CSE 250: Machine Learning Theory Spring 2017

Part 3

Instructor: Raef Bassily Scribe: Andrew Leverentz

3.1 Vapnik-Chervonenkis (VC) Dimension, continued

Definition 3.1 (Set Shattering)**.** A hypothesis class $\mathcal{H}$ shatters a finite set $T_n = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ if the set $C_{\mathcal{H}}(T_n) = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\} = \{0,1\}^n$. (This means that it can generate all 2^n labelings of T_n.)

Definition 3.2 (VC Dimension)**.** The VC dimension of $\mathcal{H}$ is the size of the largest set of domain points that can be shattered by $\mathcal{H}$. The VC dimension is denoted $\mathrm{VC}(\mathcal{H})$ or $K_{\mathcal{H}}$.

The VC dimension is infinite if, for all n, there is a set T_n of size n that can be shattered by $\mathcal{H}$.

Remark 3.3. Previously, we defined

$$\hat{C}_{\mathcal{H}}(n) = \max_{T_n \subset \mathcal{X} : |T_n| = n} |C_{\mathcal{H}}(T_n)|.$$

This is called the "growth function."

For example, recall the class of binary thresholds. We cannot shatter any set of size 2, so VC(thresholds) = 1.

With linear classifiers in $\mathbb{R}^2$, we cannot shatter any set of size 4. (To see why, note that for any 4 points, there must be 1 point that lies outside of the convex hull of the others, and there is an associated labeling that we cannot achieve.) So, VC(linear classifiers) = 3.

Example 3.4 (Intervals)**.** Consider $\mathcal{H}$ = indicator functions for intervals on $\mathbb{R}$. We can shatter a set of 2 distinct points, but we cannot shatter any set of 3 points. Hence, VC(intervals) = 2.

Note that a set of n points divides the real line into $n+1$ regions. If we place an interval such that its endpoints lie between two adjacent data points, we get just one labeling (namely, all 0's). If we place an interval such that its endpoints lie within distinct regions (i.e., it spans at least one data point), we have $\binom{n+1}{2}$ choices. In total, we have $1 + \binom{n+1}{2}$ possible labelings. So,

$$\hat{C}_{\mathcal{H}}(n) = O(n^2).$$

We will see that the earlier definition of VC dimension (in terms of the degree of the polynomial upper bound on the growth function) is equivalent to today's definition (in terms of shattering sets).

What if we have "bi-directional" intervals? That is, we pick an interval and choose whether to assign 0 or 1 to the interior. The VC dimension of this hypothesis class is 3, because we can shatter some sets of size 3, but no sets of size 4.

Example 3.5 (Returning to axis-aligned rectangles)**.** There exist some sets of size 4 (not necessarily all sets of size 4) that we can shatter. This means 4 is a lower bound on the VC dimension.

With 5 points, at least one point must not be "extremal" (i.e., not at the extreme left, right, top, or bottom). We cannot generate the labeling where this point gets assigned "−" and all others get assigned "+". So *no* collection of 5 points can be shattered.

Hence, VC(axis-aligned rectangles) = 4.

What about the VC dimension of finite hypothesis classes?

Theorem 3.6. For any finite hypothesis class $\mathcal{H}$,

$$\mathrm{VC}(\mathcal{H}) \leq \log_2(|\mathcal{H}|).$$

Proof. Let $T \subset \mathcal{X}$ be the largest set that can be shattered by $\mathcal{H}$. Note that for each labeling in $C_{\mathcal{H}}(T)$, there is at least an $h \in \mathcal{H}$ that has generated this labeling. Then

$$|\mathcal{H}| \geq |C_{\mathcal{H}}(T)| = 2^{|T|} = 2^{\mathrm{VC}(\mathcal{H})}.$$

Therefore, $\mathrm{VC}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$. □

Example 3.7 (Thresholds chosen from a finite set)**.** Suppose

$$\mathcal{X} = \{1, \ldots, \ell\},$$
$$\mathcal{H} = \{\text{threshold functions } h_t : t \in \mathcal{X}\}.$$

In this example, $\log_2(|\mathcal{H}|) = \log_2 \ell$, but the VC dimension is 1. This example shows that the VC dimension of a finite hypothesis class can be significantly smalles than log the size of the class. This can lead to improved sample bounds (i.e., better $n_{\mathcal{H}}(\varepsilon, \delta)$).

