Machine Learning, Spring 2019
EN 2105
MWF 9-9:50am

Chao Lan
EN 4084A
clan@uwyo.edu

Sample Complexity and PAC Theory

Sample complexity is the number of training instances sufficient for learning an accurate model. It can be quantified by the Probably Approximately Correct (PAC) theory.

Consider classification task. Recall the generalization error of f is

$$er(f) = \Pr_{x \sim D}[f(x) \neq y] = E_{x \sim D}[\mathbf{1}_{f(x) \neq y}], \tag{1}$$

and the empirical error of f on sample S is

$$\hat{er}_S(f) = \frac{1}{|S|} \sum\nolimits_{(x,y) \in S} \mathbf{1}_{f(x) \neq y}. \tag{2}$$

[*Exercise*] Prove $E_{S \sim D^m}[\hat{er}(f)] = er(f)$.

A (candidate) model f is a hypothesis, and the set of hypotheses forms a hypothesis set $\mathcal{F}$. A target hypothesis f_* is a hypothesis one wants to learn.[1]

Let C be a collection of target hypotheses. We say C is PAC learnable if there exists a learning algorithm such that for any $f_* \in C$, ϵ, $\delta > 0$ and sample distribution D, if f is learned from a sample S of size $m = ploy(1/\epsilon, 1/\delta)$, then

$$\Pr_{S \sim D^m}[er(f) \leq \epsilon] \geq 1 - \delta. \tag{3}$$

PAC theory is called 'probably' because error is bounded in probability, and is called 'approximately correct' because it only gives a bound of the error (instead of identifying the error). PAC theory is distribution-free, in a sense that it applies to any sample distribution D.

Example 1: PAC Bound for Finite Hypothesis Space in the Consistent Case

Finite hypothesis space means $\mathcal{F}$ is finite. Consistent case means $f_* \in \mathcal{F}$. We say f is consistent on S if $\hat{er}_S(f) = 0$. The following theorem specifies the sample complexity of supervised learning under the PAC framework.

Theorem 1. *Let $\mathcal{F} : X \to Y$ be a finite hypothesis set, $S \subseteq X$ be an i.i.d. sample, and $f_S \in \mathcal{F}$ be any hypothesis consistent on S. For any ϵ, δ, $er(f_S) \leq \epsilon$ with probability at least $1 - \delta$ if*

$$|S| \geq \frac{1}{\epsilon}\left(\log|\mathcal{F}| + \log\frac{1}{\delta}\right). \tag{4}$$

This implies with probability at least $1 - \delta$,

$$er(f_S) \leq \frac{1}{|S|}\left(\log|\mathcal{F}| + \log\frac{1}{\delta}\right). \tag{5}$$

[1] It is often assumed $y = f_*(x)$. It is not always true that $f_* \in \mathcal{F}$.

Proof. The theorem says consistent hypotheses are likely to be accurate. We can prove this by bounding the chance that consistent hypotheses are not accurate.

First, for any fixed f, we have

$$\begin{aligned}
&\mathrm{Pr}_S\{f : \hat{er}_S(f) = 0 \wedge er(f) > \epsilon\} \\
&= \mathrm{Pr}\{f : \hat{er}_S(f) = 0 \mid er(f) > \epsilon\} \cdot \mathrm{Pr}\{er(f) > \epsilon\} \\
&\leq \mathrm{Pr}\{f : \hat{er}_S(f) = 0 \mid er(f) > \epsilon\} \\
&\leq (1-\epsilon)^{|S|}
\end{aligned} \tag{6}$$

[*Exercise*] Derive (6).

Next, we extend (6) to $\mathcal{F}$. Define set $B = \{f;\ \hat{er}_S(f) = 0 \wedge er(f) > \epsilon\}$. Then,

$$\begin{aligned}
&\mathrm{Pr}_S\{f \in B\} \\
&= \mathrm{Pr}\{f_1 \in B \vee f_2 \in B \vee \ldots \vee f_{|\mathcal{F}|} \in B\} \\
&\leq \sum\nolimits_{k=1}^{|\mathcal{F}|} \mathrm{Pr}\{f_k \in B\} \\
&\leq \sum\nolimits_{k=1}^{|\mathcal{F}|} (1-\epsilon)^{|S|} \\
&= |\mathcal{F}| \cdot (1-\epsilon)^{|S|} \\
&\leq |\mathcal{F}| \cdot e^{-\epsilon|S|},
\end{aligned} \tag{7}$$

where the first inequality is based on the union bound, and the last inequality is based on the fact that any $a \in [0,1]$ has

$$1 - a \leq e^{-a}. \tag{8}$$

[*Exercise*] Verify (7) and (8).

