CS 2416 - Machine Learning Theory

Lecture #6: 20 October 2005

Lecturer: Toniann Pitassi

Scribe Notes by: Toni Pitassi

VC Dimension

Recall from last class the following definitions.

Definition Let S be a set of points/examples. Define $\mathcal{C}[S]$ as the number of ways to split S using a concept from the concept class $\mathcal{C}$. Define $\mathcal{C}[m]$ as the maximum over all sets S, $|S| = m$, of $\mathcal{C}[S]$. That is, $\mathcal{C}[m]$ is the maximum number of ways to split m points (examples) using a concept from the concept class $\mathcal{C}$. (The textbook calls this $\pi_{\mathcal{C}}(m)$.) Clearly $\mathcal{C}[m] \leq 2^m$.

Definition A set of points S is *shattered* by $\mathcal{C}$ if there are concepts in $\mathcal{C}$ that split S in all of the $2^{|S|}$ possible ways. (That is to say that if you can split every possible combination of input with some concepts from C, then you have shattered S.)

Definition The VC Dimension of $\mathcal{C}$ is the size of the **largest** set of points that can be shattered by $\mathcal{C}$. This is denoted by $VCdim(\mathcal{C})$. $VCdim(\mathcal{C}) = d$ means that there exists a set S, $|S| = d$, that is shattered by $\mathcal{C}$. It also means that there is no set of size $d+1$ can be shattered.

Properties of VC Dimension

Theorem 1 *(This is similar to Theorem 3.1 and 3.2 in the book.)*

$$C[m] \leq \sum_{i=0}^{VCdim(\mathcal{C})} \binom{m}{i} = O(m^{VCdim(\mathcal{C})})$$

Recall $\mathcal{C}[m]$ is the total number of different values realizable by the concepts in $\mathcal{C}$. This means that if the VC dimension is infinite, then m can be split in all possible ways ($\mathcal{C}[m]$ is unbounded) If the VC dimension is finite, then $\mathcal{C}[m]$ eventually stops going quickly and converges to m^d. ($\mathcal{C}[m]$ is polynomially bounded in d).

Proof Let $(\binom{m}{d})$ be the number of ways of choosing d or fewer items out of m. Therefore, $(\binom{m}{d})$ $= \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{d}$.

Fact $(\binom{m}{d}) = \left(\binom{m-1}{d}\right) + \left(\binom{m-1}{d-1}\right)$

The fact can be seen by noting that the the number of ways to choose d or fewer elements out of the set $a_1, \ldots, a_m$ is to either pick a_1 and then choose $d-1$ or fewer elements out of $a_2, \ldots, a_m$, or to not pick a_1, then then choose d or fewer out of $a_2, \ldots, a_m$.

The proof of the theorem is by induction on both d and m. When either $d = 0$ or $m = 0$ then the statement holds. Now assume by induction that the statement holds for all m' and d' such that $m' \leq m$ and $d' \leq d$ and at least one of the two inequalities is strict. We now show that this inductive assumption establishes the desired statement for d and m. Let S be any set of points of size m. $\mathcal{C}[S]$ is, by definition, the number of splittings of S by concepts in $\mathcal{C}$, and we define $Set_{\mathcal{C}}(S)$ to be this set of splittings. Pick an $x \in S$. Let $d = VCdim(C)$. We know by using induction that $\mathcal{C}[S - \{x\}]$ has at most $\left(\binom{m-1}{d}\right)$ different partitions. How many more splittings are there of S than there are of $S - \{x\}$?

The answer is that the increase is exactly the number of pairs in $Set_{\mathcal{C}}(S)$ that differ only on x, since these collapse when you remove x. For each such pair, focus on the one that labels x negative. Let $\mathcal{C}' = \{c$ in $Set_{\mathcal{C}}(S)$ such that it labels x negative, for which there is another concept in $Set_{\mathcal{C}}(S)$ that is identical to c except it labels x positive.$\}$ That is, $|\mathcal{C}'|$ counts the number of pairs of sets from $Set_{\mathcal{C}}(S)$ that are collapsed to a single representative in $Set_{\mathcal{C}}(S - \{x\})$. We want to show that $VCdim(\mathcal{C}') \leq d - 1$ to finish the proof.

Proof by contradiction: if $VCdim(\mathcal{C}') = d$, then there is a set S' of d points in S that can be shattered, and further, x is not in S'. But then $\{S' \cup x\}$ is a set of size $d+1$ that is shattered by $\mathcal{C}$, contradicting $VCdim(\mathcal{C}) = d$.

Thus our total count is $\left(\binom{m-1}{d}\right) + \left(\binom{m-1}{d-1}\right)$.

