Lecture-09: PAC Learning

1 PAC learning model

Definition 1.1 (PAC-learning). Consider a concept class $C \subseteq \mathcal{Y}^{\mathcal{X}}$ where 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 size(c). The concept class C is said to be PAC-learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function poly$(\cdot,\cdot,\cdot,\cdot)$ such that $P\{R(h_z) \leqslant \epsilon\} \geqslant 1-\delta$ for any
(a) $\epsilon, \delta > 0$,
(b) distribution $D \in \mathcal{M}(\mathcal{X})$,
(c) target concept $c \in C$,
(d) hypothesis h_z returned by the algorithm $\mathcal{A}$,
(e) sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ of size m *i.i.d.* generated under distribution D, and
(f) of sample size $m \geqslant \text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c))$.
If $\mathcal{A}$ further runs in $poly(1/\epsilon, 1/\delta, n, size(c))$, 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. A concept class C is thus PAC-learnable if the hypothesis returned by the algorithm after observing a sample of size polynomial in $1/\epsilon$ and $1/\delta$ 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 2. Note that if the running time of the algorithm is polynomial in $1/\epsilon$ and $1/\delta$, then the sample size m must also be polynomial if the full sample is received by the algorithm.

Remark 3. We make the following observations for the PAC framework.
(a) It is a distribution-free model.
(b) The training sample and the test examples are drawn from the same distribution D.
(c) 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 labeled sample $z \in (\mathcal{X} \times \mathcal{Y})^m$, empirical risk is $\hat{R}(h) \triangleq \frac{1}{m}\sum_{i=1}^m \ell(x_i, y_i)$ 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

Assumption 2.1 (Consistent hypothesis set). We assume that $c \in H$ and hence for any sample z, there exists $h_z \in H$ such that empirical risk $\hat{R}(h_z) = 0$.

Definition 2.2. Consider the probability space $(\Omega, \mathcal{F}, D)$. Fix $\epsilon > 0$, and define events $E_h \triangleq \{R(h) \leqslant \epsilon\} \cup \{\hat{R}(h) \neq 0\}$ for each hypothesis $h \in H$.

Theorem 2.3 (Learning bound). *For any $\epsilon, \delta > 0$ and sample size $m \geqslant \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$, we have the inequality $P(\cap_{h \in H} E_h) \geqslant 1-\delta$ holds.*

Proof. 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 $\mathcal{A}$. We fix a hypothesis $h \in H$ and observe that

$$\mathbb{1}_{\{\hat{R}(h)=0\}} = \mathbb{1}_{\cap_{i=1}^m \{h(X_i)=Y_i\}} = \prod_{i=1}^m \mathbb{1}_{\{h(X_i)=Y_i\}} = \prod_{i=1}^m (1-\ell(X_i, Y_i)).$$

Since $R(h) = \mathbb{E}\ell(X_i, Y_i)$, for any *i.i.d.* labeled training sample $Z \in (\mathcal{X} \times \mathcal{Y})^m$, the probability of getting zero empirical risk is

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

Using this bound and union bound to sum the probability of union of events, we can bound the probability of getting a consistent hypothesis with the generalization risk exceeding ϵ as

$$P(\cup_{h\in H}E_h^c) \leqslant \sum_{h\in H} P(E_h^c) \leqslant |H|(1-\epsilon)^m \leqslant |H|e^{-m\epsilon}.$$

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

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

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

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

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

Remark 4. 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. We note that $a < 0 < b$ since $\mathbb{E}X = 0$. Any $x \in [a,b]$ can be written as $x = \lambda a + (1-\lambda)b$ for $\lambda \triangleq \frac{b-x}{b-a} \in [0,1]$. We fix $t > 0$ and observe that the map $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) \triangleq e^{tx}$ for each $x \in \mathbb{R}$, is convex. From the convexity of the function f, we have $f(x) \leqslant \lambda f(a) + (1-\lambda) f(b)$. It follows that for any random variable $X \in [a,b]$ and $t > 0$, we have

$$e^{tX} = f(X) \leqslant \frac{b-X}{b-a} e^{ta} + \frac{X-a}{b-a} e^{tb}.$$

Taking expectation on the both sides of the above equation, from the linearity of the expectations, and the fact that $\mathbb{E}[X] = 0$, we get

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

where the function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is defined as $\phi(t) \triangleq ta + \ln\left(\frac{b}{b-a} + \frac{-a}{b-a} e^{t(b-a)}\right)$ for each $t > 0$. We can write the first two derivatives of this function $\phi(t)$ as

$$\phi'(t) = a - \frac{a e^{t(b-a)}}{\frac{b}{b-a} - \frac{a}{b-a} e^{t(b-a)}} = a - \frac{a}{\frac{b}{b-a} e^{-t(b-a)} - \frac{a}{b-a}},$$

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

where we have denoted $\alpha = \frac{-a}{b-a} \geqslant 0$. The result follows from the second order expansion of $\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(\theta)$ for some $\theta \in [0,t]$. This implies that $\phi(t) \leqslant \frac{t^2(b-a)^2}{8}$ and the result follows. □

Theorem A.2 (Hoeffding). *Consider an independent random vector $X : \Omega \to \mathbb{R}^m$ such that $X_i \in [a_i, b_i]$ for each $i \in [m]$ 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

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

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 vector $-X$ and $\epsilon > 0$. □