Computational + Statistical Learning Theory

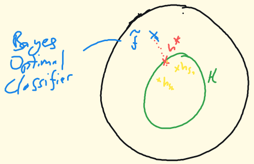
supervised vs. unsupervised
active vs. passive
random vs. st. else
batch vs. online

domain set $\mathcal{X}$
instance $x \in \mathcal{X}$
label set $\mathcal{Y}$ $(\mathbb{R}, \{-1,1\})$
label $y \in \mathcal{Y}$
$S = ((x_1,y_1),\dots,(x_m,y_m)) \subset \mathcal{X} \times \mathcal{Y}$
$|S| = m$

$\mathcal{D}: \mathcal{X} \to [0,1]$, $f: \mathcal{X} \to \mathcal{Y}$
hypothesis $h: \mathcal{X} \to \mathcal{Y}$
$h(x) = f(x)$
risk: $L_{\mathcal{D},f}(h) = P_{x\sim\mathcal{D}}[h(x) \neq f(x)]$
unknown

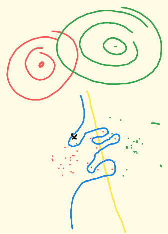
Empirical Risk Minimization (ERM):
training error: (empirical risk) $L_S(h) := \frac{1}{m}\left|\{i \in [m] : h(x_i) \neq \underbrace{f(x_i)}_{=y_i}\}\right|$
hypothesis class $\mathcal{H}$ (e.g.: $\mathcal{F} := \{f_w : \mathbb{R}^d \to \mathbb{R},\ f_w(x) = \langle w, x\rangle, w \in \mathbb{R}^d\}$)
$= \mathrm{sign}(\langle w, x\rangle)$
$h_S = \mathrm{ERM}_{\mathcal{H}}(S) = \underset{h\in\mathcal{H}}{\mathrm{argmin}}\, L_S(h) \to \{-1,1\}$

$L_{\mathcal{D},f}(h_S) - L_{\mathcal{D},f}(h^*)$ $\quad$ $\underset{h\in\mathcal{H}}{\mathrm{argmin}}\, L_{\mathcal{D},f}(h^*)$



Overfitting

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



Generalization Bound

finite $\mathcal{H}$
realizability: $\exists h^* \mathcal{H}$ s.t. $L_{\mathcal{D},f}(h^*) = 0$
iid: S 1) draw $x_1,\dots,x_m$ iid $\sim \mathcal{D}$
2) $y_i = f(x_i)$

Corollary: finite $\mathcal{H}$, $\delta \in (0,1)$, $\varepsilon > 0$, $m \geq \frac{\log\left(\frac{|\mathcal{H}|}{\delta}\right)}{\varepsilon}$
$\Rightarrow$ for any $\mathcal{D}, f$ (realizable) with prob $1-\delta$
for iid S of size m
$h_S \in \mathrm{ERM}_{\mathcal{H}}(S)$ it holds that

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

Proof (sketch): $\mathcal{D}^m(\{S|_x : L_{\mathcal{D},S}(h_S) > \varepsilon\})$
prob. of "misleading" sample of size m

bad hyp. $\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \varepsilon\}$
set of misleading sampling $M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$

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

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

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

S iid

$$\leq \sum_{h\in\mathcal{H}_B} \prod_{i=1}^m \mathcal{D}(\{x_i : \underbrace{h(x_i) = f(x_i)}_{\leq 1-\varepsilon \text{ since } h \in \mathcal{H}_B}\}) \leq (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

$\Leftrightarrow \quad \delta \leq |\mathcal{H}| e^{-\varepsilon m}$