COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Georgy Noarov

Lecture #6
February 20, 2019

1 Recap from last time

In this lecture, we are going to finish the proof of the theorem that we began last time.

Theorem 1. *With probability at least $1-\delta$, for every $h \in \mathcal{H}$ if h is consistent with m random training examples then $err(h) \leq \epsilon$, where $\epsilon = O\left(\frac{\ln(\Pi_{\mathcal{H}}(2m))+\ln(1/\delta)}{m}\right)$.*

Recall how we went about proving this theorem. We first applied the double sample trick, by considering, along with the original sample S, a second, "ghost" sample S' which also contains m elements. Thus, together S and S' consist of $2m$ iid random variables. Then, these two samples were randomly permuted into samples T and T', correspondingly.

We then defined $M(h, S)$ as the number of mistakes that h makes on S and considered three events:

$$B : \exists h \in \mathcal{H} : M(h,S) = 0 \wedge err(h) > \epsilon$$
$$B' : \exists h \in \mathcal{H} : M(h,S) = 0 \wedge M(h,S') \geq \frac{m\epsilon}{2}$$
$$B'' : \exists h \in \mathcal{H} : M(h,T) = 0 \wedge M(h,T') \geq \frac{m\epsilon}{2}$$

Additionally, the event that $M(h,T) = 0 \wedge M(h,T') \geq \frac{m\epsilon}{2}$ for a particular hypothesis h was denoted $b(h)$.

The driving idea behind the proof of Theorem 1 is that all we need to do in order to see how a consistent (on S) hypothesis h generalizes, is to look at combinatorial properties of the entire hypothesis space $\mathcal{H}$. Thus, we defined the *growth function* of a hypothesis space as

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

where

$$\Pi_{\mathcal{H}}(S) := \{\langle h(x_1), \ldots, h(x_m)\rangle : h \in \mathcal{H}\}.$$

To apply this measure in the context of our proof, we needed to restrict our attention to how h behaves on a finite set of points, namely, the $2m$ iid random variables from the two samples S, S'.

Then the proof of Theorem 1 proceeded in steps. We covered the following ones last time:

- Step 2: $Pr[B] \leq 2Pr[B']$
- Step 3: $Pr[B'] = Pr[B'']$
- Step 4: Fix h, S, S'. Then, $Pr[b(h)|S, S'] \leq 2^{-m\epsilon/2}$.

2 Finishing the proof of Theorem 1

Now, we finish proving the theorem.

Step 5: Fix S, S'. Then $Pr[B''|S, S'] \leq \Pi_{\mathcal{H}}(2m) \cdot 2^{-m\epsilon/2}$.

We have set everything up so that in order to study the error of a consistent hypothesis h, we are only working with the samples S, S' rather than the entire underlying space. At this point, recall that the hypotheses from the space $\mathcal{H}$ exhibit $|\Pi_{\mathcal{H}}(S; S')|$ possible behaviours on combined samples S and S'. Imagine picking $|\Pi_{\mathcal{H}}(S; S')|$ hypotheses out of $\mathcal{H}$, one per each possible behaviour on the samples S and S' combined. Thus, let

$$\mathcal{H}' := \{\text{one representative hypothesis from } \mathcal{H} \text{ for every labelling on } S, S'\}.$$

Then

$$|\mathcal{H}'| = \left|\Pi_{\mathcal{H}}(S; S')\right| \leq \Pi_{\mathcal{H}}(2m).$$

Observe that $b(h)$ only depends on the behaviour of h on S and S', and does not depend on how h labels the rest of the space. This justifies the second equality in the following chain:

$$\begin{aligned} Pr[B''|S, S'] &= Pr[\exists h \in \mathcal{H} : b(h)|S, S'] \\ &= Pr[\exists h \in \mathcal{H}' : b(h)|S, S'] \\ &\leq \sum_{h \in \mathcal{H}'} Pr[b(h)|S, S'] \\ &\leq |\mathcal{H}'| 2^{-m\epsilon/2} \\ &\leq \Pi_{\mathcal{H}}(2m) \cdot 2^{-m\epsilon/2}. \end{aligned}$$

The first inequality is due to the union bound. This concludes the proof of Step 5.

Step 6: $Pr[B''] \leq \Pi_{\mathcal{H}}(2m) \cdot 2^{-m\epsilon/2}$.

The proof here relies on marginalization. By this, we mean in the present context the following special case of the Tower Property of expectations. Let a be an event and X a random variable. Then

$$Pr[a] = \mathbb{E}_X[Pr[a|X]].$$

Applying this property, and then using Step 5, we see that

$$Pr[B''] = \mathbb{E}_{S,S'}[Pr[B''|S, S']] \leq \Pi_{\mathcal{H}}(2m) \cdot 2^{-m\epsilon/2}.$$

Bringing Steps 1-6 together: We observe that

$$Pr[B] \leq 2Pr[B'] = 2Pr[B''] \leq 2\Pi_{\mathcal{H}}(2m) \cdot 2^{-m\epsilon/2}.$$

Here, the transitions hold due to Step 2, Step 3, and Step 6, respectively.

