6.858/18.428 Machine Learning	**Lecture 10: October 17, 1994**
Lecturer: Ron Rivest	*Scribe: Mikael Goldmann*

10.1 Outline

- A lower bound on sample complexity through VC-dimension.
- Cover's coin problem.

10.2 A lower bound on the sample complexity through VC-dimension

In previous lectures it has been shown that the VC-dimension of a concept class gives an upper bound on the number of samples needed to PAC-learn concepts from the class. In particular it has been shown that for VC-dimension d it is possible to PAC-learn with parameters δ and ϵ given that the number of samples m is at least

$$m \geq c_0 \left(\frac{d}{\epsilon} \log \frac{1}{\epsilon} + \frac{1}{\epsilon} \log \frac{1}{\delta} \right) \tag{10.1}$$

where c_0 is a fixed constant. So, as long as VC-dim($\mathcal{C}$) is finite, it is possible to PAC-learn concepts from $\mathcal{C}$ even though $|\mathcal{C}|$ might be infinite.

One might ask if there are concept classes of infinite VC-dimension that can be learned from finite sample size, or if at least the upper bound of (10.1) can be improved in any substantial way to, say, $O(d \log(1/\epsilon))$. The answer to both these questions is "No." It is pretty clear that one cannot PAC-learn a concept class $\mathcal{C}$ of infinite VC-dimension. Say that we have an algorithm that makes m queries to the examples oracle (m depends only on δ and ϵ), and let Y be a subset of the domain of size $\geq 2m$ that is shattered by $\mathcal{C}$. After m examples A knows nothing about at least half the points in Y, and therefore it cannot predict the labels on those points better than a random coin toss. In fact, for finite VC-dimension equation (10.1) is nearly tight, and we will show that for VC-dimension d, achieving error rate ϵ requires $c_1 d/\epsilon$ samples for some fixed constant c_1.

Let $X = \{x_1, \ldots, x_d\}$ and let $\mathcal{C} = 2^X$. Obviously, VC-dim$(\mathcal{C}) = d$. We will show that PAC-learning $\mathcal{C}$ requires $\Omega(d/\epsilon)$ examples.

Theorem 1 *Let $X = \{x_1, \ldots, x_d\}$ and let $\mathcal{C} = 2^X$. Any algorithm A that PAC-learns $\mathcal{C}$ with parameters ϵ and $\delta \leq 1/15$ must use more than $(d-1)/(64\epsilon)$ queries to the examples oracle.*

This is in fact a general result since PAC-learning $\mathcal{C}$ is a sub-problem of PAC-learning any concept class of VC-dimension d.

Corollary 1 *Let $\mathcal{C}$ be an arbitrary concept class and let $d =$ VC-dim$(\mathcal{C})$. To PAC-learn $\mathcal{C}$ with parameters ϵ and $\delta \leq 1/15$ requires more than $(d-1)/(64\epsilon)$ queries to the examples oracle.*

Proof of Theorem 1: Let $m = (d-1)/(64\epsilon)$, and A be an algorithm that makes at most m queries to the examples oracle and then produces a hypothesis h. We need to show that there are a distribution p on X and a concept $c \in \mathcal{C}$ such that $er(h) > \epsilon$ with probability at least 1/15.

Let us first fix p independently of A:

$$p(x_1) = 1 - 16\epsilon$$
$$p(x_2) = p(x_3) = \cdots = p(x_d) = \frac{16\epsilon}{d-1}$$

In what follows S is a random sample of X that we get by m independent draws from p. We want to establish that there is a c so that $\Pr_S[er(h) > \epsilon] > \frac{1}{15}$.

Let $X' = \{x_2, \ldots, x_d\}$. For any fixed $c \in \mathcal{C}$ and hypothesis h, let $er'(h) = \Pr[c(x) \neq h(x) \wedge x \in X']$. For technical reasons, it is easier to prove that $\Pr_S[er'(h) > \epsilon] > 1/15$, which is enough since $er'(h) \leq er(h)$.

We pick a random $c \in \mathcal{C}$ and show that with positive probability c is hard to learn for A, thereby showing that there must be some fixed c that is hard to learn for A.

Let B be the event that A's sample S contains less than $(d-1)/2$ points in X'. We establish three things:

$$\Pr_S[B] \geq 1/2 \qquad (10.2)$$
$$\mathrm{E}_{c,S}[\, er'(h) \mid B\,] > 4\epsilon \qquad (10.3)$$
$$\mathrm{E}_S[\, er'(h) \mid er'(h) > \epsilon\,] \leq 16\epsilon \text{ for any fixed } c. \qquad (10.4)$$

Inequality (10.4) is obvious since for any h we have $er'(h) \le \Pr[x \in X'] = 16\epsilon$. However, if we consider $er(h)$ rather than $er'(h)$, then (10.4) does not necessarily hold, which is the reason for introducing $er'(h)$. Both (10.2) and (10.3) make intuitive sense: if A only sees a few samples it is likely to miss a large fraction of X' (this is (10.2)), and if A knows nothing about a large fraction of X' then the error rate of the hypothesis is likely to be high (this is (10.3)). Before formally proving (10.2) and (10.3) we show that they imply that there is a $c_* \in \mathcal{C}$ so that $\Pr_S[er'(h) > \epsilon] > 1/15$.

