

## LECTURE 4: PAC LEARNING: SIMPLE EXAMPLES

For a finite training sample  $\mathcal{D}_n$ , the predictor  $\hat{h}_n$  can be thought as the output of the learning algorithm given the training data and the hypothesis space, i.e.  $\hat{h}_n = \mathcal{A}(\mathcal{D}_n, \mathcal{H})$ . Its risk  $R(\hat{h}_n) = \mathbb{E}_{X,Y}[I(Y \neq \hat{h}_n(X))]$  is a random variable which depends on  $\mathcal{D}_n$ ,  $\mathcal{A}$  and  $\mathcal{H}$ .

Consistency of the learning algorithm has focused on the mean of this random variable, i.e.  $\mathbb{E}_{\mathcal{D}_n}[R(\hat{h}_n)]$ . In PAC learning we are interested in its tail distribution, i.e. finding a bound which holds with large probability:

$$P\left(\sup_{h \in \mathcal{H}} [R(h) - \hat{R}_n(h)] \geq \epsilon\right) \leq \delta.$$

The basic idea is to set the probability of being misled to  $\delta$  and thus solve the  $\epsilon$ .

**Example 1 (single classifier).** Consider the special case  $\mathcal{H} = \{h\}$ , i.e. we only have a single function. Furthermore, we assume that it can achieve 0 training error over  $\mathcal{D}_n$ , i.e.  $\hat{R}_n(h) = 0$ . Then what is the probability that its generalization error  $R(h) \geq \epsilon$ ? We have

$$\begin{aligned} P\left(\hat{R}_n(h) = 0, R(h) \geq \epsilon\right) &= (1 - R(h))^n \\ &\leq (1 - \epsilon)^n \\ &\leq \exp(-n\epsilon). \end{aligned}$$

Setting the RHS to  $\delta$  and solve for  $\epsilon$  we have  $\epsilon = \frac{1}{n} \log \frac{1}{\delta}$ . Thus with probability  $(1 - \delta)$ ,

$$P\left(\hat{R}_n(h) = 0, R(h) < \frac{1}{n} \log \frac{1}{\delta}\right).$$

Note that we can also utilize the Hoeffding's inequality to obtain  $P(|\hat{R}_n(h) - R(h)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$ , which leads to

$$P\left(|\hat{R}_n(h) - R(h)| \geq \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}\right) \leq \delta.$$

This is more general but not as tight as the previous one since it does not utilize the fact  $\hat{R}_n(h) = 0$ .  $\square$

Although the result in Example 1 is very simple, it has very limited practical meaning. The main reason is that it only applies to a single fixed function  $h$ . Essentially, it says that for each fixed function  $h$ , there is a set  $S$  of samples (whose measure  $P(S) \geq 1 - \delta$ ) for which  $|\hat{R}_n(h) - R(h)|$  is bounded. However, such  $S$  sets could be different for different functions. To handle this issue we need to obtain the *uniform deviations* since:

$$\hat{R}_n(\hat{h}_n) - R_n(\hat{h}_n) \leq \sup_{h \in \mathcal{H}} (\hat{R}_n(h) - R(h)).$$

The idea is to utilize the *union bound* as shown in the following example.

**Example 2 (finite number of classifiers).** Consider the case  $\mathcal{H} = \{h_1, \dots, h_m\}$ . Define

$$B_k := \left\{ (x_1, y_1) \dots, (x_n, y_n) : R(h_k) - \hat{R}_n(h_k) \geq \epsilon \right\}, k = 1, \dots, m.$$

Each  $B_k$  is the set of all **bad** samples for  $h_k$ , i.e. the samples for which the bound fails for  $h_k$ . In other words, it contains all **misleading** samples. If we want to measure the probability of the samples which are bad for any  $h_k$  ( $k = 1, \dots, m$ ), we could apply the *Bonferroni inequality* to simply obtain:

$$P(B_1 \cup \dots \cup B_m) \leq \sum_{k=1}^m P(B_k).$$

Thus we have

$$\begin{aligned}
 P\left(\exists h \in \mathcal{H} : R(h) - \hat{R}_n(h) \geq \epsilon\right) &= P\left(\bigcup_{k=1}^m \left\{R(h_k) - \hat{R}_n(h_k) \geq \epsilon\right\}\right) \\
 &\stackrel{\text{union bound}}{\leq} \sum_{k=1}^m P\left(R(h_k) - \hat{R}_n(h_k) \geq \epsilon\right) \\
 &\leq m \exp(-2n\epsilon^2).
 \end{aligned}$$