Picking

$$\epsilon = \frac{2}{m}\left(\lg \Pi_{\mathcal{H}}(2m) + \lg 1/\delta + 1\right) = O\left(\frac{\ln(\Pi_{\mathcal{H}}(2m)) + \ln(1/\delta)}{m}\right),$$

we have, as a consequence of $Pr[B] \leq 2\Pi_{\mathcal{H}}(2m) \cdot 2^{-m\epsilon/2}$, that $Pr[B] \leq \delta$, which concludes the proof of the theorem.

3 The Vapnik-Chervonenkis (VC) dimension

The bound on the true error of a consistent hypothesis given in Theorem 1 reflects the complexity of the hypothesis space $\mathcal{H}$ in terms of the logarithm of its growth function evaluated at $2m$, namely, $\ln(\Pi_{\mathcal{H}}(2m))$. However, due to the purely combinatorial nature of the growth function $\Pi_{\mathcal{H}}(\cdot)$, its asymptotic behaviour is far from obvious. Once we can get hold of it, the bound in Theorem 1 will become much more useful. A very important concept that captures the intrinsic complexity of a hypothesis space is the so-called VC-dimension. It will help us study the behaviour of the growth function. We begin our discussion of the VC-dimension with an auxiliary definition.

Definition 2. *A sample S of size m is* shattered *by $\mathcal{H}$ if the hypotheses from $\mathcal{H}$ realize all possible labellings of S. That is, $|\Pi_{\mathcal{H}}(S)| = 2^m$.*

For example, if we consider $\mathcal{H} = \{\text{intervals on the real line}\}$, then clearly $\mathcal{H}$ shatters any subset of $\mathbb{R}$ of size 1 or 2. Indeed, $\mathcal{H}$ shatters any single point since depending on whether that point is labelled $+$ or $-$, we can find an interval that, respectively, includes or does not include that point. And $\mathcal{H}$ shatters any two points. Indeed, suppose the points are x_1 and x_2. Then the labelling $-,-$ is realized by any interval that excludes both points, the labelling $+,+$ is realized by any interval containing both points, and the labelling $+,-$ ($-,+$ respectively) corresponds to any interval that only contains x_1 (x_2 respectively).

Definition 3. *The* Vapnik-Chervonenkis (VC) dimension *$VCdim(\mathcal{H})$ of a hypothesis set $\mathcal{H}$ is the cardinality of the largest set shattered by $\mathcal{H}$.*

Continuing the previous example of $\mathcal{H} = \{\text{intervals on the real line}\}$, we see that $\mathcal{H}$ has VC-dimension 2. Indeed, we have shown that $\mathcal{H}$ shatters sets of 2 points. Thus, to prove that the VC-dimension of $\mathcal{H}$ is equal to 2, it suffices to show that $\mathcal{H}$ does not shatter any set of 3 points (why?). Consider any points $x_1 < x_2 < x_3$ labelled as follows: $+\ -\ +$. Then any interval that contains both x_1 and x_3 must contain x_2. Therefore, it is impossible to find an interval that labels x_1 and x_3 as $+$ and x_2 as $-$. Therefore, we have shown for an arbitrary set of 3 points that not all labellings are realizable by hypotheses from $\mathcal{H}$. This proves our claim.

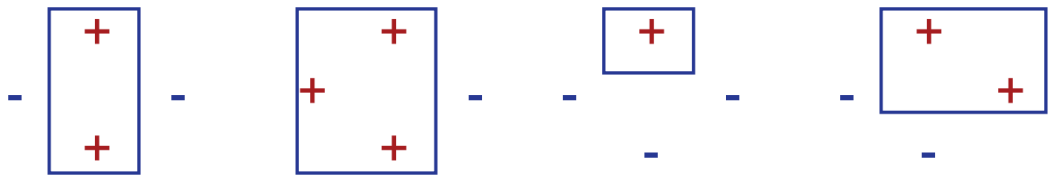

Figure 1: [Mohri et al. textbook, 2nd edition, 2018, pg. 39] A diamond-shaped set in $\mathbb{R}^2$ and some of its possible labellings, along with the axis-aligned rectangles that realize them.

To work out yet another example, consider $\mathcal{H} = \{2d \text{ axis-aligned rectangles}\}$. We claim that $VCdim(\mathcal{H}) = 4$.

