10.1

Let $\varepsilon, \delta \in (0,1)$. pick k "chunks" of size $m_{\mathcal{H}}(\varepsilon/2)$

Apply A on every k chunks, to obtain $\hat{h}_1, \dots, \hat{h}_k$.

Apply ERM over $\hat{H} := \{\hat{h}_1, \dots, \hat{h}_k\}$

training data $\rightsquigarrow$ last chunk of size $\left\lceil \frac{2\log(4k/\delta)}{\varepsilon^2} \right\rceil$

with probability at least $1-\delta/2$, $L_D(\hat{h}) \leq \min_{i\in[k]} L_D(\hat{h}_i) + \varepsilon/2$

Apply the union bound $\rightsquigarrow$ probability at least $1-\delta$

$$L_D(\hat{h}) \leq \min_{i\in[k]} L_D(\hat{h}_i) + \varepsilon/2$$

$$\leq \min_{h\in\mathcal{H}} L_D(h) + \varepsilon$$

10.4

(a) Let X be finite set of size n.

Let B be the class of all func $X \to \{0,1\}$

$L(B,T) = B$. for any $T \geq 1$,

$$\text{vc dim}(B) = \text{VCdim}(L(B,T)) = \log 2^n = n$$

(b) B class of decision stump in $\mathbb{R}^d$

$B = \{h_{j,b,\theta} : j \in [d], b \in \{-1,1\}, \theta \in \mathbb{R}\}$

where $h_{j,b,\theta}(x) = b \cdot \text{sign}(\theta - x_j)$

for each $j \in [d]$ let $B_j = \{h_{b,\theta} : b \in \{-1,1\}, \theta \in \mathbb{R}\}$

$h_{b,\theta}(x) = b \cdot \text{sign}(\theta - x_j)$, $(\text{VCdim}(B_j) = 2)$

$B = \bigcup_{j=1}^{d} B_j \Rightarrow \text{VCdim}(B) \leq 16 + 2\log d$

$d = 2^k$ for some $k \in \mathbb{N}$. $A \in \mathbb{R}^{k \times d} \Rightarrow$ matrix whose columns range over set $\{0,1\}^k$. for each $i \in [k]$ let $x_i = A_{i,\to}$. ~~we claim~~ $C = \{x_i, \dots, x_k\}$ is shattered.

Let $I \subseteq [k]$. the instances $[k] \setminus I$ are labeled negatively. exist an index j such that $A_{i,j} = x_{i,j} = 1$ iff $i \in I$. Then, $h_{j,-1,1/2}(x_i) = 1$ iff $i \in I$.

(c) for each $i \in [Tk/2]$, let $x_i = [i/k]\, A_{i,\to}$.

set $C = \{x_i : i \in [Tk/2]\}$ is shattered by $L(B_d, T)$

Let $I \subseteq [Tk/2]$. $I = I_1 \cup \dots \cup I_{T/2}$. I_t is subset of $\{(t-1)k+1, \dots, tk\}$. let j_t be the corresponding column of A. $h(x) = \text{sign}((h_{j_1,-1,1/2} + h_{j_1,1,3/2} + \dots h_{j_{T/2},-1,T/2-1/2}(x))$

Then $h(x_i) = 1$ iff $i \in I \Rightarrow$ observe that $h \in L(B_d, T)$

11.1

let h be the output of the described learning algorithm. $L_D(h) = 1/2$.

calculating Estimate $L_V(h)$. parity of s is 1.

1. Fix some Fold $\{(x,y)\} \subseteq S$.

• parity of $S \setminus \{x\}$ is 1. follows thats $y = 0$. trained using $S \setminus \{x\}$, outputs the constant Predictor $h(x) = 1$. Hense the leave-one-out estimate using this fold is 1.

• parity of $S \setminus \{x\}$ is 0. it follows that $y = 1$. when being trained using $S \setminus \{x\}$. the algorithm outputs the constant predictor $h(x) = 0$.

Averaging over folds, the estimate of the error of h is 1. cossequently, the diffrence between the esitimate and true error is 1/2. parity of S is 0 is analyzed analogously.