Theorem 2 *For any concept class $\mathcal{C}$ and input distribution $\mathcal{D}$, if*

$$m > \frac{2}{\varepsilon}[\log_2(2C[2m]) + \log_2(\frac{1}{\delta})]$$

where m is the number of samples, then with probability $(1-\delta)$ all bad (error greater than ε) hypotheses in $\mathcal{C}$ are inconsistent with data.

This is like the Occam theorem from before, which says "if you use at least this many samples, you should be OK"

Proof Fix the target concept $c \in \mathcal{C}$ and fix a distribution $\mathcal{D}$. For each concept $c' \in \mathcal{C}$, the error region, $error_c(c')$ is defined to be equal to the set of inputs where c' and c disagree. It is not hard to see that the VC dimension of $\mathcal{C}$ is equal to the VC dimension of the set of error regions, $ERR_c = \{error_c(c') \mid c' \in \mathcal{C}\}$. (To see this, we want to show that if a set S is shattered over $Set_{\mathcal{C}}(S)$, then S is still shattered over ERR_c. Notice that we go from a vector v from $Set_{\mathcal{C}}(S)$ to a vector v' in ERR_c by taking the bitwise parity of v and c. Since this map is one-to-one, the size of $Set_{\mathcal{C}}(S)$ will be equal to the size of Err_c, for any c.)

Now we can further refine the set of error regions to consider only those error regions with weight at least ϵ with respect to the distribution $\mathcal{D}$. Let $ERR_{\epsilon,c}$ denote this subset of ERR_c. Since $ERR_{\epsilon,c}$ is a subset of ERR_c, it follows that the VC dimension of $ERR_{\epsilon,c}$ is at most the VC dimension of $\mathcal{C}$.

Definition For any $\epsilon > 0$, we say that a set S is an ϵ-net for ERR_c if every region in $ERR_{\epsilon,c}$ is hit by a point in S. (Here S is a subset of the set of all examples/points.)

Our goal is therefore to show that if we choose a set S of examples of size at least m uniformly at random from distribution $\mathcal{D}$ (where m is the quantity defined in the theorem), then S forms an ϵ-net for $ERR_{\epsilon,c}$ with probability at least $(1-\delta)$.

Given a set S_1 of m random examples from $\mathcal{D}$, let A denote the event that the elements of S_1 fail to form an ϵ-net for ERR_c. Let B be the combined event that in $2m$ samples, where S_1 is the first m samples, and S_2 is the next m samples, that A occurs on the draw of S_1 (so S_1 is not an ϵ-net), and that S_2 has at least $\epsilon m/2$ hits in a region of $ERR_{\epsilon,c}$ that is missed by S_1. Since the definition of event B already requires that A occurs on S_1, $Pr[B] = Pr[B \wedge A] = Pr[A]Pr[B \mid A]$.

We claim that $Pr[A] \leq 2Pr[B]$. To prove this claim, we can use Markov or Chernoff bounds to argue that $Pr[B \mid A] > 1/2$. (Given that event A occurs, S_1 misses some region $r \in ERR_{\epsilon,c}$. Fix this missed region r, and now draw an additional multiset S_2 of m random examples from $\mathcal{D}$. Since each element of S_2 has probability at least ϵ of hitting r, if $m = O(1/\epsilon)$, the probability that S_2 hits r at least $\epsilon m/2$ times is at least $1/2$.)

By the above claim, we can upper bound $Pr[A]$ by upper bounding $Pr[B]$; it therefore remains to show how to upper bound $Pr[B]$.

As explained in the book, the principle advantage of bounding event B rather than event A is as follows. In order to directly analyze event A, we must consider all regions of the uncountably infinite class $ERR_{\epsilon,c}$ that S_1 might miss. On the other hand, the occurrence of event B is equivalent to saying that there is some region r in $Set_{ERR_{\epsilon,c}}(S_1 \cup S_2)$ such that $|r| \geq \epsilon m/2$ and r does not intersect S_1. Thus, to analyze event B, we only have to consider the number of different regions in $Set_{ERR_{\epsilon,c}}(S_1 \cup S_2)$ and this is upper bounded by $\mathcal{C}[2m]$. We now proceed to the formal analysis.

Let S be a randomly drawn set of size $2m$. We will choose S_1 and S_2 by randomly splitting S into two parts each of equal size. Once S is drawn and fixed, we may also fix a region $r \in Set_{\mathcal{C}}(S)$ such that $|r| \geq \epsilon m/2$. For this fixed S and r, we now analyze the probability (with respect to the random partitioning of S into S_1 and S_2) that r and S_1 are disjoint. We then obtain our overall bound by summing over all possible distinct regions $r \in Set_{\mathcal{C}}(S)$.