- First, $VCdim(\mathcal{H}) \geq 4$ because it can be easily seen that e.g. for a "diamond-shaped" set of 4 points in $\mathbb{R}^2$ (see Figure 1), there is an axis-aligned rectangle for every possible labelling of these points.

- Second, to show that $VCdim(\mathcal{H}) \leq 4$ we need to prove that for any arrangement of 5 points in $\mathbb{R}^2$, there is a labelling that no axis-aligned rectangle can realize. To sketch the proof of this, consider any 5 points $x_1, \ldots, x_5$ in $\mathbb{R}^2$. Choose a top-most, a bottom-most, a left-most, and a right-most point from among these points. Some of the chosen points may be equal, e.g. if one of the points is the top-most and also the left-most point, but in total we will have chosen at most 4 points. Since we have 5 points in total, there will be at least one "leftover" point. Now, label the chosen points as $+$ and the leftover points as $-$. No axis-aligned rectangle can both contain the chosen points and exclude the leftover points. Hence, no set of 5 points is shattered by $\mathcal{H}$, proving our claim.

In general, it can be shown that $VCdim(\{\text{hyper-rectangles in } \mathbb{R}^n\}) = 2n$. See the textbook for more details.

Finally, there is a very important hypothesis set whose VC-dimension we want to describe.

Definition 4. *A* linear threshold function $h(\cdot)$ *in* $\mathbb{R}^n$ *with parameters* $\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}$ *is defined as* $h(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{w} \cdot \mathbf{x} \geq b \\ 0, & \text{otherwise} \end{cases}$.

Any hyperplane of the form $\mathbf{w} \cdot \mathbf{x} = b$ gives rise to a split of the space $\mathbb{R}^n$ into two halves, and points are classified as $+$ or $-$ depending on which of the two half-spaces they belong to.

Lemma 5. *The hypothesis set* $\mathcal{H} = \{$*linear threshold functions in* $\mathbb{R}^n\}$ *has* $VCdim(\mathcal{H}) = n + 1$*. Furthermore, the VC-dimension of the set of linear threshold functions in* $\mathbb{R}^n$ *that are based on hyperplanes that go through the origin, that is, where* $b = 0$*, is equal to* n*.*

For the proof of this lemma, consult the textbook.

Remark 1. As seen from the previous examples, it is often the case that the VC-dimension of a hypothesis set $\mathcal{H}$ is the same as the number of parameters that are needed to define $\mathcal{H}$. In particular, intervals in $\mathbb{R}$ are given by their two ends, and the VC-dimension is 2. Axis-aligned rectangles are given by their 4 corners, and the VC-dimension is 4. Linear threshold functions in $\mathbb{R}^n$ have $n+1$ parameters (n coordinates of $\mathbf{w}$ and the one-dimensional b), and this is the same as their VC-dimension. Moreover, setting $b = 0$ reduces the effective number of parameters by 1, and accordingly the VC-dimension reduces to n. However, this is not always the case. For instance, there exist single-parameter environments whose VC-dimension is ∞.

Remark 2. An interesting question is that of the relation between $VCdim(\mathcal{H})$ and $\lg|\mathcal{H}|$ for finite hypothesis spaces. We would like to note the property that $VCdim(\mathcal{H}) \leq \lg|\mathcal{H}|$ in that case. Indeed, $VCdim(\mathcal{H}) = d$ holds only if there is a set of d examples that is shattered by $\mathcal{H}$. This implies, however, that there are at least 2^d hypotheses in $\mathcal{H}$, at least one per labelling of the d examples. So $|\mathcal{H}| \geq 2^d$, and taking logs gives the above inequality.

4 Sauer's lemma

We begin the proof of the following result.

Theorem 6 (Sauer's lemma)**.** *Let $\mathcal{H}$ be a hypothesis space, $d = VCdim(\mathcal{H})$. Then*

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

This theorem has a very surprising and far-reaching consequence:

Corollary 7. *One has the following surprising dichotomy:*

- *If $d < \infty$, $\Pi_{\mathcal{H}}(m) = O(m^d)$ for all $m \in \mathbb{N}$;*
- *If $d = \infty$, $\Pi_{\mathcal{H}}(m) = 2^m$ for all $m \in \mathbb{N}$.*

Rather than immediately proving the corollary, note its interpretation. For any $\mathcal{H}$, there are essentially only two possibilities. One of them is that the VC-dimension of $\mathcal{H}$ is finite. This case is "very good": the growth function of $\mathcal{H}$ is polynomial in m, and the bound on the error of consistent hypotheses from Theorem 1 becomes $\epsilon = O\left(\frac{d \ln m + \ln(1/\delta)}{m}\right)$. The other possibility is that the VC-dimension of $\mathcal{H}$ is infinite. It is the "worst possible case" in terms of bounding the error of consistent hypotheses: the growth function of $\mathcal{H}$ becomes exponential in m, and so for fixed δ the bound in Theorem 1 does not go to zero as $m \to \infty$. It is also the worst case in the sense that 2^m is the maximum number of labellings of any m points.

