

ECE 695 Lecture 6 Notes*

Prof. Shai Ben-David
Cornell University

October 8, 2003

1 Introduction

The aim of this lecture is to begin to show how the VC dim characterizes the rate of uniform convergence of $Err^S(h)$ to $Err^P(h)$ for a hypothesis class H . This lecture is focused on negative (lower bound) results pertaining to this problem.

As a reminder, the basic setup is as follows. We consider a domain X , a (typically unknown) distribution P over $X \times \{0, 1\}$, a family of subsets $H \subseteq 2^X$ ¹, and a sample $S \subseteq (X \times \{0, 1\})^m$. There are two relevant measures of error of a hypothesis $h \in H$: the *sample error*, $Err^S(h)$, and the *true error*, $Err^P(h)$, which are defined as follows:

$$Err^S(h) = \frac{1}{m} |\{(x, y) \in S : h(x) \neq y\}|$$

$$Err^P(h) = Pr_P[h(x) \neq y]$$

2 Basic Lower Bounds

The first claim we prove shows that if the size m of the sample S is not large enough, then there will exist a ‘bad’ distribution whereby the *real* and *sample* error differ by some specified amount.

Claim *Let $VCdim(H) = d$ and $|S| = m < d(1 - \epsilon)$. Then $\exists P \exists h \in H : |Err^S(h) - Err^P(h)| > \epsilon$.*

Proof Let $A \subseteq X$ be a set of size d that is shattered by H and pick any $h_0 \in H$. We can define a ‘bad’ distribution P as follows:

$$P(x, y) = \begin{cases} 1/d & x \in A \text{ and } y = h_0(x) \\ 0 & \text{o.w.} \end{cases}$$

Note that since S is drawn i.i.d. according to P that the sample is in fact a subset of $\{(x, y) : x \in A \text{ and } y = h_0(x)\}$.

*Scribed by Philip Zigoris

¹Remember that we can also consider the elements of H as functions mapping X to $\{0, 1\}$, as in $h(x) = 1$ iff $x \in h$.

If $|S| < d$ then there must be some points in A that are not represented in S . Now we choose h such that $\forall x \in A$:

$$h(x) = \begin{cases} h_0(x) & \text{if } (x, h_0(x)) \in S \\ 1 - h_0(x) & \text{o.w.} \end{cases}$$

We know such an h exists because we specifically picked an A that was shattered by H . Note that h is designed such that it correctly labels any point in the domain of S and mislabels all other points.

By definition the sample error is 0. And we can easily see that $Err^P(h) = Pr[(x, y) \in A \times \{0, 1\} : x \notin Domain(S)]$. Since we can draw at most m distinct elements of $Domain(S)$ and by assumption we have $m < d(1 - \epsilon)$ so:

$$\begin{aligned} Err^P(h) &\geq \frac{d - m}{d} \\ &> \frac{d - d(1 - \epsilon)}{d} \\ &= \epsilon \end{aligned}$$

This completes the proof.

In the previous claim our sample size was smaller than d . In most settings, however, m can be considerably larger than d . The following claim allows for a much larger sample while providing similar negative results.

Claim Let $VCdim(H) = d$ and $|S| = m < \frac{d-1}{4\epsilon}$. Then $\exists P$ such that w.p. $> \frac{1}{3}$ over all S , $\exists h \in H : |Err^S(h) - Err^P(h)| > \epsilon$.

Proof Choose some $A \subseteq X$ s.t. $|A| = d$ and A is shattered by H . Again, we want to construct a ‘bad’ P . To do this we pick $x_0 \in A$ and $h_0 \in H$ and define P as follows:

$$P(x, y) = \begin{cases} 1 - 2\epsilon & x = x_0 \text{ and } y = h_0(x) \\ \frac{2\epsilon}{d-1} & x \in A_{x_0} \text{ and } y = h_0(x) \\ 0 & \text{o.w.} \end{cases}$$

Where $A_{x_0} = A \setminus \{x_0\}$. Again, we have a distribution where all examples are labeled according to h_0 , but instead of the distribution being uniform over A , as in the previous proof, we concentrate the distribution on $(x_0, h_0(x_0))$.

