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 any $\varepsilon > 0$ and $\delta > 0$, for all distributions D on input space $\mathcal{X}$ and for any target concept $c \in C$, the following holds for any sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ of size $m \geqslant \text{poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, n, \text{size}(c))$:

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

If $\mathcal{A}$ further runs in $poly(\frac{1}{\varepsilon}, \frac{1}{\delta}, n, size(c))$, 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 number of points polynomial in $\frac{1}{\varepsilon}$ 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 $\varepsilon > 0$ the accuracy $1-\varepsilon$. Note that if the running time of the algorithm is polynomial in $\frac{1}{\varepsilon}$ and $\frac{1}{\delta}$, then the sample size m must also be polynomial if the full sample is received by the algorithm.

Remark 3. 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 — consistent case

Theorem 2.1 (Learning bounds — finite H, consistent case). *Let $H \subset \mathcal{Y}^{\mathcal{X}}$ be a finite set of functions. Let $\mathcal{A}$ be an algorithm that for any target concept $c \in H$ and* i.i.d. *sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ returns a consistent hypothesis $h_z \in H$ such that $\hat{R}(h_z) = 0$. Then, for any $\varepsilon, \delta > 0$, the inequality $P\{R(h_z) \leqslant \varepsilon\} \geqslant 1 - \delta$ holds if*

$$m \geqslant \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right).$$

This sample complexity result admits the following equivalent statement as a generalization bound, for any $\varepsilon, \delta > 0$, with probability at least $1-\delta$

$$R(h_z) \leqslant \frac{1}{m}\left(\ln|H| + \ln\frac{1}{\delta}\right).$$

Proof. Fix $\varepsilon > 0$. We provide a **uniform convergence bound** for all consistent hypotheses $h_z \in H$, since we don't know which of these is selected by the algorithm $\mathcal{A}$. For a given hypothesis h and any unlabeled training sample $X \in \mathcal{X}^m$ drawn *i.i.d.* from the same distribution D, the probability of getting zero empirical risk is

$$P\left\{\hat{R}(h) = 0\right\} = P(\cap_{i=1}^m \{h(X_i) = Y_i\} = \prod_{i=1}^m P\{h(X_i) = Y_i\} = (1 - R(h))^m.$$

Consider any $h \in H$ such that $R(h) = \mathbb{E}\mathbb{1}_{\{h(X) \neq Y\}} > \varepsilon$, then the probability for any sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ drawn *i.i.d.* from the same distribution D with zero empirical risk is

$$P\left(\cup_{h \in H}\left\{\hat{R}(h) = 0\right\}\right) \leqslant \sum_{h \in H} P\left\{\hat{R}(h) = 0\right\}.$$

We can upper bound the probability of a hypothesis being consistent in terms of its generalization risk. Consider any $h \in H$ such that $R(h) = \mathbb{E}\mathbb{1}_{\{h(X) \neq Y\}} > \varepsilon$, then $P\left\{\hat{R}(h_z) = 0\right\} < (1-\varepsilon)^m$. The result follows from substituting this bound in the union bound. □

3 Guarantees for finite hypothesis sets — inconsistent case

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

Corollary 3.1 (Hoeffding). *Fix $\varepsilon > 0$ and let $z \in (\mathcal{X} \times \{0,1\})^m$ be an* i.i.d. *sample of size m. Then, for any hypothesis $h : \mathcal{X} \to \{0,1\}$*

$$P\left\{\hat{R}(h) - R(h) \geqslant \varepsilon\right\} \leqslant \exp(-2m\varepsilon^2), \qquad P\left\{\hat{R}(h) - R(h) \leqslant -\varepsilon\right\} \leqslant \exp(-2m\varepsilon^2).$$

By the union bound, we have $P\left\{\left|\hat{R}(h) - R(h)\right| \geqslant \varepsilon\right\} \leqslant 2\exp(-2m\varepsilon^2)$.