3.2 Fundamental Theorem of Statistical Learning

Theorem 3.8 (Characterization of learnability via VC dimension)**.** Let $\mathcal{H}$ be any hypothesis class. If $\mathcal{H}$ has a finite VC dimension, then $\mathcal{H}$ is PAC (and agnostic-PAC) learnable.

The converse is also true, but we will not prove it in this course.

Remark 3.9 (Preview of proofs)**.** We will see that

$$P[\mathrm{err}(h_S; \mathcal{D}) - \mathrm{err}(h^*; \mathcal{D}) > \varepsilon] \leq \underbrace{\hat{C}_{\mathcal{H}}(n)}_{O(n^{\mathrm{VC}(\mathcal{H})})} e^{-\Omega(\varepsilon^2 n)} \triangleq \delta.$$

Fact 3.10. Let $\mathcal{H}_1, \mathcal{H}_2$ be two hypothesis classes. Suppose that for every $T \subset \mathcal{X}$ that is shattered by $\mathcal{H}_1$, T is also shattered by $\mathcal{H}_2$. Then $\mathrm{VC}(\mathcal{H}_1) \leq \mathrm{VC}(\mathcal{H}_2)$.

Proof. Let $\mathrm{VC}(\mathcal{H}_2) = K$. For the sake of contradiction, suppose $\mathrm{VC}(\mathcal{H}_1) \geq K + 1$. This means $\mathcal{H}_1$ can shatter a set of size $K + 1$. By the premise, this implies that $\mathcal{H}_2$ can shatter this set, which implies $\mathrm{VC}(\mathcal{H}_2) \geq K + 1$. This is a contradiction. □

Fact 3.11. Let $\mathcal{H}_1, \mathcal{H}_2$ be two hypothesis classes over domain $\mathcal{X}$. Suppose that $\mathrm{VC}(\mathcal{H}_1) < |\mathcal{X}|$. Suppose that for every $W \subset \mathcal{X}$ that is shattered by $\mathcal{H}_1$, there is some element $x \in \mathcal{X} \setminus W$ such that $W \cup \{x\}$ is shattered by $\mathcal{H}_2$. Then $\mathrm{VC}(\mathcal{H}_1) \leq \mathrm{VC}(\mathcal{H}_2) - 1$.

Proof. Let W be the largest set that can be shattered by $\mathcal{H}_1$. That is, $\mathrm{VC}(\mathcal{H}_1) = |W|$. By the premise, there is $x \in \mathcal{X} \setminus W$ s.t. $\tilde{W} = W \cup \{x\}$ is shattered by $\mathcal{H}_2$. Hence, $\mathrm{VC}(\mathcal{H}_2) \geq |W| + 1 = \mathrm{VC}(\mathcal{H}_1) + 1$. □

3.3 Sauer's Lemma

Lemma 3.12 (Sauer's Lemma)**.** If $\mathrm{VC}(\mathcal{H}) = k$, then for any $T \subset \mathcal{X}$ of size n, we have

$$\begin{aligned} |C_{\mathcal{H}}(T)| &\leq \sum_{i=0}^{k} \binom{n}{i} \\ &\leq \left(\frac{en}{k}\right)^k \qquad \text{for } n > k \\ &= O(n^k) \qquad\qquad \text{for } n > k. \end{aligned}$$

Remark 3.13. This implies

$$\hat{C}_{\mathcal{H}}(n) \leq O\left(\left(\frac{en}{k}\right)^k\right).$$

Proof. First we prove the first inequality. The proof is by induction on n and k.

Base case 1: $n = 0$, k is arbitrary. Note: by convention $\binom{x}{y} = 0$ when $y > x$, and $\binom{x}{0} = 1$. So, the left-hand side is 0, the right-hand side is 1, and the inequality is true.

Base case 2: $k = 0$, n is arbitrary. Here, the left-hand side is 1, the right-hand side is $\binom{n}{0} = 1$, and the inequality is true.