Also similar to the last proof we define some h that correctly labels all $x \in S$ but mislabels everything else in A . As above, this h is made available by the fact that H shatters A . So for all $x \in A$, h is defined as:

$$h(x) = \begin{cases} h_0(x) & \text{if } (x, h_0(x)) \in S \\ 1 - h_0(x) & \text{o.w.} \end{cases}$$

It should be clear that for every S the sample error is 0 and the real error is $Pr_P[(x, y) : (x, y) \notin S]$.

Figuring out this probability is a little bit trickier than before. First note that the probability of drawing (x, y) where $x \in A_{x_0}$ is 2ϵ . So $\text{Exp}_S[\#\{(x, y) \in S : x \in A_{x_0}\}] = 2m\epsilon < \frac{d-1}{2}$. In other words, we expect that w.p. $> \frac{1}{3}$ S hits A_{x_0} less than $\frac{d-1}{2}$ times. Consequently, we expect that there are at least $\frac{d-1}{2}$ elements of A not covered by S . Each of those elements has weight $\frac{2\epsilon}{d-1}$ according to P . Thus the probability of drawing (x, y) not in S is greater than ϵ and the proof is complete.

3 Stronger Negative Results

The results of the previous section were based on showing that there exists a distribution and a hypothesis such that the real error exceeded ϵ . But in that setting, the criteria for choosing h was a low sample error. Now we want to consider the situation where you have some learning algorithm L which chooses h according to any criteria. For instance, SVMs optimize a balance between empirical error and distance of the training sample from the hyperplane (called the margin). Formally speaking, L is any function mapping a labeled sample S and to a function h which maps X to $\{0, 1\}$. Note that this function need not be an element of H .

Claim *Let $VCdim(H) = d$ then $\forall L, \exists P$ s.t. with probability $> \frac{1}{10}$, $\text{Err}^P(L(S)) > \epsilon$ whenever $m < \frac{d-1}{16\epsilon}$.*

Proof One thing to notice is that we do not consider the sample error in this claim. In this setting the sample error can be trivially set to zero so it really has so significance.

As in previous proofs, we want to construct an adversarial distribution, but here we must consider a family of distributions. Choose some set $A \subseteq X$ of size d that can be shattered by H and choose some element $x_0 \in A$ and let $A_{x_0} = A \setminus \{x_0\}$. For every $h \in H$:

$$P_h^A(x, y) = \begin{cases} 1 - 8\epsilon & x = x_0 \text{ and } y = h_0(x) \\ \frac{8\epsilon}{d-1} & x \in A_{x_0} \text{ and } y = h_0(x) \\ 0 & \text{o.w.} \end{cases}$$

The claim now is that for any learning algorithm at least one of these distributions is bad.

There are a few observations worth making. First note that the distribution over A_{x_0} is uniform for all h and so the domain of S is independent of h . With this in mind consider the expected error of $L(S)$, for samples of fixed size m , and make the following claim:

$$\forall S \text{Exp}_h[\text{Err}^{P_h^A}] \geq \frac{8\epsilon |A_{x_0} \setminus S|}{2 |A|}$$

To see this is true, take any element of $A_{x_0} \setminus S$. Exactly half of the hypotheses will label this element differently than the hypothesis returned by the learning

algorithm. So for any sample S there must exist a hypothesis h such that the true error of $L(S)$ is greater than or equal to $\frac{4\epsilon}{d-1}|A_{x_0} \setminus S|$.

Now that we have, in some sense, chosen our 'bad' hypothesis h we want to look at the expected value of $|A_{x_0} \setminus S|$. From the previous section we know this value to be $> \frac{d-1}{2}$. So to finish out the proof:

$$\begin{aligned} \text{Exp}_S[\text{Err}^{P_h^A}(L(S))] &\geq \text{Exp}_S\left[\frac{4\epsilon}{d-1}|A_{x_0} \setminus S|\right] \\ &> \frac{4\epsilon}{d-1} \frac{d-1}{2} \\ &= 2\epsilon > \epsilon \end{aligned}$$

As in the previous proof we limited the domain to a set A that can be shattered. Then we defined an adversarial distribution such that the weight of the distribution was focused on an arbitrary element x_0 . In this way we could show that a significant portion of A_{x_0} would not be seen in the sample. And because A is shatterable, the true labeling of these elements could be chosen in a way opposite to the hypothesis returned by L . And in this way we can make sure the true error is large.