Hoeffding's Inequality:

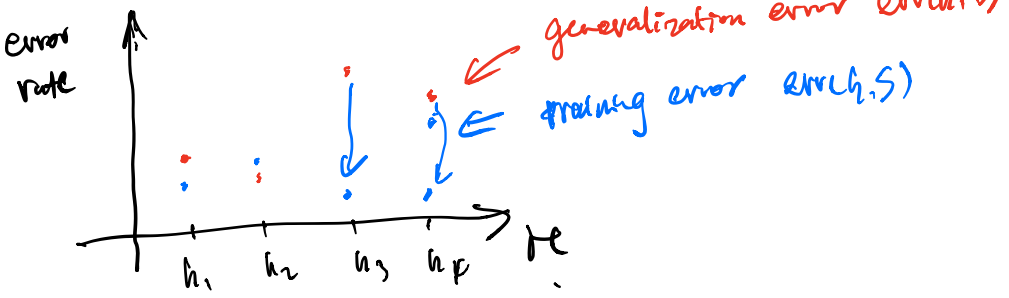
fix h. S = dataset of size m drawn i.i.d from D. (independent from h), then:

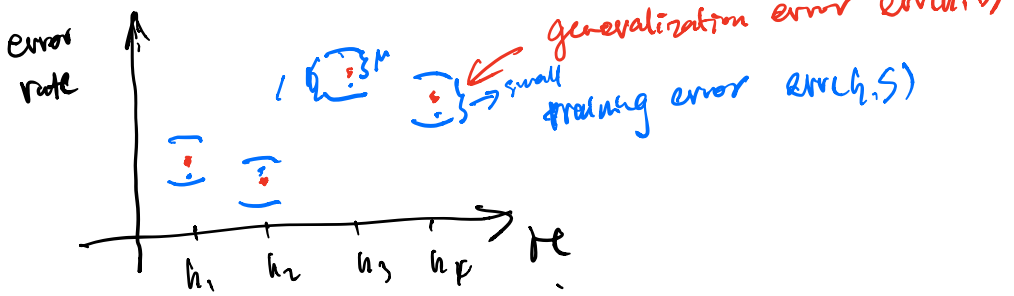
$$P\left(\left|err(h,S) - err(h,D)\right| \geq \sqrt{\frac{\ln 2/\delta}{2m}}\right) \leq \delta.$$

Agnostic PAC def. D $\min_{h \in \mathcal{H}} err(h,D)$ may or may not be 0.

ERM input S: return $\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} err(h,S)$.

Analysis of ERM:





$\forall h \in \mathcal{H}$: $\left|err(h,S) - err(h,D)\right| \leq \mu$ ... (*)

$\gamma = \min_{h \in \mathcal{H}} err(h,D)$ $\quad h^* = \operatorname{argmin}_{h \in \mathcal{H}} err(h,D)$

$$\text{err}(\hat{h}, S) \le \text{err}(h^*, S) \quad \text{(optimality of ERM)}$$

$$\le r + \mu \quad (*)$$

$$\text{err}(\hat{h}, D) \overset{?}{\le} \text{err}(\hat{h}, S) + \mu$$

$$\le (r + \mu) + \mu = r + 2\mu.$$

Next: find μ. ($(*)$ true with high probability $(1-\delta)$)

$$(*) = \bigcap_{h \in H} \{ |\text{err}(h,S) - \text{err}(h,D)| \le \mu \}$$

$$\overline{(*)} = \bigcup_{h \in H} \{ |\text{err}(h,S) - \text{err}(h,D)| > \mu \}$$

$$P(\overline{(*)}) \le \sum_{h \in H} P(\{ |\text{err}(h,S) - \text{err}(h,D)| > \mu \})$$

Hoeffding $\le \sum_{h \in H} 2e^{-2m\cdot\mu^2} = |H| \cdot 2e^{-2m\mu^2}$

choose μ s.t. $|H| 2e^{-2m\mu^2} = \delta$

$$\mu = \sqrt{\frac{\ln \frac{2|H|}{\delta}}{2m}}$$

In summary: w.p. $1-\delta$.

$(*)$ happens w/ $\mu = \dots$

$$\Rightarrow \text{err}(\hat{h}, D) \le \min_{h \in H} \text{err}(h, D) + 2 \cdot \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}.$$

have $m \le D\Sigma$

Theorem: ERM agnostic PAC learns $\mathcal{H}$, with a simple complexity function of $f(\varepsilon, \delta) = \frac{2}{\varepsilon^2}\left(\ln|\mathcal{H}| + \ln\frac{2}{\delta}\right)$

$$2\sqrt{\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2m}} = \varepsilon$$

consistency

$$\frac{1}{\varepsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

$$\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2m} = \frac{\varepsilon^2}{4} \longrightarrow$$

$$m = \frac{2}{\varepsilon^2}\left(\ln|\mathcal{H}| + \ln\frac{2}{\delta}\right)$$

Infinite hypothesis classes; Vapnik Chervonenkis Theory (VC)

$|\mathcal{H}| = \infty \Rightarrow \mathcal{H}$ is not PAC learnable ?