Since the induction involves two variables, we can think of it as filling in a "grid" of statements corresponding to integer-valued points (k, n).

Induction assumption: We will assume that the inequality holds for the cases $(k-1, n-1)$ and $(k, n-1)$, and we must show that the inequality holds for (k, n).

Let $\Psi(k, n)$ refer to the right-hand side. Then we want to show that

$$|C_{\mathcal{H}}(T)| \leq \Psi(k, n).$$

Given a set $T = \{x_1, \ldots, x_n\}$, consider the restriction $\mathcal{H}_T$ of $\mathcal{H}$ to T. (We will pretend that T is our "new" domain).

For example, consider a set T of five points:

$\mathcal{H}_T$	x_1	x_2	x_3	x_4	x_5
h_1	0	1	1	0	0
h_2	0	1	1	0	1
h_3	0	1	1	1	0
h_4	1	0	0	1	0
h_5	1	0	0	1	1
h_6	1	1	0	0	1

We can group hypotheses based on how they label all but the last point:

$$\{h_1, h_2\}, \{h_3\}, \{h_4, h_5\}, \{h_6\}.$$

Put representatives from each group into a new hypothesis class $\mathcal{H}_1$, and put any "duplicates" into $\mathcal{H}_2$. Here, $\mathcal{H}_1 = \{h_1, h_3, h_4, h_6\}$ and $\mathcal{H}_2 = \{h_2, h_5\}$. Note that $\mathcal{H}_T = \mathcal{H}_1 \cup \mathcal{H}_2$.

More generally, we partition $\mathcal{H}_T$ into two subsets and also obtain a partition of $\mathcal{H}_T$ into $\mathcal{H}_1$ and $\mathcal{H}_2$. The set $\mathcal{H}_1$ contains hypotheses in $\mathcal{H}_T$ that can generate the maximum number of labelings over $T \setminus \{x_n\}$. Whenever there are two hypotheses in $\mathcal{H}_T$ that generate the same labeling on $T \setminus \{x_n\}$, we put one in $\mathcal{H}_1$ and the other in $\mathcal{H}_2$.

Note that if T is shattered by $\mathcal{H}_1$, then T is also shattered by $\mathcal{H}_T$, because $\mathcal{H}_1 \subseteq \mathcal{H}_T$. This implies

$$\mathrm{VC}(\mathcal{H}_1) \leq \mathrm{VC}(\mathcal{H}_T) \leq \mathrm{VC}(\mathcal{H}).$$

Note also that for every subset $W \subset T \setminus \{x_n\}$ that is shattered by $\mathcal{H}_2$, $W \cup \{x_n\}$ must be shattered by $\mathcal{H}_T$. This is because each labeling of W using $\mathcal{H}_2$ corresponds to two distinct labelings of $W \cup \{x_n\}$ using $\mathcal{H}_T$. Hence, using Fact 3.11, we have

$$\mathrm{VC}(\mathcal{H}_2) \leq \mathrm{VC}(\mathcal{H}_T) - 1 \leq \mathrm{VC}(\mathcal{H}) - 1.$$

Because we've created a partition of $C_{\mathcal{H}}(T)$, we also have

$$|C_{\mathcal{H}}(T)| = |C_{\mathcal{H}_T}(T)| = |C_{\mathcal{H}_1}(T \setminus \{x_n\})| + |C_{\mathcal{H}_2}(T \setminus \{x_n\})|.$$

Then, using the induction assumption,

$$\begin{aligned}
|C_{\mathcal{H}}(T)| &\leq \Psi(k, n-1) + \Psi(k-1, n-1) \\
&= \sum_{i=0}^{k} \binom{n-1}{i} + \sum_{i=0}^{k-1} \binom{n-1}{i} \\
&= \sum_{i=0}^{k} \binom{n-1}{i} + \sum_{i=1}^{k} \binom{n-1}{i-1} \\
&= \sum_{i=0}^{k} \left[\binom{n-1}{i} + \binom{n-1}{i-1} \right] \\
&= \sum_{i=0}^{k} \binom{n}{i} \\
&= \Psi(k, n).
\end{aligned}$$