Now we turn to the proof of Sauer's lemma. Let us note the following helpful properties of binomial coefficients that we use.

1. $\binom{m}{k} = \frac{m(m-1)\ldots(m-k+1)}{k!}$. From this property, it follows that $\binom{m}{k} = O(m^k)$ for fixed k. In particular, this demonstrates Corollary 7, since for $d < \infty$, the bound in Sauer's lemma becomes $O(m^d)$.
2. $\binom{m}{k} = 0$ if $k < 0$ or $k > m$.
3. $\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}$. This is often referred to as Pascal's triangle property.

We also introduce a useful notation: $\Phi_d(m) := \sum_{i=0}^{d} \binom{m}{i}$.

Proof of Sauer's lemma. We will prove Sauer's lemma by induction on $m + d$. Let us start with the base cases.

- $m = 0$. Here, $\Pi_{\mathcal{H}}(m) = 1 = \sum_{i=0}^{d} \binom{0}{i} = \Phi_d(0)$. This is because there is exactly 1 way to label m points, namely the empty sequence of labels.
- $d = 0$. Here, $\Pi_{\mathcal{H}}(m) = 1 = \sum_{i=0}^{d} \binom{m}{0} = \Phi_0(m)$.

Now we assume $m \geq 1, d \geq 1$, and that the inductive hypothesis holds for any m', d' such that $m' + d' < m + d$. We consider an arbitrary sample $S = \langle x_1, \ldots, x_m \rangle$. For the inductive step, we want to show that $|\Pi_{\mathcal{H}}(S)| \leq \Phi_d(m)$.

Now, we look at all labellings induced by $\mathcal{H}$ on S. Let us define $\mathcal{H}_1$ to be the set of labellings induced by $\mathcal{H}$ on $S' = \langle x_1, \ldots, x_{m-1} \rangle$. We can equivalently think of members of $\mathcal{H}_1$ as hypotheses defined on a restricted domain which is S'. Also, we define $\mathcal{H}_2$ to be the set of all labellings l of S' such that there are exactly two different labellings of S by $\mathcal{H}$ that induce the labelling l on S'. We call this a "collapse". In other words, labellings l of S' that belong to $\mathcal{H}_2$ have the following property: The possible labellings of S such that

						$\mathcal{H}_1$			
x_1	x_2	x_3	x_4	x_5		x_1	x_2	x_3	x_4
0	1	1	0	0	$\rightarrow$	0	1	1	0
0	1	1	0	1	$\nearrow$				
0	1	1	1	0	$\rightarrow$	0	1	1	1
1	0	0	1	0	$\rightarrow$	1	0	0	1
1	0	0	1	1	$\nearrow$				
1	1	0	0	1	$\rightarrow$	1	1	0	0

Figure 2: Example illustrating the relationship between $\Pi_{\mathcal{H}}(S)$ and $\mathcal{H}_1$. Here $S = (x_1, \ldots, x_5)$ and $S' = (x_1, \ldots, x_4)$. Note the collapse in rows $1-2$ and $4-5$. Because of that, $\mathcal{H}_2 = \{(0\,1\,1\,0), (1\,0\,0\,1)\}$ since these are the labellings on $x_1, \ldots, x_4$ that correspond to the collapses.

$x_1, \ldots, x_{m-1}$ are labelled according to l (the one where x_m is labelled as a 0 and the one where x_m is labelled as a 1) are both realized by $\mathcal{H}$ on S. See Figure 2 for an illustrative example.

Claim. $|\mathcal{H}_1|+|\mathcal{H}_2| = |\Pi_{\mathcal{H}}(S)|$. This follows from the definitions of $\mathcal{H}_1, \mathcal{H}_2$, since counting all possible labellings of S ($|\Pi_{\mathcal{H}}(S)|$) is equivalent to counting all labellings of S' ($|\mathcal{H}_1|$) and then adding 1 whenever there are two labellings of S that correspond to a particular labelling of S', i.e. whenever there is a "collapse" ($|\mathcal{H}_2|$).

The idea of what follows in the proof of Sauer's lemma is to look at $\mathcal{H}_1, \mathcal{H}_2$ and their VC-dimensions, connect those to $VCdim(\mathcal{H}) = d$, and apply the inductive hypothesis. Here is the first step in that direction:

Claim. $VCdim(\mathcal{H}_1) \leq d$. Indeed, any set of examples T that is shattered by $\mathcal{H}_1$ is also shattered by $\mathcal{H}$, since $\mathcal{H}$ includes all the same labellings of S' as $\mathcal{H}_1$. $\square$