COL774

Machine Learning

Nov 9, 2021

Last class:-

Learning Theory

$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$ $\quad y^{(i)} \in \{0,1\}$

$h \in \mathcal{H}$

$$\Pr\left(|\underbrace{\varepsilon(h)}_{\text{generalization error}} - \underbrace{\hat{\varepsilon}(h)}_{\text{training error}}| > \gamma \right) \le \delta \quad (1)$$

PAC:- Probably Approximately correct

↳ Hoeffding inequality

$$\delta \ge 2e^{-2\gamma^2 m}$$

$$\Rightarrow \Pr\left[\exists h \in \mathcal{H}, \ |\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma \right) \le \delta \quad (2)$$

Property of SLB

$$P\left(\bigcup_{l=1}^{K} A_l\right) \le \sum_{l=1}^{K} P(A_l) \quad K = |\mathcal{H}|$$

$$\Rightarrow \quad \delta \ge 2K e^{-2\gamma^2 m} \quad (3)$$

$(x^{(i)}, y^{(i)}) \sim \text{Dist}$

$(x, y) \sim \text{Dist}$

$\Rightarrow$ Uniform convergence Bound

Implications of the above Result:-

Four quantities of interest:-

δ, γ, K, m

(a) $\delta \geq 2k\, e^{-2\gamma^2 m}$ -

(b) $\dfrac{\delta}{2k} \geq e^{-2\gamma^2 m}$

$\log \delta/2k \geq -2\gamma^2 m$

$\log \dfrac{2k}{\delta} \leq 2\gamma^2 m$

$\Rightarrow m \geq \dfrac{1}{2\gamma^2} \log\left(\dfrac{2k}{\delta}\right)$

(c) $\gamma \geq \sqrt{\dfrac{1}{2m} \log \dfrac{2k}{\delta}}$

Finally:- → optimal hypothesis on training data

have access to ← $\hat{h} := \underset{h \in \mathcal{H}}{\arg\min}\ \hat{\varepsilon}(h)$ → training error

interested in ← $h^* := \underset{h \in \mathcal{H}}{\arg\min}\ \varepsilon(h)$ → generalization error

Bound the difference between:

$\varepsilon(\hat{h})$ $\quad$ $\varepsilon(h^*)$

Uniform convergence holds

$\varepsilon(\hat{h}) \overset{①}{\leq} \hat{\varepsilon}(\hat{h}) + \gamma \overset{②}{\leq} \hat{\varepsilon}(h^*) + \gamma$

↳ using 3(b) $\quad$ ↳ $\hat{h}$ is optimal on training data

uniform convergence (3(ii)) $\overset{③}{\leq} (\varepsilon(h^*) + \gamma) + \gamma$

bound

With probability at least $(1-\delta)$

$\nexists h \in \mathcal{H} \quad |\mathcal{E}_e(h) - \hat{\mathcal{E}}_D(h)| > \gamma$ (3a)

$\checkmark\ \forall h \in \mathcal{H} \quad |\mathcal{E}_e(h) - \hat{\mathcal{E}}_e(h)| \le \gamma$ (3b)

uniform convergence bound

A

$Pr(\underline{\exists h \in \mathcal{H} : |\mathcal{E}_e(h) - \hat{\mathcal{E}}_e(h)| > \gamma}) \le \delta$ — (1)

$\Rightarrow Pr(\nexists h \in \mathcal{H} : |\mathcal{E}_e(h) - \hat{\mathcal{E}}_e(h)| > \gamma) \ge 1-\delta$ (2)

