COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Senem Velipasalar

Lecture # 4
February 13, 2003

1 From Last Time

H is finite. Then with probability $\geq 1-\delta$, if $h \epsilon H$ is consistent then

$$err(h) \leq \frac{ln|H| + ln|\frac{1}{\delta}|}{m}. \tag{1}$$

We saw last time that this bound works if $|H|$ is finite. What if $|H|$ is infinite?

2 Intuition and examples

Even if we have infinitely many possible hypotheses, learning is possible from a finite sample.
Example 1:
Let's say we have 3 examples. Then there are infinitely many possible hypotheses but only four possible labelings. Labelings are also called ***behaviors*** or ***dichotomies***. In Fig. 1, all the possible labelings for the possible hypotheses are shown.

In such a case if we have m samples, there are $m+1$ possible labelings.

Example 2 - Learning Intervals: In this case, there are $\frac{m(m-1)}{2} + m + 1 = \binom{m}{2} + m + 1$ possible labelings, where $+m$ is for the intervals having just single points. As it can be seen, the number of labelings is $O(m^2)$ for this example.

3 An upper bound for err(h) when $|H|$ is not finite

3.1 Notation

The following notation was introduced:

$$S = < x_1, x_2, ..., x_m >,$$
$$\Pi_H(S) = \{< h(x_1), h(x_2), ..., h(x_m) >: h \epsilon H\},$$
$$\Pi_H(m) = \max_{\substack{S \\ |S|=m}} |\Pi_H(S)| \leq 2^m.$$

Note: $\Pi_H(S)$ is the set of all possible labelings for all possible hypotheses and $\Pi_H(m)$ is the number we computed in the above examples.

3.2 Finding the upper bound

For any H, there are 2 possible cases:

1. Either $\forall m, \Pi_H(m) = 2^m$, which is the worst case,

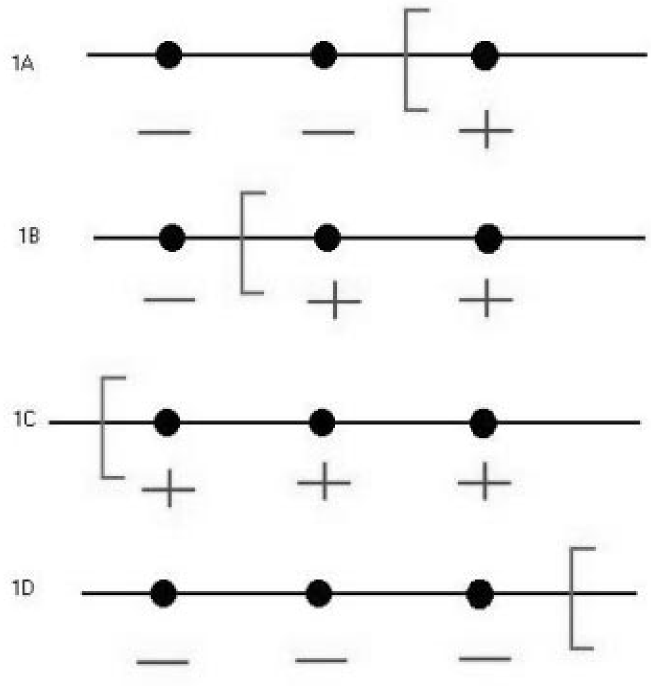

Figure 1: Possible labelings when we have 3 samples

2. or, $\Pi_H(m) = O(m^d)$, which is a really nice case. Here, d is the VC-dimension of H where VC stands for Vapnik-Chervonenkis. VC-dimension will be defined in next lecture.

Step 1: Derive an error bound in which $\ln|H|$ is replaced by $\ln|O(m^d)|$ so, we will get a result analogous to Occam's razor result.

Theorem: With probability $\geq 1-\delta$, $\forall h \epsilon H$, if h is consistent, then

$$err(h) \leq O\left(\frac{\ln \Pi_H(2m) + \ln(\frac{1}{\delta})}{m}\right) \tag{2}$$

Proof: First, we will try to show that with probability $\geq 1-\delta$

$$(\forall h \epsilon H : \text{h is consistent, err(h)} \leq \epsilon). \tag{3}$$

Let's define event B and Pr[B] as follows:

$$\Pr_S[\underbrace{\exists h \epsilon H : \text{h is consistent on S but } err(h) > \epsilon}_{eventB}].$$

Note that event B is the negation of the event defined in (3). We are trying to bound $\Pr_S[B]$. Because, if $\Pr_S[B] < \delta$, then Pr[event defined in (3)] $\geq 1-\delta$ which is what we want to show.