First we use (10.2) and (10.3) to get a lower bound on $\mathrm{E}_{c,S}[er'(h)]$.

$$\mathrm{E}_{c,S}[er'(h)] \ge \Pr_S[B] \cdot \mathrm{E}_{c,S}[\, er'(h) \mid B\,] > \frac{1}{2} \cdot 4\epsilon = 2\epsilon.$$

In particular, there is some $c_* \in \mathcal{C}$ such that $\mathrm{E}_S[er'(h)] > 2\epsilon$. We take c_* as the target concept and show that A is likely to produce a hypothesis with high error rate. Using conditional expectation we get

$$\begin{aligned} 2\epsilon &< \mathrm{E}_S[er'(h)] \\ &= \Pr_S[er'(h) > \epsilon] \cdot \mathrm{E}_S[\, er'(h) \mid er'(h) > \epsilon\,] \\ &\quad + (1 - \Pr_S[er'(h) > \epsilon]) \cdot \mathrm{E}_S[\, er'(h) \mid er'(h) \le \epsilon\,]. \end{aligned}$$

Next we apply (10.4) to get

$$\begin{aligned} 2\epsilon < \mathrm{E}_S[er'(h)] &\le \Pr_S[er'(h) > \epsilon] \cdot 16\epsilon + (1 - \Pr_S[er'(h) > \epsilon]) \cdot \epsilon \\ &= 15\epsilon \Pr_S[er'(h) > \epsilon] + \epsilon, \end{aligned}$$

which implies $\Pr_S[er'(h) > \epsilon] > 1/15$.

It remains to prove (10.2) and (10.3).

Proof of (10.2). Let T be the number of times A gets a sample point in X'. Clearly, $\mathrm{E}[T] = 16\epsilon m = (d-1)/4$. We have $\Pr_S[B] \ge 1 - \Pr[T \ge (d-1)/2]$, and by Markov's inequality,[1] $\Pr[T \ge (d-1)/2] \le 1/2$.

Proof of (10.3). Let S be the set of sample points that A gets. Choosing a random c is equivalent to flipping a fair coin for each point in X to determine its label. Since h is independent of the labeling of $X' - S$, the contribution to $er'(h)$ is expected to be $16\epsilon/(2(d-1))$ for each point in $X' - S$. When B occurs, we have $|X' - S| > (d-1)/2$ and thus the expected value of $er'(h)$ given B is strictly greater than 4ϵ.

The proof is now complete. ■

[1] Markov's inequality: if X is a positive random variable, then $\Pr[X \ge t] \le \mathrm{E}[X]/t$.

10.3 Cover's coin problem

Cover's coin problem is somewhat connected to the idea of "learning the bias of a coin." The precise statement of the problem is as follows:

Cover's coin problem
INPUT: A coin.
OUTPUT: "rational" if Pr[heads] is rational and "irrational" otherwise.

At first glance the question might look strange, especially since the maximum likelihood estimate $\bar{p}$ of $p = \Pr[\text{heads}]$ after a finite number of coin flips is always rational ($\bar{p} = \#\text{heads}/\#\text{flips}$). On the face of it, the problem appears to be impossible. In order to get a better understanding, let us look at a simpler problem and see in what sense it can be solved.

Fair coin problem
INPUT: A coin.
OUTPUT: "fair" if Pr[heads] = 1/2 and "biased" otherwise.

It is possible to devise an algorithm F for the "Fair coin problem" that flips the coin a few times, guesses "fair" or "biased," then flips again and guesses, and so on. With probability 1, the infinite sequence of guesses will contain only a finite number of mistakes.

```
F:
    For i = 1, 2, 3, ...
        δ_i = 2^{-i}
        N_i = 4^i i
        Flip N_i times and let H be the number of "heads" observed
        If |H − N_i/2| > δ_i N_i then output "fair" else output "biased"
    End for
```

Consider first the case that the coin is fair, and let B_i be the event that the i^{th} output is "biased." Using the additive version of the Chernoff bounds we get

$$\Pr[B_i] = \Pr[|H - N_i/2| > \delta_i N_i] \leq 2e^{-2N_i\delta_i^2} = 2e^{-2i}.$$

The following result is very useful when studying the behavior of F.

Theorem 2 (Borel–Cantelli lemmas) *Let $A_1, A_2, \ldots$ be a countable sequence of events, and let $A = \cap_{n\geq 1} \cup_{m\geq n} A_i$ be the event that infinitely many of the A_i's occur.*

Then

$$\Pr[A] = 0 \quad \textit{if} \quad \sum_{i\geq 1} \Pr[A_i] < \infty,$$
$$\Pr[A] = 1 \quad \textit{if} \quad \sum_{i\geq 1} \Pr[A_i] < \infty \quad \textit{and } A_1, A_2, \ldots \textit{ are independent events.}$$

In our case $\sum_{i\geq 1} \Pr[B_i] < \infty$, which implies that F ouputs "fair" all but finitely often.

