CS269: Machine Learning Theory
Lecture 4: Infinite Function Classes
October 6, 2010

Lecturer: Jennifer Wortman Vaughan
Scribe: Luca Valente, Palash Agrawal

1 General Learning Bound for the Unrealizable Setting

In the last lecture, we had established the following theorem.

Theorem 1. *For any concept class H, suppose that we have access to an algorithm that, for any joint distribution D over input-label pairs, given access to m pairs drawn i.i.d. from D, outputs $\hat{h} = \arg\min_{h \in H} \hat{\text{err}}(h)$ (where $\hat{\text{err}}(h)$ is the empirical error of h on the m points). Then for any $\epsilon \in$ (0,1), $\delta \in$ (0,1), with probability $\geq 1 - \delta$,*

$$\text{err}(\hat{h}) \leq \min_{h \in \mathcal{H}} \text{err}(h) + O\left(\sqrt{\frac{ln|\mathcal{H}| + ln\frac{2}{\delta}}{m}}\right)$$

We proved this theorem in the previous lecture using *Hoeffding's Inequality*. We state and prove this inequality now.

2 Hoeffding's Inequality

Before stating *Hoeffding's Inequality*, we recall two intermediate results that we will use in order to prove it. One is *Markov's Inequality* and other is *Hoeffding's Lemma*. (Note that in class we did not cover Hoeffding's Lemma, and only gave a brief outline of the Chernoff Bounding Techniques and how they are used to prove Hoeffding's Inequality. Here we give a full proof of Hoeffding's Inequality for completeness.)

Theorem 2. *(Markov's Inequality)*
Let X be a non negative random variable, for any $K > 0$,

$$Pr[X \geq K] \leq \frac{\boldsymbol{E}[X]}{K}$$

Lemma 1. *(Hoeffding's Lemma)*
Let Z be a random variable so that $Z \in [a, b]$. Then, for any $t \geq 0$,

$$\boldsymbol{E}(e^{tZ}) \leq e^{\frac{t^2(b-a)^2}{8}}$$

The proof of this lemma will not be stated here, but can be found, for example, in the course notes from Peter Bartlett's Statistical Learning Theory course at Berkeley [1].

We now state *Hoeffding's Inequality*.

[1] http://www.cs.berkeley.edu/∼bartlett/courses/281bsp08/13.pdf

Theorem 3. *(Hoeffding Inequality)*
Let $Z_1, Z_2, ..., Z_m$ be independent random variables with $Z_i \in [a_i, b_i]\ \forall\ i$. Define

$$\hat{p} = \frac{1}{m}\sum_{i}^{m} Z_i \text{ and } p = \boldsymbol{E}[\hat{p}] = \frac{1}{m}\sum_{i}^{m} E[Z_i],$$

Then,

$$Pr(\mid p - \hat{p} \mid \geq \epsilon) \leq 2e^{\frac{-2\epsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}}$$

We prove this by using the Chernoff Bounding Techniques and Hoeffding's Lemma.
Proof: For any $t > 0$, e^{tx} is non-negative and monotone increasing with respect to x. This very simple observation allows us to improve on the Markov's Inequality.
For any random variable X(that doesn't have to be non-negative), Markov's Inequality leads to

$$\Pr(X \geq \epsilon) = \mathbf{P}(e^{tX} \geq e^{t\epsilon}) \leq \frac{\mathbf{E}(e^{tX})}{e^{t\epsilon}}, \text{ for any } t > 0$$

Substituting $X = p - \hat{p}$,

$$\begin{aligned} \Pr(p - \hat{p} \geq \epsilon) &\leq e^{-t\epsilon}\mathbf{E}(e^{t(p-\hat{p})}) \\ \Pr(p - \hat{p} \geq \epsilon) &\leq e^{-t\epsilon}\mathbf{E}(e^{t(\frac{1}{m}\sum_{i=1}(Z_i - \mathbf{E}Z_i))}) \end{aligned}$$

Since $Z_1, Z_2, ..., Z_m$ are independent, the random variables $e^{t(\frac{1}{m}\sum_{i=1}(Z_i - \mathbf{E}Z_i)}$ for each i are also independent. Then,

$$\begin{aligned} \Pr(p - \hat{p} \geq \epsilon) &\leq e^{-t\epsilon}\mathbf{E}(\prod_{i=1} e^{t(\frac{1}{m}(Z_i - \mathbf{E}Z_i))}) \\ \Pr(p - \hat{p} \geq \epsilon) &\leq e^{-t\epsilon}\prod_{i=1}\mathbf{E}(e^{t(\frac{1}{m}(Z_i - \mathbf{E}Z_i))}) \end{aligned}$$