Trick: Replace the error with error on another sample. In this new sample, there will be finitely many errors we need to consider. Let

$$S' = \text{ second sample of m examples.}$$

The data is independent identically distributed. We will argue that it is unlikely to see many errors on one sample, and no errors on the second sample.

$$S =< x_1, x_2, ..., x_m > \text{ all i.i.d,}$$
$$S' =< x'_1, x'_2, ..., x'_m > \text{ all i.i.d,}$$
$$S; S' \text{ has 2m samples.}$$

NOTATION:

$M(h) = |\{i : h(x'_i) \neq c(x'_i)\}|$. (number of mistakes)

$B' \equiv \exists h \epsilon H$: h is consistent on S and $M(h) \geq \frac{m\epsilon}{2}$. (We have m samples and probability of making error for each sample is ϵ.)

Claim: $\Pr[B'|B] \geq \frac{1}{2}$ i.e. if you are in bad case B, the probability that you are in case B' is $\geq \frac{1}{2}$.

If you know B happens, i.e., if h is consistent on S and err(h) $> \epsilon$, then $M(h) \geq \frac{m\epsilon}{2}$ with probability $\geq \frac{1}{2}$ which implies $\Pr[B'|B] \geq \frac{1}{2}$. (This will be proven later.)

$$\begin{aligned} \Pr[B'] &\geq \Pr[B' \wedge B] \\ &= \Pr[B] \cdot \Pr[B'|B] \\ &\geq \frac{1}{2}\Pr[B] \end{aligned} \tag{4}$$

(4) implies $\Pr[B] \leq 2\Pr[B']$. So, if probability of event B' happening is small, then the probability of event B happening is also small. Thus, instead of bounding probability of event B, we can start working with event B' and bound its probability.

Experiment I: Draw S at random and then draw S' at random.

Experiment II: Draw S, S'. With probability 1/2 interchange x_i and x'_i and with probability 1/2 leave them as they are. Doing this will not change the sample distribution.

As Experiment I and Experiment II will give the same distribution of examples, we can work with experiment II. So,

FIX h, S, S'. We will try to bound $Pr[B'|S, S']$.
Recall, $B' \equiv \exists h \epsilon H : h$ is consistent on S and $M(h) \geq \frac{m\epsilon}{2}$.

$$\begin{array}{ccccccccc} & & x_1 & x_2 & \dots & & & \\ S & : & 0 & 1 & 0 & 1 & 0 & 0 \\ S' & : & 1 & 1 & 0 & 0 & 1 & 1 \\ & & x_1\prime & x_2\prime & \dots & & & \end{array}$$

S : *0 0 0 0 ...* means h is consistent with S.

If $\exists i$ such that both x_i and x'_i are 1, then there is no way we can have all zeros in S. So,

$$Pr[\text{h is consistent on S and } M(h) \geq \frac{m\epsilon}{2}] = 0. \tag{5}$$

If there are $M(h)$ i's where exactly one of x_i or x'_i is 1, then,

$$Pr[\text{h is consistent on S}] \leq 2^{-M(h)} \text{ (this is the probability of all the 1s ending up in } S'). \tag{6}$$

We can think of this as follows: If x_1 is 0 and x'_1 is 1, w.p 1/2 x_1 will remain 0. If x_2 is 0 and x'_2 is 0, w.p 1 x_2 will remain 0 ...etc. So; the probability of all x_i's being 0 is:

$$\frac{1}{2}.1.\frac{1}{2}.... = (\frac{1}{2})^{\text{number of i's for which only one of } x_i \text{ or } x'_i \text{ is } 1}$$

unless there is an i for which both x_i and x'_i are 1 in which case the probability is zero.

Let $H'(S)$ = one representative from H for each dichotomy in S. Then;

$$B' \equiv \left(\exists h \epsilon H'(S; S') : \underbrace{\text{h is consistent on S and } M(h) \geq \frac{m\epsilon}{2}}_{e(h)} \right)$$

$$\begin{aligned} Pr[B'|S, S'] &= Pr[\exists h \epsilon H'(S; S') : e(h)|S, S'] \\ &= Pr[e(h_1) \vee e(h_2) \vee \cdots \vee e(h_N)|S, S'] \\ &\leq \sum_{i=1}^{N} Pr[e(h_i)|S, S'] \\ &\leq |H'(S, S')| 2^{-m\epsilon/2} \\ &= |\Pi_H(S, S')| 2^{-m\epsilon/2} \end{aligned}$$

The last equality comes from the fact that there is one representative for each labeling. So; number of representatives is equal to number of labelings.

In the next lecture; $Pr[B']$ will be written as an expectation and the bound, found above, for $Pr[B'|S, S']$ will be used to bound $Pr[B']$ which will in turn give a bound for $Pr[B]$. Because $Pr[B] \leq 2Pr[B']$.