(7) suggests bounding $|\mathcal{F}|e^{-\epsilon|S|}$ suffices for bounding $\mathrm{Pr}\{f \in B\}$. Setting $|\mathcal{F}|e^{-\epsilon|S|} \leq \delta$ and solving for $|S|$, we have

$$|S| \geq \frac{1}{\epsilon}\left(\log|\mathcal{F}| + \log\frac{1}{\delta}\right). \tag{9}$$

[*Exercise*] Derive (9). □

Example 2: PAC Bound for Finite Hypothesis Space in the Inconsistent Case

Inconsistent case means $f_* \notin \mathcal{F}$. In this case, arguments in Example 1 are inapplicable (e.g., we may not find any consistent f_S). We can use the following bounding technique.

Theorem 2 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_m$ be independent random variables, where $X_i \in [a_i, b_i], \forall i \in [1, m]$, and let $S_m = \sum_{i=1}^m X_i$. Then, for any ϵ, there are*

$$\begin{aligned}
\mathrm{Pr}[S_m - E[S_m] \geq \epsilon] &\leq e^{-2\epsilon^2/\sum_{i=1}^m (b_i - a_i)^2}, \\
\mathrm{Pr}[S_m - E[S_m] \leq -\epsilon] &\leq e^{-2\epsilon^2/\sum_{i=1}^m (b_i - a_i)^2}.
\end{aligned} \tag{10}$$

Applying Hoeffding's inequality, we have for any fixed f, there is

$$\mathrm{Pr}\{\ |\hat{er}_S(f) - er(f)| \geq \epsilon\ \} \leq 2e^{-2|S|\epsilon^2}. \tag{11}$$

[*Exercise*] Derive (11).

Extending (11) to $\mathcal{F}$ by a union bound, we have the following result.

Theorem 3. *Let $H : X \to \{0, 1\}$ be a finite hypothesis space, $S \subseteq X$ be an i.i.d. sample. Then, for any $\delta > 0$, with probability at least $1 - \delta$, every $f \in \mathcal{F}$ satisfies*

$$er(f) \leq \hat{er}_S(f) + \sqrt{\frac{\log |\mathcal{F}| + \log \frac{2}{\delta}}{2|S|}}. \tag{12}$$

[*Exercise*] Prove (3).

Example 3: PAC Bound for Infinite Hypothesis Space

If $\mathcal{F}$ is infinite, arguments in Examples 1 and 2 are inapplicable. In this case, we can measure the complexity of $\mathcal{F}$ using the Vapnik-Chervonenkis (VC) Dimension.

Consider binary classification. Given a sample S, there are $2^{|S|}$ ways to label its instances. We say S is shattered by $\mathcal{F}$ if every of its $2^{|S|}$ labeling ways is realized by some $f \in \mathcal{F}$. The maximum size of an arbitrary S that can be shattered by $\mathcal{F}$ is the VC-Dimension of $\mathcal{F}$.

[*Discussion*] Give an example of S being shattered by $\mathcal{F}$.

[*Discussion*] What is the VC-Dimension of $\mathcal{F} = \mathbb{R}^2$?

Theorem 4. *Let $H : X \to \{0, 1\}$ be a hypothesis set with VC dimension d, and $S \subseteq X$ be an i.i.d. sample. Then, for any $\delta > 0$, with probability at least $1 - \delta$, every $f \in \mathcal{F}$ satisfies*

$$er(f) \leq \hat{er}_S(f) + \sqrt{\frac{2d \log \frac{em}{d}}{|S|}} + \sqrt{\frac{\log \frac{1}{\delta}}{2|S|}}. \tag{13}$$

The proof[2] is very complicated. See more in [Foundations of Machine Learning, Ch 2,3].

[2] `http://www.cs.princeton.edu/courses/archive/spring13/cos511/scribe_notes/0219.pdf`