Proof. Recall that $\hat{R}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{Y_i \neq h(X_i)\}}$ and $R(h) = \mathbb{E}\hat{R}(h)$. We get the results by taking the random variables $\mathbb{1}_{\{Y_i \neq h(X_i)\}} \in \{0,1\}$, and applying Theorem A.2 with $\sigma^2 = m$. □

Corollary 3.2 (Generalization bound — single hypothesis). *For a hypothesis $h : \mathcal{X} \to \{0,1\}$ and any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$*

$$R(h) \leqslant \hat{R}(h) + \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

Theorem 3.3 (Learning bound — finite H, inconsistent case). *Let H be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$R(h) \leqslant \hat{R}(h) + \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}, \text{ for all } h \in H.$$

Proof. Let $h_1, \ldots, h_{|H|}$ be the elements of H. Using the union bound and applying the generalization bound, we get

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

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.

4 Generalities

4.1 Deterministic versus stochastic scenarios

Consider the **stochastic scenario** where the distribution D is defined over $\mathcal{X} \times \mathcal{Y}$. The training data is a labeled sample $T = ((X_i, Y_i) : i \in [m])$ drawn i.i.d. from the distribution D. The learning problem is to find a hypothesis $h \in H$ with small generalization error

$$R(h) = P\{h(X) \neq Y\} = \mathbb{E}[\mathbb{1}_{\{h(X) \neq Y\}}].$$

Definition 4.1 (Agnostic PAC-learning). Let H be a hypothesis set. An algorithm $\mathcal{A}$ is an agnostic PAC-learning algorithm if there exists a polynomial function $poly(\cdot,\cdot,\cdot,\cdot)$ such that for any $\varepsilon > 0$ and $\delta > 0$, for all distributions D over $\mathcal{X} \times \mathcal{Y}$, the following holds for any sample size $m \geqslant poly(\frac{1}{\varepsilon}, \frac{1}{\delta}, n, size(c))$

$$P\left\{R(h_S) - \min_{h \in H} R(h) \leqslant \varepsilon\right\} \geqslant 1 - \delta.$$

Further, if the algorithm $\mathcal{A}$ runs in $poly(\frac{1}{\varepsilon}, \frac{1}{\delta}, n, size(c))$, then it is said to be an efficient agnostic PAC-learning algorithm.

4.2 Bayes error and noise

In the deterministic case, by definition, there exists a target function $c : \mathcal{X} \to \mathcal{Y}$ with no generalization error $R(h) = 0$. In the stochastic case, there is a minimal non-zero error for any hypothesis.

Definition 4.2 (Bayes error). Given a distribution D over $\mathcal{X} \times \mathcal{Y}$, the Bayes error R^* is defined as the infimum of the errors achieved by measurable functions $h : \mathcal{X} \to \mathcal{Y}$

$$R^* \triangleq \inf_{h \text{ measurable}} R(h).$$

A hypothesis h with $R(h) = R^*$ is called a Bayes hypothesis or Bayes classifier.

In the deterministic case, we have $R^* = 0$, however $R^* \neq 0$ in the stochastic case. Recall that

$$R(h) = \mathbb{E}\mathbb{1}_{\{h(X) \neq Y\}} = \int_{x \in \mathcal{X}} dP(x) \sum_{y \in \mathcal{Y}} P(y|x) \mathbb{1}_{\{h(x) \neq y\}}.$$

The Bayes classifier h_B can be defined in terms of the conditional probabilities as

$$h_B(x) = \arg\max_{y \in \mathcal{Y}} P(y|x), \text{ for all } x \in \mathcal{X}.$$

The average error made by h_B on $x \in \mathcal{X}$ is thus $\min\left\{\sum_{z \in \mathcal{Y}: z \neq y} P(z|x)\right\}$, and this is the minimum possible error.

Definition 4.3 (Noise). For binary classification $\mathcal{Y} = \{0, 1\}$, given a distribution D over $\mathcal{X} \times \mathcal{Y}$, the noise at point $x \in \mathcal{X}$ is defined by

$$n(x) = \min\{P(1|x), P(0|x)\}.$$

The average noise or the noise associated to D is $\mathbb{E}[n(X)]$.

Remark 5. The average noise is the Bayes error, i.e. $\mathbb{E}[n(X)] = R^*$. The noise determines the difficulty of the learning task.

4.3 Estimation and approximation errors

For a hypothesis set H, we let h^* be the **best-in-class hypothesis** in the H with minimal error. Then, the difference between the generalization risk and Bayes error can be written as

$$R(h) - R^* = R(h) - R(h^*) + R(h^*) - R^*.$$

Definition 4.4. The second term $R(h^*) - R^*$ is called the **approximation error**, and is a measure of how well the Bayes error can be approximated by the class H.