Now let's prove the second upper bound (here we assume $n > k$ since otherwise $\Psi(k, n) = 2^n$):

$$\begin{aligned}
\Psi(k, n) &= \sum_{i=0}^{k} \binom{n}{i} \\
&\leq \left(\frac{n}{k}\right)^k \sum_{i=0}^{k} \binom{n}{i} \left(\frac{k}{n}\right)^i \\
&\leq \left(\frac{n}{k}\right)^k \sum_{i=0}^{n} \binom{n}{i} \left(\frac{k}{n}\right)^i.
\end{aligned}$$

The last expression can be rewritten using the binomial theorem, and so

$$\begin{aligned}
\Psi(k, n) &\leq \left(\frac{n}{k}\right)^k \left(1 + \frac{k}{n}\right)^n \\
&\leq \left(\frac{n}{k}\right)^k e^k \\
&= \left(\frac{en}{k}\right)^k.
\end{aligned}$$

□

Lemma 3.14 (Bound on the generalization error)**.** Let $\mathcal{H}$ be any hypothesis class, and let S be a training set of n i.i.d. examples from the unknown distribution $\mathcal{D}$. Then

$$P(\exists\, h \in \mathcal{H} : |\widehat{\mathrm{err}}(h; S) - \mathrm{err}(h; \mathcal{D})| > \varepsilon) \leq 4\, \hat{C}_{\mathcal{H}}(2n)\, e^{-\varepsilon^2 n/8}.$$

The proof is omitted, but it is related to the "bonus quiz" posted on the course website. It uses a "double sampling" trick.

This lemma and Sauer's lemma together imply the following theorem.

Theorem 3.15 (Characterization of agnostic PAC learning)**.** Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VC}(\mathcal{H}) = k$. Then $\mathcal{H}$ is agnostic PAC learnable with sample complexity at most

$$n_{\mathcal{H}}(\varepsilon, \delta) \leq O\left(\frac{k \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon^2}\right).$$

In particular, $\mathcal{H}$ is agnostic PAC learnable via the *ERM approach*.

Outline of proof. How do we prove the characterization of agnostic PAC learnability? First, we obtain $h_S \in \mathcal{H}$ via empirical risk minimization (ERM). Then

$$\begin{aligned} P(|\mathrm{err}(h_S; \mathcal{D}) - \mathrm{err}(h^*; \mathcal{D})| > \varepsilon) &\leq P(2 \sup_{h \in \mathcal{H}} |\widehat{\mathrm{err}}(h; S) - \mathrm{err}(h; \mathcal{D})| > \varepsilon) \\ &= P(\exists\, h \in \mathcal{H} : |\widehat{\mathrm{err}}(h; S) - \mathrm{err}(h; \mathcal{D})| > \varepsilon/2) \\ &\leq 4\, \hat{C}_{\mathcal{H}}(2n)\, \exp(-\varepsilon^2 n/32) \\ &\leq 4 \left(\frac{2en}{k}\right)^k \exp(-\varepsilon^2 n/32). \end{aligned}$$

Then for sufficiently large n, we can get this to be less than δ. Then, we can show that $n_{\mathcal{H}}(\varepsilon, \delta)$ is as given in the theorem above. □

Lemma 3.16 (From the "bonus quiz")**.** Let $\mathcal{H}$ be any hypothesis class. Let S be a training set of n i.i.d. examples from $\mathcal{D}$. Then,

$$P(\exists\, h \in \mathcal{H} : \mathrm{err}(h; \mathcal{D}) > \varepsilon \text{ and } h \text{ is consistent with } S) \leq 2\, \hat{C}_{\mathcal{H}}(2n)\, \exp(-\varepsilon n/4).$$

Using this lemma and Sauer's lemma, we get the following theorem.

Theorem 3.17 (Characterization of PAC learnability)**.** Let $\mathcal{H}$ be a hypothesis class where $\mathrm{VC}(\mathcal{H}) = k$. Then $\mathcal{H}$ is PAC learnable with sample complexity at most

$$n_{\mathcal{H}}(\varepsilon, \delta) \leq O\left(\frac{k \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon}\right).$$

In particular, $\mathcal{H}$ is PAC learnable by an algorithm that outputs a hypothesis that is *consisten*t with a set of $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ training examples.