Applying Hoeffding's Lemma to $Z_i - \mathbf{E}Z_i$ for any i, we obtain

$$\Pr(p - \hat{p} \geq \epsilon) \leq e^{-t\epsilon}\prod_{i=1} e^{\frac{t^2(b_i - a_i)^2}{8m^2}} = e^{-t\epsilon}e^{\frac{t^2\sum_{i=1}(b_i - a_i)^2}{8m^2}}$$

Since this inequality is true for any $t > 0$, it is also true for the t that minimizes the bound, which is $t = \frac{4\epsilon m^2}{\sum_{i=1}(b_i - a_i)^2}$. This value for t allows us to get the best possible bound. We then obtain

$$\Pr(p - \hat{p} \geq \epsilon) \leq e^{-\frac{2\epsilon^2 m^2}{\sum_{i=1}(b_i - a_i)^2}}$$

By a symmetric argument, we also have

$$\Pr(\hat{p} - p \geq \epsilon) \leq e^{-\frac{2\epsilon^2 m^2}{\sum_{i=1}(b_i - a_i)^2}}$$

which, by the union bound, gives

$$\Pr(|\hat{p} - p| \geq \epsilon) \leq 2e^{-\frac{2\epsilon^2 m^2}{\sum_{i=1}(b_i - a_i)^2}}$$

and concludes the proof. □

3 Case $\mathcal{H}$ Infinite

We have proved useful theorems for the case when $\mathcal{H}$ is finite. But there are many cases where we encounter hypothesis classes having infinite number of functions. The general learning bounds we have considered grow to infinity when $|\mathcal{H}|$ grows to infinity. In this section, we discuss cases when $\mathcal{H}$ is infinite and try finding the general learning bounds.

We have already encountered examples of infinite hypothesis case. One was the threshold problem which involved finding an optimal separation in $(0, 1)$. The set of all the hypothesis might be infinite, but it is still simple to describe as only one parameter (the threshold) is enough to characterize each hypothesis.

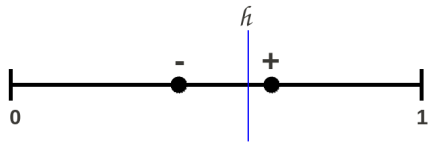


Figure 1: 1-D Threshold

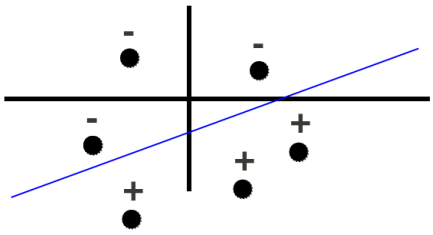


Figure 2: 2-D Threshold

We initially present a 'bad' argument to introduce the domain. *Better arguments* exist (we discuss these later in the notes), but we still discuss the 'bad' argument to make the concept more receptive to intuition. Suppose we have an $\mathcal{H}$ parameterized by d real numbers. We can think of 2-D threshold function where an input vector $\overrightarrow{x} = \{x_1, x_2\}$ can be labeled 1 or 0 by evaluating $w_0 + w_1 x_1 + w_2 x_2 \geq 0$. We can see that $\mathcal{H}$ is infinite in this case and has three parameters for two dimensions. We can generalize this argument and say we have d parameters for $d - 1$ dimensions. Suppose we would like to store a representation of such a function on a computer with finite memory. If each of these parameters is represented by say b bits, we have a total of db bits. This modified hypothesis class $\mathcal{H}'$ would then consist of a finite set of only 2^{db} different

hypothesis, so $|\mathcal{H}'| = 2^{db}$.

If we substitute this value in the bound we obtained for number of samples m that we need to guarantee error less than ϵ with probability $1 - \delta$, [2] we observe that $m = O\left(\frac{1}{\epsilon}(d + \ln(1/\delta))\right)$. We are able to show that the number of training samples needed is at most *linear* in the number of parameters of the class.

The argument we have used may not be completely satisfying, but the conclusion we have arrived at is roughly accurate. If our goal is to minimize training error, then in order to learn a hypothesis class with d parameters *well*, we will frequently (though not always) need order of number of training examples to be linear in d.

3.1 The Growth Function

We now move onto giving more *concrete* arguments for cases where we encounter infinite $\mathcal{H}$. Let's now give some definitions for the general infinite case.

We will assume that we are working in a model of computation in which we can store and manipulate real numbers in constant space and time. This is crucial if we want to say anything about efficient algorithms in this setting.

