6.858/18.428 Machine Learning	Lecture 8 : October 5, 1994
Lecturer: Ron Rivest	*Scribe: Lisa Zhang*

In today's lecture we show the PAC learnability of infinite concept classes of finite VC-dimension.

8.1 Background and Definitions

Last time we introduced notations $\Pi_C(S)$ and $\Pi_C(m)$ defined by

$$\begin{aligned}\Pi_C(S) &= \{c \cap S : c \in C\}, \\ \Pi_C(m) &= \max\{|\Pi_C(S)| : |S| = m\}\end{aligned}$$

for any concept class C over instance class X and any $S \subseteq X$. The value of $\Pi_C(m)$ is upper bounded by function Φ, that is $\Pi_C(m) \leq \Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$.

To show an infinite concept class C of finite VC-dimension d is PAC learnable, the learning algorithm L

1. draws a large sample S of size $|S| = m$, where m is a function of d, ε and δ.
2. returns any concept $h \in C$ that is consistent with the sample.

We would like a similar Occam algorithm for infinite concept classes.

Hypothesis h is *bad* if $er(h) \geq \varepsilon$; otherwise h is *good*. Let $c\Delta h = \{x : c(x) \neq h(x)\}$ be the region where concept c and hypothesis h differ. Notice that $Pr_{x \in D}[x \in c\Delta h] = er(h)$. For a fixed target concept $c \in C$, let $\Delta(c) = \{h\Delta c : h \in C\}$ be the set of *error regions* with respect to c and C. Furthermore, let $\Delta_\varepsilon(c) = \{h\Delta c : h \in C \wedge Pr_{x\in D}[x \in h\Delta c] \geq \varepsilon\}$ be the error regions with weight at least ε under the fixed target distribution D. We can now make the following definition:

Definition 1 *For any $\varepsilon > 0$ a set S is an* ε-net *for $\Delta(c)$ if every region in $\Delta_\varepsilon(c)$ is hit by a point in S, that is $\forall r \in \Delta_\varepsilon(c)$, $S \cap r \neq \emptyset$.*

As an example, let us look at the concept class C consisting of all closed intervals on $[0,1]$ under the uniform distribution. For any hypothesis h, $h\Delta c = I_1 \cup I_2$, where I_1 and I_2 are two closed intervals and $I_1 = \{x : x \in h \wedge x \notin c\}$, $I_2 = \{x : x \notin h \wedge x \in c\}$. Under the uniform density $Pr_{x\in D}[x \in I]$ for any interval I is just the length of I. So if h is bad, either $|I_1|$ or $|I_2|$ is at least $\varepsilon/2$. Hence the set $S = \{x = k\varepsilon/2 : k = 0,1,\ldots,\lceil 2/\varepsilon\rceil\}$ forms an ε-net.

8.2 The Double Sampling Method

Notice that any hypothesis h consistent with an ε-net is good. Hence if we can upper bound the probability that a random sample S fails to form an ε-net, then we have upper bounded the probability of $er(h) \geq \varepsilon$. In order for $Pr[S$ to be an ε-net$] \geq 1-\delta$ we adopt the method of double sampling.

Let S_1 be a random sample of size m, and let A be the event that S_1 fails to be an ε-net. Clearly, we want $Pr[A] \leq \delta$. Let S_2 be a second random sample of size m. If event A happens, then there exists region $r \in \Delta_\varepsilon(c)$ such that $S_1 \cap r = \emptyset$ by the definition of the ε-net. For a fixed region r missed by S_1, each element in S_2 has probability $\varepsilon' \geq \varepsilon$ to hit r. By the multiplicative form of the Chernoff bound (See Appendix)

$$
\begin{aligned}
& Pr[|S_2 \cap r| > \frac{\varepsilon m}{2}] \\
\geq\ & Pr[|S_2 \cap r| > \frac{\varepsilon' m}{2}] \\
=\ & 1 - Pr[|S_2 \cap r| \leq \frac{\varepsilon' m}{2}] \\
\geq\ & 1 - e^{-\frac{\varepsilon' m}{8}} \\
\geq\ & \frac{1}{2}
\end{aligned}
\tag{8.1}
$$