Hence for  $\mathcal{H} = \{h_1, \dots, h_m\}$ , with probability at least  $(1 - \delta)$ ,

$$\forall h \in \mathcal{H}, \quad R(h) - \hat{R}_n(h) \leq \sqrt{\frac{\log m + \log \frac{1}{\delta}}{2n}}.$$

Since this is a uniform upper bound, it can be applied to  $\hat{h}_n \in \mathcal{H}$ .  $\square$

Note that we could also bound the expected value of  $\mathbb{E}[\sup_{h \in \mathcal{H}} |R(h) - \hat{R}_n(h)|]$  by using the fact that for any nonnegative random variable  $Z$ ,  $\mathbb{E}[Z] = \int_0^\infty P(Z > t) dt$ .

From the above PAC learning examples we can see that

- It requires assumptions on data generation, i.e. samples are iid.
- The error bounds are valid with respect to repeated samples of training data.
- For a fixed function we roughly have  $R(h) - \hat{R}_n(h) \approx 1/\sqrt{n}$ .
- If  $|\mathcal{H}| = m$  then  $\sup_{h \in \mathcal{H}} (R(h) - \hat{R}_n(h)) \approx \sqrt{\log m/n}$ . The term  $\log m$  can be thought as the complexity of the hypothesis space  $\mathcal{H}$ .

There are several things which can be improved:

- Hoeffding's inequality does not utilize the variance information. So the results could be improved by utilizing such information.
- The union bound could be quite loose. For instance, it is as bad as if all the functions in  $\mathcal{H}$  were independent.
- The supremum over  $\mathcal{H}$  might be too conservative.

The bound in Example 2 becomes meaning less when  $m$  is infinite. The following example generalizes it to the case of countably many classifiers.

**Example 3 (countable number of classifiers).** Consider the case  $\mathcal{H} = \{h_1, h_2, \dots, h_m, \dots\}$ . Since we need to bound the probability of the set of misleading samples (which could mislead any  $h \in \mathcal{H}$ ) by  $\delta$ , we need budget the probability of being misled by  $h_m$  to  $w_m \delta$  such that  $\sum_{k=1}^\infty w_k \leq 1$ . So in order to find  $\epsilon > 0$  which satisfies

$$P\left(\exists h \in \mathcal{H} : R(h) - \hat{R}_n(h) \geq \epsilon\right) \leq \delta,$$

we only need to make sure that for any  $k$ ,

$$P\left(R(h_k) - \hat{R}_n(h_k) \geq \epsilon\right) \leq w_k \delta$$

since

$$\begin{aligned}
 P\left(\exists h \in \mathcal{H} : R(h) - \hat{R}_n(h) \geq \epsilon\right) &= P\left(\bigcup_{k=1}^{\infty} \left\{R(h_k) - \hat{R}_n(h_k) \geq \epsilon\right\}\right) \\
 &\stackrel{\text{union bound}}{\leq} \sum_{k=1}^{\infty} P\left(R(h_k) - \hat{R}_n(h_k) \geq \epsilon\right) \\
 &= \sum_k w_k \delta \\
 &\leq \delta.
 \end{aligned}$$

Again the first inequality comes from the Bonferroni inequality. By a similar argument, we solve  $\epsilon$  by setting  $\exp(-2n\epsilon^2) = w_k \delta$ , which leads to  $\epsilon = \sqrt{\frac{1}{2n} \log \frac{1}{w_k \delta}}$ . Thus we have with probability  $(1 - \delta)$ ,

$$\forall h \in \mathcal{H}, R(h) \leq \hat{R}_n(h) + \sqrt{\frac{\log \frac{1}{w_k} + \log \frac{1}{\delta}}{2n}}.$$

□

Note that  $w_k$ 's have to be specified before seeing the training data, otherwise the result will not hold. One way to interpret  $w_k$ 's is that they can be thought as the “prior” knowledge about the functions (such as in Bayesian inference).