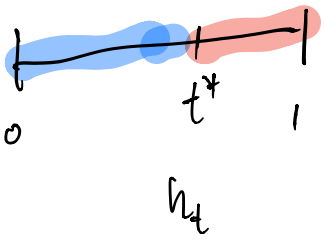
$|\mathcal{H}| < \infty \Rightarrow \mathcal{H}$ is PAC learnable by ERM. ✓

Ex Infinite class can be PAC learnable

$\mathcal{X} = [0,1]$

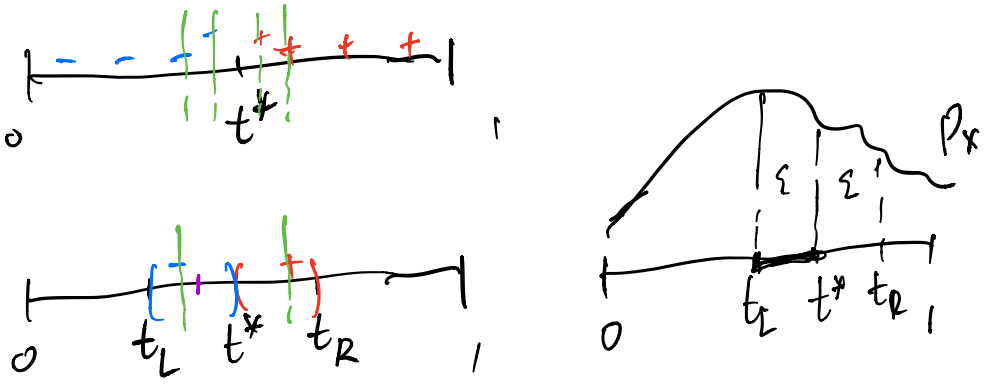
$\mathcal{Y} = \{+1, -1\}$

$\mathcal{H} = \{ h_t : 2\mathbb{I}(x > t) - 1 \; ; \; t \in [0,1] \}$



D realizable by h_{t^*} $\quad X \sim \text{uniform}([0,1])$.

How does the consistency alg. do ?



Observation: if $\exists$ training example in $[t_L, t^*]$, and (E_L)

$\exists$ training example in $[t^*, t_R]$, (E_R)

then $\hat{h}$: consistent classifier will have

$err(\hat{h}, D) \leq \varepsilon$.

want:

$P(E_L \cap E_R) \geq 1-\delta$

$\Leftarrow P(\bar{E}_L \cup \bar{E}_R) \leq \delta$

$P(\bar{E}_L) = P($ all training examples are outside $[t_L, t^*])$

$= \prod_{i=1}^{m} P(x_i$ outside $[t_L, t^*])$

$\leq (1-\varepsilon)$

$= (1-\varepsilon)^m$

$$P(\bar{E}_R) \le (1-\varepsilon)^m$$

$$P(\bar{E}_L \cup \bar{E}_R) \le 2\cdot(1-\varepsilon)^m \quad \text{want this to be } \le \delta.$$

$$(1-x \le e^{-x}$$

$$\Leftarrow 2\cdot e^{-m\varepsilon} \le \delta$$

$$\Leftarrow m \ge \frac{1}{\varepsilon}\ln\frac{2}{\delta}.$$

Thm: the consistency algorithm learns $\mathcal{H}$ (threshold class) with a sample complexity function of $f(\varepsilon,\delta) = \frac{1}{\varepsilon}\ln\frac{2}{\delta}$.

VC dimension

characterizing the complexity / expressiveness of hypothesis classes.

Def: hypothesis class $\mathcal{H}$, $\subseteq (X \to \{+1,-1\})$

sequence of unlabeled examples $S = (x_1 \cdots x_n)$, define the projection of $\mathcal{H}$ on S as

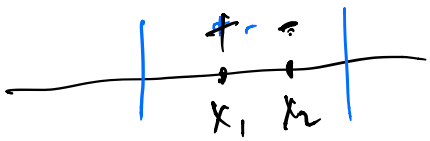
$$\Pi_{\mathcal{H}}(S) = \{(h(x_1)\cdots h(x_n)) : h \in \mathcal{H}\}$$

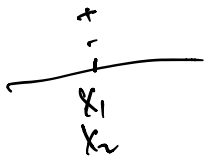
$h(x_i) \in \{+1,-1\}$

$$|\Pi_{\mathcal{H}}(S)| \overset{?}{\le} 2^n$$

Def: $\mathcal{H}$ shatter S if $|\Pi_{\mathcal{H}}(S)| = 2^n$.

Ex





$\mathcal{H}$: threshold class

$S = (x_1)$ ✓

$S = (x_1, x_2)$ ✗

Def: The VC dimension of $\mathcal{H}$ (abbrev. $VC(\mathcal{H})$) is $\max\{n \in \mathbb{N}: \mathcal{H}$ can shatter n points$\}$.

$VC(\mathcal{H}_{\text{threshold}}) = 1$

Lemma $d \in \mathbb{N}$.

$VC(\mathcal{H}) = d \iff$ { $\mathcal{H}$ can shatter d points, and
$\mathcal{H}$ cannot shatter $d+1$ points.

$\iff$ { $\exists$ d points shatterable by $\mathcal{H}$, and
$\forall$ $d+1$ points not shatterable by $\mathcal{H}$

Comment if $\forall n$, can find n points shatterable by $\mathcal{H}$, $VC(\mathcal{H})$ is defined as ∞.

∀ dataset of size 2 (x_1, x_2)

case 1 $x_1 \neq x_2$