(11.2) H_k in the agnostic-Pac model provides the following bound for ERM hypothesis h:

$$L_D(h) \leq \min_{h \in H_k} L_D(h) + \sqrt{\frac{2(k+1+\log(1/\delta))}{m}}.$$

Assume that j is the minimal index which contains a hypothesis $h^* \in \operatorname{argmin}_{h \in H} L_D(h)$. Fix some $r \in [k]$. By hoeffding's inequality, with probability at least $1-\delta/2k$

$$|L_D(\hat{h}_r) - L_V(\hat{h}_r)| \leq \sqrt{\frac{1}{2\alpha m} \log 4/\delta}.$$

Applying the uinon bound, we obtain that with probability at least $1-\delta/2$, the following inequality holds for $\forall r \in [k]$:

$$L_D(\hat{h}) \leq L_V(\hat{h}) + \sqrt{\frac{1}{2\alpha m} \log \frac{4k}{\delta}}.$$

$$\leq L_D(\hat{h}_r) + \sqrt{\frac{2}{\alpha m} \log \frac{4k}{\delta}}.$$

I with probability $1-\delta/2$ we have:

$$L_D(\hat{h}) \leq L_D(\hat{h}_j) + \sqrt{\frac{2}{\alpha m} \log \frac{4k}{\delta}}.$$

$$\rightarrow L_D(\hat{h}_j) \leq L_D(h^*) + \sqrt{\frac{2}{(1-\alpha) m} \log \frac{4|H_j|}{\delta}}$$

$$= L_D(h^*) + \sqrt{\frac{2}{(1-\alpha) m} \log \frac{4|H_j|}{\delta}}$$

Scanned with CamScanner

combining the two last inequalities:

$$L_D(\hat{h}) \leq L_D(h^*) + \sqrt{\frac{2}{\alpha m} \log 4k/\delta} + \sqrt{\frac{2}{(1-\alpha)m} \log \frac{4|H_j|}{\delta}}$$

we ~~ent~~ conclude that

$$L_D(\hat{h}) \leq L_D(h^*) + \sqrt{\frac{2}{\alpha m} \log \frac{4k}{\delta}} + \sqrt{\frac{2}{(1-\alpha)m} \left(j + \log \frac{4}{\delta}\right)}$$

comparing the two bounds.

if j is logarithmic in k, we achieve a logarithmic improvement

18.2

(a) information ~~For~~ gain for feature 1 is:

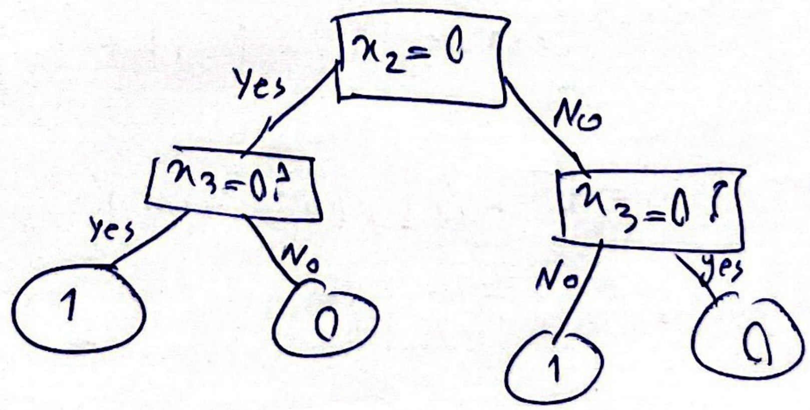
$$H(1/2) - \left(\frac{3}{4} H\left(\frac{2}{3}\right) + \frac{1}{4} H(0)\right) \approx 0{,}22$$

the information gain for feature 2 as well as feature 3

$$H\left(\frac{1}{2}\right) - \left(\frac{1}{2} H(1/2) + 1/2\, H(1/2)\right) = 0$$

we wont be able to classify all three example perfectly. since we have 4 examples in the training set, it follows that the training error is at least 1/4.

(b)



Scanned with CamScanner