for $m \geq \frac{8}{\varepsilon}\ln 2 \geq \frac{8}{\varepsilon'}\ln 2$. Now let B be the event that A happens and S_2 has at least $\frac{\varepsilon m}{2}$ hits in some region $r \in \Delta_\varepsilon(c)$ that is missed by S_1. We have just proved that $Pr[B|A] \geq \frac{1}{2}$. Since $Pr[B] = Pr[A \wedge B] = Pr[B|A]Pr[A]$, $Pr[A] \leq 2Pr[B]$. Therefore, we want $Pr[B] \leq \frac{\delta}{2}$.

Equivalently, B is the event that there is some $r \in \Pi_{\Delta_\varepsilon(C)}(S_1 \cup S_2)$ such that $|r| \geq \varepsilon m/2$ and $r \cap S_1 = \emptyset$. Thus, instead of directly analyzing the probability of event A by considering all regions of the infinite class $\Delta_\varepsilon(c)$ that S_1 might miss, we can now analyze the probability of event B by only considering the regions of $\Pi_{\Delta_\varepsilon(C)}(S_1 \cup S_2)$.

To bound $Pr[r \in \Pi_{\Delta_\varepsilon(C)}(S_1 \cup S_2) : |r| \geq \varepsilon m/2 \wedge r \cap S_1 = \emptyset]$, we draw a random sample of size $2m$ and randomly divide it into S_1 and S_2 of equal size. The resulting distribution of S_1 and S_2 is the same as drawing 2 samples of size m randomly and independently. The probability for a fixed region $r \in S_1 \cup S_2$ of size $l = |r| \geq \varepsilon m/2$ to be entirely in S_2 is $\frac{\binom{m}{l}}{\binom{2m}{l}} \leq 2^{-l} \leq 2^{-\varepsilon m/2}$. Therefore,

$$\begin{aligned}
Pr[B] &= Pr[r \in \Pi_{\Delta_\varepsilon(C)}(S_1 \cup S_2) : |r| \geq \varepsilon m/2 \wedge r \cap S_1 = \emptyset] \\
&\leq |\Pi_{\Delta_\varepsilon(c)}(S_1 \cup S_2)| 2^{-\varepsilon m/2} \\
&\leq |\Pi_{\Delta(c)}(S_1 \cup S_2)| 2^{-\varepsilon m/2}.
\end{aligned}$$

We need the following lemma to finish the analysis:

Lemma 1 *VC-dimension($\Delta(c)$)=VC-dimension(C).*

To see the correctness of Lemma 1, for any set S we can map each element $c' \in \Pi_C(S)$ to $c'\Delta(c \cup S) \in \Pi_{\Delta(c)}(S)$. Since this is a bijective mapping of $\Pi_C(S)$ to $\Pi_{\Delta(c)}(S)$ for any S, $|\Pi_C(S)| = |\Pi_{\Delta(c)}(S)|$, and VC-dimension($\Delta(c)$)=VC-dimension($C$) follows.

Hence we have

$$\begin{aligned}
Pr[B] &\leq |\Pi_{\Delta(c)}(S_1 \cup S_2)| 2^{-\varepsilon m/2} \\
&\leq \Phi_d(2m) 2^{-\varepsilon m/2} \\
&\leq \left(\frac{2em}{d}\right)^d 2^{-\varepsilon m/2}.
\end{aligned}$$

Lemma 2 *The inequality $\left(\frac{2em}{d}\right)^d 2^{-\varepsilon m/2} \leq \frac{\delta}{2}$ is satisfied for $m = \max(\frac{4}{\varepsilon}\lg\frac{2}{\delta}, \frac{8d}{\varepsilon}\lg\frac{13}{\epsilon})$.*