Approximation error is a measure of the richness of the hypothesis set H, and not available in general.

Definition 4.5. The first term $R(h) - R(h^*)$ is called the **estimation error**, and measures the performance of hypothesis h with respect to best-in-class hypothesis.

The definition of agnostic PAC-learning is also based on the estimation error. The estimation error of the hypothesis h_S returned by the algorithm $\mathcal{A}$ after training on a sample S, can sometimes be bounded in terms of the generalization error.

Example 4.6 (Empirical risk minimization (ERM)). Let h_T^E denote the hypothesis $h \in H$ that minimizes the empirical risk for the labeled sample T. In particular, $\hat{R}_{h_T^E} \leqslant R(h^*)$ and we can write

$$R(h_T^E) - R(h^*) = R(h_T^E) - \hat{R}(h_T^E) + \hat{R}(h_T^E) - R(h^*) \leqslant R(h_T^E) - \hat{R}(h_T^E) + \hat{R}(h^*) - R(h^*) \leqslant 2 \sup_{h \in H} \left|R(h) - \hat{R}(h)\right|.$$

The upper bound can be bounded by the learning bounds and increases with the size of the hypothesis set $|H|$, while the Bayes error $R(h^*)$ decreases with $|H|$.

4.4 Model selection

Example 4.7 (Structural risk minimization (SRM)). Consider an infinite sequence of hypothesis sets with increasing sizes $H_n \subset H_{n+1}$ for all $n \geqslant 0$. For each H_n, we can find the ERM solution h_n^E and complexity term $c(H_n, m)$. Then,

$$h_T^S = \arg \min_{h \in H_n, n \in \mathbb{N}} (\hat{R}_T(h) + c(H_n, m)).$$

If $\hat{R}_T(h) = 0$ for some $h \in H_n$, then $\hat{R}_T(h) = 0$ for all $H_m,\ m \geqslant n$.

Example 4.8 (Regularized risk minimization). An alternative family of algorithms is based on a more straightforward optimization that consists of minimizing the sum of the empirical error and a regularization term that penalizes more complex hypotheses. The regularization term is typically defined as $\|h\|^2$ for some norm $\|\cdot\|$ when H is a vector space, and

$$h_T^R = \arg\min_{h \in H} \hat{R}_T(h) + \lambda \|h\|^2,$$

where $\lambda \geqslant 0$ is a regularization parameter, which can be used to determine the trade-off between empirical error minimization and control of the complexity. In practice, λ is typically selected using n-fold cross-validation.

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 $\sigma^2 = \sum_{i=1}^m (b_i - a_i)^2$. Then, for any $\varepsilon > 0$ and $S_m \triangleq \sum_{i=1}^m X_i$, we have*

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

Proof. From the definition of indicator sets and for any increasing function $\phi : \mathbb{R} \to \mathbb{R}$, we can write for any random variable X

$$\phi(X) \geqslant \phi(X)\mathbb{1}_{\{X \geqslant \varepsilon\}} = \phi(X)\mathbb{1}_{\{\phi(X) \geqslant \phi(\varepsilon)\}} \geqslant \phi(\varepsilon)\mathbb{1}_{\{X \geqslant \varepsilon\}}.$$

Taking the random variable $S_m - \mathbb{E}[S_m]$ and $\phi(x) = e^{tx}$, and taking expectation on both sides, we get the Chernoff bound

$$\begin{aligned} P\{S_m - \mathbb{E}S_m \geqslant \varepsilon\} &\leqslant e^{-t\varepsilon}\mathbb{E}[\exp(t(S_m - \mathbb{E}S_m))] = e^{-t\varepsilon}\prod_{i=1}^{m}\mathbb{E}[\exp(t(X_i - \mathbb{E}X_i))] \\ &\leqslant e^{-t\varepsilon}\prod_{i=1}^{m}\exp(t^2(b_i - a_i)^2/8) = \exp\left(-t\varepsilon + \frac{t^2\sigma^2}{8}\right) \leqslant \exp\left(-\frac{2\varepsilon^2}{\sigma^2}\right). \end{aligned}$$

The first equality follows from the i.i.d. nature of $(X_i : i \in [m])$, the following inequality follows from Lemma A.1, the equality follows from the definition of σ^2, and the last inequality from $t^* = \frac{4\varepsilon}{\sigma^2}$. □