)\}) = 1 - \underbrace{L_{D,f}(h)}_{> \varepsilon} < 1 - \varepsilon$

$$\overset{(*)}{\leq} (1-\varepsilon)^m \leq e^{-m\varepsilon}$$

$1 + x \leq e^x \ \forall x$ by convexity of e^x and first order condition.

Putting it all together

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

<u>Lemma</u> Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ & $\varepsilon > 0$ and let $m \in \mathbb{Z}_+$ s.t. $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$. Then for any f and D for which the realizability assumption holds, w.p. at least $1-\delta$, for S iid w/ m samples, we have for any ERM hypothesis h_S

$$L_D(h_S) \leq \varepsilon$$