

# Machine learning HW 4

Understanding Machine Learning: From Theory to Algorithms [Exercises for chapters 10,11 and 18]

Mehrnaz jalili:400422061- Arash Sajjadi:400422096

Email: Mehrnazjalili1991@gmail.com

Email: arash.sajjadi@yahoo.com

Date: January 30, 2022

## chapter 10

### Exercise 10.1:

$\epsilon, \delta \in (0, 1)$  Pick  $k$  “chunks” of size  $m_{\mathcal{H}}(\epsilon/2)$ . Apply A on each of these chunks, to obtain  $\hat{h}_1, \dots, \hat{h}_k$ <sup>1</sup> Now, apply an ERM over  $\hat{\mathcal{H}}$ . Note that  $\hat{\mathcal{H}} := \{\hat{h}_1, \dots, \hat{h}_k\}$  with the training data being the last chunk of size  $\lceil \frac{2}{\epsilon^2} \cdot \log(\frac{4k}{\delta}) \rceil$  Denote the output hypothesis by  $\hat{h}$ . We also should claim that with probability at least  $1 - \delta/2$ ,  $L_{\mathcal{D}}(\hat{h}) \leq \min_{i \in [k]} L_{\mathcal{D}}(h_i) + \frac{\epsilon}{2}$ . Now we have:  $L_{\mathcal{D}}(\hat{h}) \leq \min_{i \in [k]} L_{\mathcal{D}}(h) + \epsilon$

## chapter 11

### Exercise 11.1:

Consider a case in that the label is chosen at random according to  $\mathbb{P}[y = 1] = \mathbb{P}[y = 0] = \frac{1}{2}$  Consider a learning algorithm that outputs the constant predictor  $h(x) = 1$  if the parity of the labels on the training set is 1 and otherwise the algorithm outputs the constant predictor  $h(x) = 0$ . Prove that the difference between the leave-one-out estimate and the true error in such a case is always  $\frac{1}{2}$ .

first consider S set as a i.i.d sample

we know h as the out put of learning algorithm.

because h is a constant function we have  $L_D(h) = \frac{1}{2}$  we want to calculate the  $L_V(h)$ . assume the parity of S is 1

then fix some fold  $\{(x, y)\} \subseteq S$

we have two cases bellow:

- as the  $S \setminus \{X\}$  is 1 so  $Y = 0$  and since trained using  $S \setminus \{X\}$  the algorithm outputs the predictor  $h(x) = 1$   
therefor the leave-one-out estimate using this fold is 1.

- as the  $S \setminus \{X\}$  is 0 so  $Y = 1$  and since trained using  $S \setminus \{X\}$  the algorithm outputs the predictor  $h(x) = 0$

Therefore the leave-one-out estimate using this fold is 1.

after averaging the two folds, we calculate, we find out that the estimated error of h is 1. and the difference between estimation error and the true error is  $\frac{1}{2}$ .

for the other case (the parity be 0) analyze in the same way.

### Exercise 11.2:

consider  $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \dots \subseteq \mathcal{H}_k$  and also,  $\forall i \in k, |\mathcal{H}_i| = 2^i$ . Learning  $\mathcal{H}_k$  in the Agnostic-Pac model provides the following bound for an ERM hypothesis  $h$ :

$$L_{\mathcal{D}}(h) \leq \min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) + \left( \frac{2}{m} \cdot (k + 1 + \log(\frac{1}{\delta})) \right)^{\frac{1}{2}}$$

---

<sup>1</sup>Note that the probability that  $\min_{i \in [k]} L_{\mathcal{D}}(h) \leq \min L_{\mathcal{D}}(h) + \frac{\epsilon}{2}$  is at least  $1 - \delta_0^k \geq 1 - \frac{\delta}{2}$

Alternatively, we can use model selection as we describe next. let us to assume that  $j$  is the minimal index which contains a hypothesis  $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ . In this stage we try to fix some  $r \in [k]$ . Now according to Hoeffding's inequality, with probability at least  $\frac{1-\delta}{2k}$ , we have:

$$|L_{\mathcal{D}}(\hat{h}_r) - L_V(\hat{h}_r)| \leq \left( \frac{\log(4/\delta)}{2\alpha m} \right)^{\frac{1}{2}}$$

by applying the union bound we can claim that, In particular, with probability at least  $1 - \delta$  we have  $L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(\hat{h}_j) + \sqrt{\frac{2 \cdot \log(4k/\delta)}{\alpha m}}$  Using similar arguments, we obtain that with probability at least  $1 - \delta/2$ ,

$$L_{\mathcal{D}}(\hat{h}_j) \leq L_{\mathcal{D}}(h^*) + \sqrt{\frac{2 \log(4|\mathcal{H}_j|/\delta)}{m - m\alpha}}$$

Combining the two last inequalities with the union bound, we obtain that with probability at least  $1 - \delta$ :

$$L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(h^*) + \sqrt{\frac{2 \log(4k/\delta)}{\alpha m}} + \sqrt{\frac{2(j + \log(4/\delta))}{(1 - \alpha)m}}$$

Comparing the two bounds, we see that when the “optimal index”  $j$  is significantly smaller than  $k$ , the bound achieved using model selection is much better. Being even more concrete, if  $j$  is logarithmic in  $k$ , we achieve a logarithmic improvement.

