Lecture-07: PAC Learning

1 PAC learning model

Definition 1.1 (PAC-learning). A concept class $C \subseteq \mathcal{Y}^{\mathcal{X}}$ is said to be PAC-learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function $\text{poly}(\cdot,\cdot,\cdot,\cdot)$ such that for (a) any $\epsilon,\delta > 0$, (b) any distribution $D \in \mathcal{M}(\mathcal{X})$, (c) any target concept $c \in C$, (c) any hypothesis h_z returned by the algorithm $\mathcal{A}$, and (d) any sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ of size $m \geqslant \text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c))$ generated under distribution D, the following holds

$$P\{R(h_z) \leqslant \epsilon\} \geqslant 1 - \delta.$$

If $\mathcal{A}$ further runs in $poly(\frac{1}{\epsilon}, \frac{1}{\delta}, n, size(c))$, then C is said to be efficiently PAC-learnable. When such an algorithm $\mathcal{A}$ exists, it is called a PAC-learning algorithm for C.

Remark 1. The cost of computational representation of an input vector $x \in \mathcal{X}$ is of order n, and of a concept c is of order $\text{size}(c)$.

Remark 2. A concept class C is thus PAC-learnable if the hypothesis returned by the algorithm after observing a sample of size polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ is approximately correct (error at most ϵ) with high probability (at least $1-\delta$), which justifies the PAC terminology. The $\delta > 0$ is used to define the confidence $1-\delta$ and $\epsilon > 0$ the accuracy $1-\epsilon$.

Remark 3. Note that if the running time of the algorithm is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, then the sample size m must also be polynomial if the full sample is received by the algorithm.

Remark 4. The following statements are true for the PAC framework.

1. It is a distribution-free model.
2. The training sample and the test examples are drawn from the same distribution D.
3. It deals with the question of learnability for a concept class C and not a particular concept.

2 Guarantees for finite hypothesis sets

Consider a binary classification problem where $\mathcal{Y} \triangleq \{0,1\}$ and a target concept $c \in C \subset \mathcal{Y}^{\mathcal{X}}$ such that $y = c(x)$ for any labeled example. Let $H \subset \mathcal{Y}^{\mathcal{X}}$ be a finite set of hypothesis functions for binary classification with loss function $\ell : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{1}_{\{h(x) \neq y\}}$, and consider an *i.i.d.* sample $z \in (\mathcal{X} \times \mathcal{Y})^m$. In this case for a hypothesis $h \in H$ and sample z, empirical risk is $\hat{R}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(x_i, y_i)$ and generalization risk $\mathbb{E}\ell(X, c(X)) = \mathbb{E}\hat{R}(h)$ for X distributed identically to an unlabeled sample.

2.1 Consistent case

We assume that $c \in H$ and hence for any sample z, there exists $h_z \in H$ such that empirical risk $\hat{R}(h_z) = 0$. Fix $\epsilon > 0$, and define events $E_h \triangleq \{R(h) \leqslant \epsilon\} \cup \{\hat{R}(h) \neq 0\}$ for each hypothesis $h \in H$. We provide a **uniform convergence bound** for all consistent hypotheses $h_z \in H$ such that $\hat{R}(h_z) = 0$, since we don't know which of these is selected by the algorithm $\mathcal{A}$.

Theorem 2.1 (Learning bound). *For $\epsilon,\delta > 0$ and $m \geqslant \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$, we have $P(\cap_{h\in H} E_h) \geqslant 1 - \delta$.*

Proof. For a given hypothesis $h \in H$ and any *i.i.d.* unlabeled training sample $X \in \mathcal{X}^m$, the probability of getting zero empirical risk is

$$P(E_h^c) = \mathbb{1}_{\{R(h)>\epsilon\}} \mathbb{E} \prod_{i=1}^{m} \mathbb{1}_{\{h(X_i)=Y_i\}} = \mathbb{1}_{\{R(h)>\epsilon\}} (1 - R(h))^m \leqslant (1-\epsilon)^m.$$