The second case is when the probability of heads is $p \neq 1/2$. In this case, with probability 1, it only happens finitely many times that $H/N_i \notin [p - \delta_j, p + \delta_j]$. When j is sufficiently large, the intervals $[p - \delta_j, p + \delta_j]$ and $[\frac{1}{2} - \delta_j, \frac{1}{2} + \delta_j]$ are disjoint, so F will say "fair" only finitely many times with probability 1.

Let us now return to Cover's coin problem. We will construct an algorithm with the following properties. If $p = \Pr[\text{heads}]$ is rational, then with probability 1 the algorithm will output "irrational" only a finite number of times. For all irrational p, except for a set of measure 0, the algorithm will output "rational" only finitely many times with probability 1.

What we would like to do is to put shrinking intervals around all rational numbers, but these intervals would of course cover $[0, 1]$ and any irrational number would always be close to *some* rational number. The key is to consider a slowly growing, but always finite, set of rationals. If p is rational, then it will eventually become a member of the set, and with probability 1 the empiric average $\bar{p}$ will be close to p all but finitely often. On the other hand, If p is irrational, the union of the intervals around the rationals in the finite set shrinks in size, and eventually $\bar{p}$ should always fall outside this union. This is almost what happens. The problem is that while the union of intervals has smaller and smaller measure, we keep adding more rationals to our finite set, and therefore an irrational number that is not in an interval at some point may be too close to a rational number that is about to be added to the set.

While the ideas are the same as in the "Fair coin problem," the technical details are a bit more involved. Let $q_1, q_2, \ldots$ be an arbitrary enumeration of $\mathbb{Q} \cap [0, 1]$, and $\mathbb{I}_k$ be $\{q_1, q_2, \ldots, q_k\}$. Define

$$\begin{aligned} \mathbb{I}_k(\delta) &= \cup_{i=1}^{k} [q_i - \delta, q_i + \delta] \\ N_0 &= 0 \\ N_k &= 2^k \\ \delta_k &= \sqrt{(\ln k)/2^k} \end{aligned}$$

```
G:
    h = 0
    n = 0
    guess = "rational"
    For k = 1 to ∞
        For n = N_{k-1} + 1 to N_k
            h_n = flip
            h = h + h_n
            p̄_n = h/n
            output guess
        End for
        If p̄_{N_k} ∈ I_k(δ_k) then guess = "rational" else guess = "irrational"
    End for
```

The following lemma is useful when analyzing G.

Lemma 1 *If* $p = \Pr[\text{heads}]$ *then* $\bar{p}_{N_k} \in [p - \delta_k, p + \delta_k]$ *for all but finitely many* k *with probability* 1.

Proof: Let A_k be the event that $p_{N_k} \notin [p - \delta_{N_k}, p + \delta_{N_k}]$. Using the additive version of the Chernoff bounds we

$$\Pr[A_k] \leq 2e^{-2N_k\delta_k^2} = 2/k^2.$$

Therefore $\sum_{k \geq 1} \Pr[A_k] < \infty$, and the lemma follows by the Borel–Cantelli lemmas. ■

We consider three cases.

Case 1: p is rational. There is an l such that $p = q_l \in \mathbb{I}_k$ for all $k \geq l$. By Lemma 1 $\bar{p}_k \in [q_l - \delta_k, q_l + \delta_k]$ for all but finitely many k with probability 1, and $[q_l - \delta_k, q_l + \delta_k] \subset \mathbb{I}_k(\delta_k)$ for $k \geq l$. This implies that G will output "rational" all but finitely often with probability 1.

Case 2: $p \in \mathbb{I}_k(2\delta_k)$ for finitely many k. In this case there is some l so that $[p - \delta_k, p + \delta_k] \cap \mathbb{I}_k(\delta_k) = \emptyset$ for $k \geq l$. Together with Lemma 1 this implies that the output will be "irrational" all but finitely often with probability 1.

Case 3: p is irrational, but $p \in \mathbb{I}_k(2\delta_k)$ for infinitely many k. This case is bad since the output may be "rational" infinitely often. However, we will show that the Lebesgue measure of the set of such p is 0.

First note that

$$\{\, p \mid p \in \mathbb{I}_k(2\delta_k) \text{ for infinitely many } k \,\} \subset \bigcup_{k \geq m} \mathbb{I}_k(2\delta_k) \quad \text{for all } m \geq 0.$$

Since $\lambda(\mathbb{I}_k(2\delta_k)) \leq 4k\delta_k$, we have

$$\lambda\left(\bigcup_{k \geq m} \mathbb{I}_k(2\delta_k)\right) \leq \sum_{k \geq m} 4k\delta_k.$$

The sum $\sum_{k \geq 1} k\delta_k$ converges, so $\sum_{k \geq m} 4k\delta_k$ tends to 0 as m tends to ∞. This implies that $\lambda\{\, p \mid p \in \mathbb{I}_k(2\delta_k) \text{ for infinitely many } k \,\} = 0$.

References

[1] T. M. Cover, On determining the irrationality of the mean of a random variable. *The Annals of Satistics*, 1(5):862–871, 1973.

[2] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–251, 1989.

[3] M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*, 52–64, 1994.