Let S be a vector of m examples $x_1, \ldots, x_m$. $x_1, \ldots, x_m$ are just m arbitrarily chosen examples that don't have anything to do with the target distribution D. Given $h \in \mathcal{H}$ we define $h(S) = (h(x_1), \ldots, h(x_m))$. There might be another $h' \in \mathcal{H}$ such that $h(S) = h'(S)$. A way to measure how complex a problem is to consider the different behavior of $h(x)$ when $h \in \mathcal{H}$.

Definition 1. *(Behavior Set)*

$$\Pi_{\mathcal{H}}(S) = \{h(S) | h \in \mathcal{H}\}$$

In the case where $h(x_i) \in \{0, 1\}$ for any x_i, we have $|\Pi_{\mathcal{H}}(S)| \leq 2^m$.

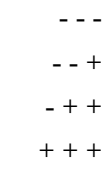
Definition 2. *(Growth Function)*

$$\Pi_{\mathcal{H}}(m) = max_{\{S:|S|=m\}} |\Pi_{\mathcal{H}}(S)|$$

We observe that $|\Pi_{\mathcal{H}}(m)| \leq 2^m$

Let's look at the growth function in the cases discussed before.

Threshold problem: $\mathcal{H}$ is the class of one-dimensional threshold functions.
Given 3 distinct points in $(0, 1)$, there are 4 different possibilities for $h(S)$, so that $\Pi_{\mathcal{H}}(3) = 4$.



[2] Refer to *Theorem 5* of *Lecture 2* notes.

If m points are given, $|\Pi_{\mathcal{H}}(m)| = m + 1$, which is far less that the general bound 2^m.

Suppose we are now given a $\mathcal{H}$ which is a class of interval functions. Each function in the class is parameterized by two threshold values, a lower threshold (call this l) and an upper threshold (call this u). A point x is labeled positive if $x \in [l, u]$ and labeled negative otherwise.

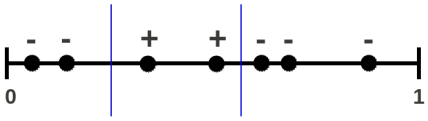


Figure 3: Intervals

If we are given m distinct points in $(0, 1)$. How many different behaviors can we observe? To answer this question, it doesn't matter where exactly the interval boundaries lie, what matters is which pairs of points they lie between. Our m points define $m + 1$ regions in $(0, 1)$. Then the number of different behaviors equals the number of ways we can choose these regions (which is $\binom{m+1}{2}$) plus 1 (which is the case obtained if the two boundaries are in the same region and so all the points are labeled negative), which is $O(m^2)$ and is again far less than the most pessimist bound.

We now show that learning depends really on the number of behaviors. Note that we are back to considering the realizable (perfect target) setting for now.

Theorem 4. *For any $\mathcal{C}, \mathcal{H}$, let A be an algorithm such that from any $c \in \mathcal{C}$ and any distribution D, for any $\delta \in (0, 1)$, if A is given m samples drawn i.i.d. from D, labeled by c, A returns a consistent hypothesis of $\mathcal{H}$. Then, with probability more than $1 - \delta$, $\mathrm{err}(h) \leq O(\frac{ln|\Pi_{\mathcal{H}}(m)| + ln\frac{1}{\delta}}{m})$*

The bound in *Theorem 4* is nice since it involves the growth function which is, as we saw in some examples, sometimes really less than the the worst case 2^m. Nevertheless, this bound is meaningless if the growth function is 2^m. We are particularly interested when value is something smaller and it tightly bounds the $\mathrm{err}(h)$. But the growth function is still hard to calculate in general and therefore we would like to find another complexity measure that we can use in its place that is easier to calculate. This leads us to introduce the *VC-dimension*.

4 Vapnik-Chervonenkis(VC) Dimension

Definition 3. *(Shattered points)*
We say that $S = (x_1, \ldots, x_m)$ of size m is shattered by class $\mathcal{H}$ if $|\Pi_{\mathcal{H}}(m)| = 2^m$, ie if all states of $\{0, 1\}^m$ can be achieved by some $h \in \mathcal{H}$.

Definition 4. *(VC Dimension)*
The VC dimension is the cardinality of the largest set S that can be shattered by $\mathcal{H}$.

The VC dimension is a quantity that we can easily calculate for most of the classes $\mathcal{H}$ that we use.

Example 1: Linear Threshold (1-D)

It is easy to see that 1 point can be shattered (trivial), but 2 cannot. Indeed, it is not possible to achieve the + - configuration. Thus, $VC(\mathcal{H}) = 1$.

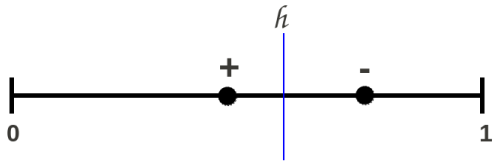


Figure 4: This configuration is not possible

Example 2: Intervals
Similarly, We observe that 2 points can be shattered, but 3 cannot. The + - + configuration is not possible. Thus, $VC(\mathcal{H}) = 2$.