$P(A) \le \delta$ $\qquad A = \bigcup_{\ell=1}^{K} A_\ell$

$P(\neg A) \ge 1-\delta$

With probability at least $(1-\delta)$

$$\boxed{\mathcal{E}_0(\hat{h}) \le \mathcal{E}_0(h^*) + 2\gamma}$$

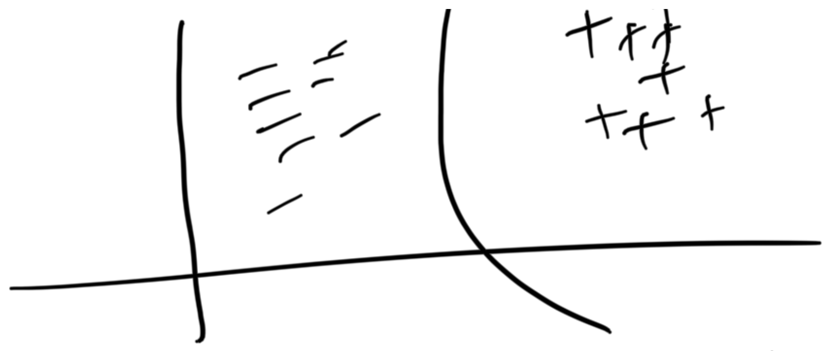
↳ VC-Dimension:-

Dealing with infinite sized hypothesis spaces:-

$\mathcal{H}_1$:- infinite sized

$h_\theta \in \mathcal{H}$:- space of all linear separators on 2D

finite sized



$\mathcal{H}_2$:- [infinite size]
↳ set of all quadratic separators in 2-D

$$\text{VC-Dim}(\mathcal{H}_1) < \text{VC-Dim}(\mathcal{H}_2)$$

Inherent representational capability of a hypothesis class

VC-Dim :- Vapnik Chervonenkis Dimension

Idea of shattering :-

$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$

$y^{(i)} \in \{0, 1\}$

$\mathcal{X}$:- instance space

$x^{(i)} \in \mathbb{R}^n$

$\mathcal{H}$:- target space

$\mathcal{H}$:- infinite sized

Consider $\{x^{(1)} \ldots x^{(d)}\} = S$

$x^{(i)} \in \mathcal{X}$

Defn:- labeling :- Given a set $S = \{x^{(1)} \ldots x^{(d)}\}$ a labeling $l = \langle l^{(1)} \ldots l^{(d)} \rangle$

$l^{(i)} \in \{0, 1\}$ is an assignment of label (target value) to point $x^{(i)}$

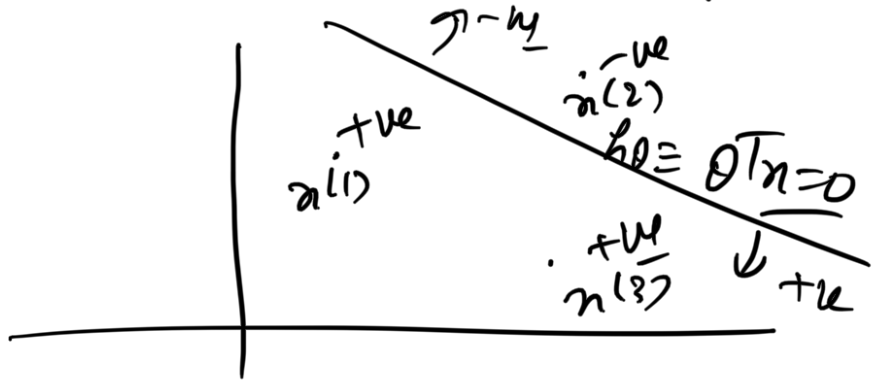
Defn:- $S = \{x^{(1)} \ldots x^{(d)}\}$ $x^{(i)} \in \mathcal{X}$,

S is shattered by $\mathcal{H}$, if

$\forall\, l \in \mathcal{L}_d, \quad \exists\, h \in \mathcal{H}$ such that

$\Downarrow$ set of all possible labelings

l is realized by $\underline{h}$. $\mathcal{H}$:- set of linear classifiers



$\underline{\mathbb{R}^2}$

$x^{(4)}$ $x^{(4)}$ $x^{(5)}$

$h_\theta(x)$:-

$h_\theta \in \underline{\mathcal{H}}$

$$\underline{h_\theta(x) = 1\{\theta^T x \geq 0\}}$$