Proof: To show that m satisfies the given inequality, we take logs and verify that it satisfies

$$d \lg \frac{2em}{d} - \frac{\varepsilon m}{2} \leq \lg \frac{\delta}{2}.$$

That is, m satisfies

$$m \geq \frac{2}{\varepsilon} \lg \frac{2}{\delta} + \frac{2d}{\varepsilon} \lg \frac{2em}{d}.$$

Since $\frac{m}{2} \geq \frac{2}{\varepsilon}\lg\frac{2}{\delta}$ by our choice of m, we simply need to verify that $\frac{m}{2} \geq \frac{2d}{\varepsilon}\lg\frac{2em}{d}$. By plugging in $m = \frac{8d}{\varepsilon}\lg\frac{13}{\varepsilon}$, we obtain the following equivalent set of inequalities:

$$\frac{m}{2} \;\geq\; \frac{2d}{\varepsilon} \lg \frac{2em}{d}$$

$$\begin{aligned}
\frac{4d}{\varepsilon}\lg\frac{13}{\varepsilon} &\geq \frac{2d}{\varepsilon}\lg\left(\frac{2e}{d}\frac{8d}{\varepsilon}\lg\frac{13}{\varepsilon}\right)\\
2\lg\frac{13}{\varepsilon} &\geq \lg\left(\frac{16e}{\varepsilon}\lg\frac{13}{\varepsilon}\right)\\
\left(\frac{13}{\varepsilon}\right)^2 &\geq \frac{16e}{\varepsilon}\lg\frac{13}{\varepsilon}\\
\frac{13^2}{16e\varepsilon} &\geq \lg\frac{13}{\varepsilon}
\end{aligned}$$

It can be easily verified that the last inequality holds for any $\varepsilon \leq 1$. Notice that $\frac{m}{2}$ grows faster than $\frac{2d}{\varepsilon}\lg\frac{2em}{d}$. Since $m = \frac{8d}{\varepsilon}\lg\frac{13}{\epsilon}$ satisfies $\frac{m}{2} \geq \frac{2d}{\varepsilon}\lg\frac{2em}{d}$, any $m \geq \frac{8d}{\varepsilon}\lg\frac{13}{\epsilon}$ satisfies the inequality. Thus, $m = \max(\frac{4}{\varepsilon}\lg\frac{2}{\delta}, \frac{8d}{\varepsilon}\lg\frac{13}{\epsilon})$ satisfies the original inequality.
■

Combining Lemma 2 with the bound $m \geq \frac{8}{\varepsilon}\ln 2$ obtained from inequality 8.1, we have proved the following result:

Theorem 1 *Let C be any concept class of VC-dimension d. Let L be any algorithm that takes as input a set S of m labeled examples of a concept in C, and produces as output a concept $h \in C$ that is consistent with S. Then L is PAC learnable for C provided it is given a random sample of m examples from EX(c,D), where m obeys*

$$m \geq \max\left(\frac{8}{\varepsilon}\lg\frac{2}{\delta}, \frac{8d}{\varepsilon}\lg\frac{13}{\epsilon}\right).$$

Appendix

The Chernoff bound is a fundamental result from probability theory that appears repeatedly in theoretical computer science.

Theorem 2 *Let $X_1, \ldots, X_m$ be a sequence of m independent Bernoulli trials (coin flips), each with probability of heads $E[X_i] = p$. Let $S = X_1 + \ldots + X_m$ be a random variable indicating the total number of heads, so $E[S] = pm$. Then for $0 \leq \gamma \leq 1$, the following bounds hold:*

(Additive Form)

$$Pr[S > (p+\gamma)m] \leq e^{-2m\gamma^2}$$

and

$$Pr[S < (p-\gamma)m] \leq e^{-2m\gamma^2}.$$

(Multiplicative Form)

$$Pr[S > (1+\gamma)pm] \leq e^{-mp\gamma^2/3}$$

and

$$Pr[S < (1-\gamma)pm] \leq e^{-mp\gamma^2/2}.$$

References

[1] M. Kearns and U. Vazirani. An Introduction to Computational Learning Theory. 57-62.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4): 929-965, 1989.