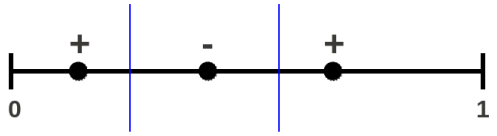


Figure 5: This configuration is not possible

Example 3: Linear Threshold (2-D)
We can observe that any 3 points that do not lie in a line can be shattered.

If the 3 points lie on the same line, one can easily see that they cannot be shattered.

This fact, however, has no influence on the *VC-dimension* since we have already proved that there is at least one set of 3 points that can be shattered. Instead, no set of four points can be shattered. We can break this down into two cases: If one of the 4 points lies in the convex hull of the other three, that point cannot get a different label than the rest.

If no point lies in the convex hull of the other three, then draw a square with the points as the four corners. Pick one pair of points that are diagonally across from each other. It is impossible to label these two points + and the other two -.

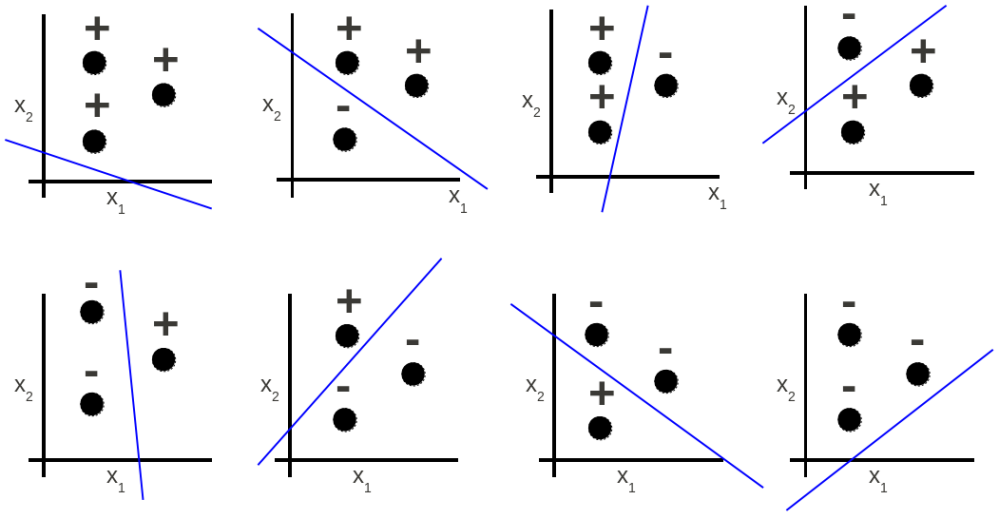


Figure 6: All the configurations possible

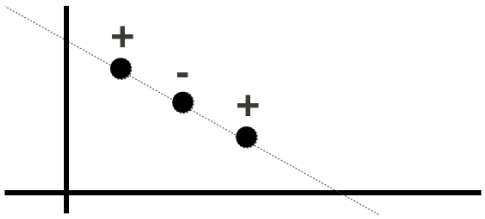


Figure 7: This configuration is not possible

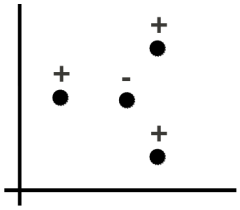


Figure 8: This configuration is not possible

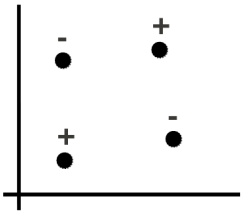


Figure 9: This configuration is not possible

We now give a classical but still a remarkable result.

Lemma 2. *If $\mathcal{H} = \{$linear thresholds in a d-dimensional space$\}$, then $VC(\mathcal{H}) = d + 1$.*

We now connect the VC-dimension with the Growth function.

5 Sauer's Lemma

Lemma 3. *(Sauer's Lemma) For any $\mathcal{H}$ with finite VC-dimension d,*

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} = O(m^d)$$

This lemma is very powerful. It tells us that all hypothesis classes fall into one of two categories: If d is infinite, then $\Pi_{\mathcal{H}}(m) = 2^m$. The bound in *Theorem 4* is then meaningless. On the other hand, if d is finite, then $\Pi_{\mathcal{H}}(m) = O(m^d)$. In this case, *Theorem 4* gives us something very nice since $\log |\Pi_{\mathcal{H}}(m)| = O(d \log m)$. Then the bound is linear in d and decreases to 0 as m goes to infinity.