We claim that the probability (with respect to the random partitioning of S into S_1 and S_2) that r and S_1 are disjoint is at most $(1/2)^{\epsilon m/2}$.

The above claim holds because we can think of the experiment as fixing the split of S into S_1 and S_2 and then picking the $l = \epsilon m/2$ points randomly. The probability that all $l = \epsilon m/2$ points lie in S_2 is $\binom{m}{l}/\binom{2m}{l} \leq 1/2^l$.

Using the above claim and applying the union bound, we get: $Pr[B] \leq C[2m] \cdot (1/2)^{\epsilon m/2}$.

Thus

$$Pr[A] \leq 2Pr[B] \leq 2C[2m](1/2)^{\epsilon m/2}.$$

We want this to be at most δ since we want the probability of a bad hypothesis (A) to not exceed δ. Solving for m gives us:

$$m = \frac{2}{\varepsilon}\left[\log_2\left(2C[2m]\right) + \log_2\left(\frac{1}{\delta}\right)\right]$$

Last class we proved our main theorem (Theorem 2), giving an upper bound on the number of samples needed in terms of $\mathcal{C}[m]$. The following theorem uses Sauer's Lemma (Theorem 1) to state the theorem directly in terms of VC dimension.

Theorem 3.(Similar to Theorem 3.3 in the text) For any concept class $\mathcal{C}$ and input distribution $\mathcal{D}$, if the number of samples, m, is at least

$$O\left(\frac{1}{\epsilon}[(VCdim(\mathcal{C})\log(1/\epsilon)+\log(1/\delta)]\right)$$

then with probability at least $(1-\delta)$ all bad (error greater than ϵ) hypotheses in $\mathcal{C}$ are inconsistent with the data.

Proof (sketch) By using Theorem 1, and Theorem 2, where d is the VC dimension of $\mathcal{C}$, we know

$$\begin{aligned}
Pr[B] &\leq C[2m]\cdot\left(\frac{1}{2}\right)^{\frac{\varepsilon m}{2}}\\
&\leq \left(\binom{2m}{d}\right)\cdot\left(\frac{1}{2}\right)^{\frac{\varepsilon m}{2}}\\
&\leq \left(\frac{2em}{d}\right)^{d}\cdot\left(\frac{1}{2}\right)^{\frac{\varepsilon m}{2}}
\end{aligned}$$

Set $m=(4d/\epsilon)\cdot(\log(1/\epsilon))+(2/\epsilon)\cdot\log(2/\delta)$.

Plugging in this value for m, we can rewrite $(1/2)^{\epsilon m/2}$ as: $(\epsilon)^{2d}\cdot(\delta/2)$. Now we can rewrite the whole thing as:

$$\begin{aligned}
Pr[B] &\leq \left(\frac{2em}{d}\right)^{d}\cdot(1/2)^{\epsilon m/2}\\
&\leq \left(\frac{2em}{d}\right)^{d}\cdot(\epsilon)^{2d}\cdot(\delta/2)
\end{aligned}$$

Now by observing that $m\leq 2(4d/\epsilon)\cdot(\log(1/\epsilon))$ we obtain:

$$\begin{aligned}
Pr[B] &\leq \left(\frac{2e8d\log(1/\epsilon)}{\epsilon d}\right)^{d}\cdot(\epsilon)^{2d}\cdot(\delta/2)\\
&= \left(\frac{2e8d\log(1/\epsilon)\epsilon^2}{\epsilon d}\right)^{d}\cdot(\delta/2)\\
&= (16e\log(1/\epsilon)\epsilon)^{d}\cdot(\delta/2)\\
&< \delta/2 \text{ for small enough } \epsilon.
\end{aligned}$$

We point out again here that the above theorem is actually a general combinatorial theorem about hitting sets.

Definition Let Q be a set of subsets over a universe U. (In our case, Q is the set of subsets $s_{c'}$ for all $c'\in\mathcal{C}$, and U is the set of all underlying inputs/points.) An ϵ-net for Q is a set $S\subseteq U$ such that for all sets $q\in Q$ such that $|q|\geq\epsilon|U|$, $q\cap S\neq\emptyset$.

Thus our theorem is saying that if we choose a set S of samples randomly, and of sufficiently large size that depends linearly on the VC dimension of Q, then with probability at least $(1-\delta)$, S will form an ϵ-net. That is, S will hit all sufficiently large sets.

The following theorem shows that our bound on the size of S is fairly tight. We will either prove this theorem later, or provide a proof in the lecture notes.

Theorem 4. (Theorem 3.5 in the text) For any learning algorithm A, there exists a distribution $\mathcal{D}$, and a distribution of target concepts $\mathcal{C}$, such that the expected error of A is greater than ϵ if A sees less than $\frac{VCdim(C)}{\delta\varepsilon}$ examples.