## chapter 11

### Exercise 18.2:

in first iteration we compute the information gain:

$$H(Y) = -\frac{1}{2} \log\left(\frac{1}{2}\right) - \frac{1}{2} \log\left(\frac{1}{2}\right) = 1$$

$$\begin{aligned} IG(X_1) &= H(Y) - H(Y|X_1) \\ &= 1 - \left[ \left(\frac{3}{4}\right) \left( \left(-\frac{2}{3}\right) \log\left(\frac{2}{3}\right) \right) - \frac{1}{3} \log\frac{1}{3} + \frac{1}{4} (-0 \log 0 - 1 \log 1) \right] \\ &= 1 - \left(\frac{3}{4}\right) \left[ -\frac{2}{3} \log\left(\frac{2}{3}\right) - \frac{1}{3} \log\left(\frac{1}{3}\right) \right] > 0 \end{aligned}$$

$$\begin{aligned} IG(X_2) &= H(Y) - H(Y|X_2) \\ &= 1 - \left[ \left(\frac{1}{2}\right) \left( \left(-\frac{1}{2}\right) \log\left(\frac{1}{2}\right) \right) - \frac{1}{2} \log\frac{1}{2} + \frac{1}{2} \left( -\frac{1}{2} \log\frac{1}{2} - \frac{1}{2} \log\frac{1}{2} \right) \right] \\ &= 1 - \left[ \frac{1}{2}(-1) + \frac{1}{2}(-1) \right] = 0 \end{aligned}$$

$$\begin{aligned} IG(X_3) &= H(Y) - H(Y|X_3) \\ &= 1 - \left[ \left(\frac{1}{2}\right) \left( \left(-\frac{1}{2}\right) \log\left(\frac{1}{2}\right) \right) - \frac{1}{2} \log\frac{1}{2} + \frac{1}{2} \left( -\frac{1}{2} \log\frac{1}{2} - \frac{1}{2} \log\frac{1}{2} \right) \right] \\ &= 1 - \left[ \frac{1}{2}(-1) + \frac{1}{2}(-1) \right] = 0 \end{aligned}$$

so we choose  $X_1 = 0$  for begin the tree. 1. for choosing the left node:

$$ID3(\{((1, 1, 1), 1), ((1, 0, 0), 1), ((1, 1, 0), 0)\}, \{x_2, x_3\})$$

we have to compute the info. gain again

$$H(Y) = -\frac{2}{3} \log\left(\frac{2}{3}\right) - \frac{1}{3} \log\left(\frac{1}{3}\right)$$

$$\begin{aligned} IG(X_2) &= H(Y) - H(Y|X_2) \\ &= H(Y) - \left[ \frac{2}{3} \left( -\frac{1}{2} \log \frac{1}{2} \right) - \frac{1}{2} \log \frac{1}{2} \right] \\ &= H(Y) - \frac{2}{3} \end{aligned}$$

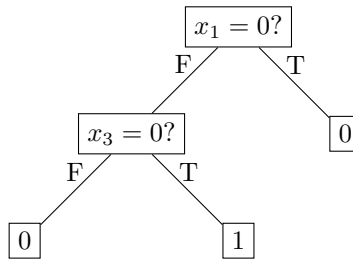
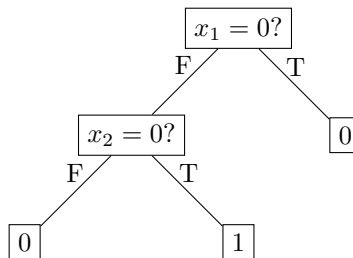
$$\begin{aligned} IG(X_3) &= H(Y) - H(Y|X_3) \\ &= H(Y) - \left[ \frac{2}{3} \left( -\frac{1}{2} \log \frac{1}{2} \right) - \frac{1}{2} \log \frac{1}{2} \right] \\ &= H(Y) - \frac{2}{3} \end{aligned}$$

its possible to choose either  $X_2$  or  $X_3$  to have 2 different trees.

2. right node:

$$ID3(\{((0, 0, 1), 0)\}, \{x_2, x_3\})$$

the only possible label is 0. training error for FIRST tree is  $\frac{1}{4}$  because the only mislabeled point is  $((1, 1, 1), 1)$



and the training error for the second tree is also  $\frac{1}{4}$  because the only mislabeled point is  $((1, 0, 0), 1)$   
 so the training error for any tree with the 2 depth with ID3 is at least  $\frac{1}{4}$

We want to show the decision tree with the 0 training error.