We next observe that the probability of getting a consistent hypothesis with the generalization risk exceeding ϵ is bounded by

$$P(\cup_{h\in H} E_h^c) = P\left(\cup_{h\in H}\left\{\hat{R}(h)=0, R(h)>\epsilon\right\}\right) \leqslant \sum_{h\in H} P\left\{\hat{R}(h)=0, R(h)>\epsilon\right\} \leqslant |H|\,(1-\epsilon)^m \leqslant |H|\,e^{-m\epsilon}.$$

Setting the right hand side to be equal to δ completes the proof. □

2.2 Inconsistent case

In many practical cases, the hypothesis set H may not consist of the target concept $c \in C$.

Theorem 2.2 (Learning bound). *Let H be a finite hypothesis set. Then, for any $\delta > 0$,*

$$P\Big(\cap_{h\in H}\left\{R(h) \leqslant \hat{R}(h) + \sqrt{\frac{1}{2m}(\ln|H| + \ln\frac{2}{\delta})}\right\}\Big) \geqslant 1-\delta.$$

Proof. Let $h \in H$ and fix $\epsilon > 0$. Recall that $\hat{R}(h) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{\{Y_i \neq h(X_i)\}}$ and $R(h) = \mathbb{E}\hat{R}(h)$. Applying Theorem A.2 to bounded random variables $\mathbb{1}_{\{Y_i\neq h(X_i)\}} \in \{0,1\}$ such that $\sigma^2 = m$, together with union bound, we get the generalization bound for single hypothesis $h \in H$, as

$$P\left\{\left|\hat{R}(h) - R(h)\right| \geqslant \epsilon\right\} = P\left\{\left|\sum_{i=1}^m (\mathbb{1}_{\{Y_i\neq h(X_i)\}} - R(h))\right| \geqslant m\epsilon\right\} \leqslant 2\exp(-2m\epsilon^2).$$

Using the union bound and applying the generalization bound, we get

$$P(\cup_{h\in H}\left\{\hat{R}(h) - R(h) > \epsilon\right\}) \leqslant \sum_{h\in H} P\left\{\hat{R}(h) - R(h) > \epsilon\right\} \leqslant 2|H|\exp(-2m\epsilon^2).$$

Setting the right-hand side to be equal to δ completes the proof. □

Remark 5. We observe the following from the upper bound on the generalized risk.

1. For finite hypothesis set H, $R(h) \leqslant \hat{R}(h) + O\left(\sqrt{\frac{\log_2|H|}{m}}\right)$.
2. The number of bits needed to represent H is $\log_2|H|$.
3. A larger sample size m guarantees better generalization.
4. The bound increases logarithmically with $|H|$.
5. The bound is worse for inconsistent case $\sqrt{\frac{\log_2|H|}{m}}$ compared to $\frac{\log_2|H|}{m}$ for the consistent case.
6. For a fixed $|H|$, to attain the same guarantee as in the consistent case, a quadratically larger labeled sample is needed.
7. The bound suggests seeking a trade-off between reducing the empirical error versus controlling the size of the hypothesis set: a larger hypothesis set is penalized by the second term but could help reduce the empirical error, that is the first term. But, for a similar empirical error, it suggests using a smaller hypothesis set.

A Hoeffding's lemma

Lemma A.1 (Hoeffding). *Let X be a zero-mean random variable with $X \in [a,b]$ for $b > a$. Then, for any $t > 0$, we have*

$$\mathbb{E}[e^{tX}] \leqslant e^{\frac{t^2(b-a)^2}{8}}.$$

Proof. From the convexity of the function $f(x) = e^{tx}$, we have for any $x = \lambda a + (1-\lambda)b \in [a,b]$ for $\lambda = \frac{b-x}{b-a} \in [0,1]$

$$e^x = f(x) \leqslant \lambda f(a) + (1-\lambda)f(b) = \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}.$$

Since $\mathbb{E}[X] = 0$, taking expectation on both sides, we get from the linearity of the expectations

$$\mathbb{E}[e^{tX}] \leqslant \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb} = e^{\phi(t)},$$

where the function $\phi(t)$ is given by

$$\phi(t) = ta + \ln\left(\frac{b}{b-a} + \frac{-a}{b-a}e^{t(b-a)}\right).$$

We can write the first two derivatives of this function $\phi(t)$ as

$$\phi'(t) = a - \frac{ae^{t(b-a)}}{\frac{b}{b-a} - \frac{a}{b-a}e^{t(b-a)}} = a - \frac{a}{\frac{b}{b-a}e^{-t(b-a)} - \frac{a}{b-a}},$$
$$\phi''(t) = \frac{-abe^{-t(b-a)}}{(\frac{b}{b-a}e^{-t(b-a)} - \frac{a}{b-a})^2} = (b-a)^2\left(\frac{\alpha}{(1-\alpha)e^{-t(b-a)} + \alpha}\right)\left(\frac{(1-\alpha)e^{-t(b-a)}}{(1-\alpha)e^{-t(b-a)} + \alpha)}\right) \leqslant \frac{(b-a)^2}{4},$$

where we have denoted $\alpha = \frac{-a}{b-a} \geqslant 0$ since $\mathbb{E}[X] = 0$. The result follows from the second order expansion of $\phi(t)$, such that we get for some $\theta \in [0, t]$

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(\theta) \leqslant t^2\frac{(b-a)^2}{8}.$$

□

Theorem A.2 (Hoeffding). *Let $(X_i \in [a_i, b_i] : i \in [m])$ be a vector of m independent random variables, and define $\sigma^2 \triangleq \sum_{i=1}^m (b_i - a_i)^2$. Then, for any $\epsilon > 0$ and $S_m \triangleq \sum_{i=1}^m X_i$, we have*

$$P\{S_m - \mathbb{E}S_m \geqslant \epsilon\} \leqslant \exp\left(-\frac{2\epsilon^2}{\sigma^2}\right), \qquad P\{S_m - \mathbb{E}S_m \leqslant -\epsilon\} \leqslant \exp\left(-\frac{2\epsilon^2}{\sigma^2}\right).$$

Proof. We define zero-mean random variables $Y_i \triangleq X_i - \mathbb{E}X_i$ for each $i \in [m]$. We observe that $(Y_i : i \in [m])$ is an independent sequence and $Y \triangleq \sum_{i=1}^m Y_i = S_m - \mathbb{E}S_m$. From the definition of indicator sets and for any increasing function $\phi : \mathbb{R} \to \mathbb{R}_+$, we can write

$$\phi(Y) \geqslant \phi(Y)\mathbb{1}_{\{Y \geqslant \epsilon\}} = \phi(Y)\mathbb{1}_{\{\phi(Y) \geqslant \phi(\epsilon)\}} \geqslant \phi(\epsilon)\mathbb{1}_{\{Y \geqslant \epsilon\}}.$$

Taking expectation on both sides for the mapping $\phi : x \mapsto e^{tx}$, we get the Chernoff bound from the independence of Y_i, as

$$P\{S_m - \mathbb{E}S_m \geqslant \epsilon\} \leqslant e^{-t\epsilon}\mathbb{E}[\exp(t(S_m - \mathbb{E}S_m))] = e^{-t\epsilon}\prod_{i=1}^m \mathbb{E}[\exp(t(X_i - \mathbb{E}X_i))].$$

We can upper-bound each term in the product by Lemma A.1 for zero-mean random variable $Y_i \in [a_i - \mathbb{E}X_i, b_i - \mathbb{E}X_i]$ and use the definition of σ^2, to get

$$P\{S_m - \mathbb{E}S_m \geqslant \epsilon\} \leqslant e^{-t\epsilon}\prod_{i=1}^m \exp(t^2(b_i - a_i)^2/8) = \exp\left(-t\epsilon + \frac{t^2\sigma^2}{8}\right).$$

First upper bound follows by observing that the upper bound is minimized for the choice of $t^* = \frac{4\epsilon}{\sigma^2}$. Second upper bound follows by repeating the same steps for bounded independent random variables $(-X_i : i \in [m])$ and $\epsilon > 0$. □