Algorithms and Uncertainty, Winter 2018/19 Lecture 22 (4 pages)

VC Dimension

Instructor: Thomas Kesselheim

Recall our setting from last time. We have to classify data points from a set X using hypothesis $h\colon X \to \{-1,1\}$. The class of all hypotheses is called $\mathcal{H}$. There is a ground truth $f\colon X \to \{-1,1\}$ and we are in the realizable case, which means that $f \in \mathcal{H}$.

By $\mathcal{H}[m]$ we indicate the maximum number of distinct ways to label m data points from X using different functions in $\mathcal{H}$. A trivial upper bound is $\mathcal{H}[m] \leq 2^m$ but the function can be much smaller.

Given m sample points $x_1, \dots, x_m$ with labels $y_1, \dots, y_m$, the *training error* of a hypothesis is

$$\mathrm{err}_S(h) := \frac{1}{m}|\{h(x_i) \neq y_i\}| \ .$$

The true error $\mathrm{err}_{\mathcal{D}}(h)$ of a hypothesis h with respect to a distribution $\mathcal{D}$ is

$$\mathrm{err}_{\mathcal{D}}(h) := \mathbf{Pr}_{X\sim\mathcal{D}}\left[h(X) \neq f(X)\right] \ .$$

For all choices of $\epsilon > 0$, $\delta > 0$, if we draw m times independently from distribution $\mathcal{D}$ such that

$$m \geq \max\left\{\frac{8}{\epsilon}, \frac{2}{\epsilon}\log_2\left(\frac{2\mathcal{H}[2m]}{\delta}\right)\right\} \ , \tag{1}$$

then with probability at least $1-\delta$, all $h \in \mathcal{H}$ with $\mathrm{err}_S(f) = 0$ have $\mathrm{err}_{\mathcal{D}}(h) < \epsilon$.

Today, we would like to better understand Condition (1). Note that is equivalent to require that

$$\epsilon \geq \max\left\{\frac{8}{m}, \frac{2}{m}\log_2\left(\frac{2\mathcal{H}[2m]}{\delta}\right)\right\} \ .$$

The question that we are interested in is if the true error $\mathrm{err}_{\mathcal{D}}(h)$ vanishes if we choose larger and larger m. This indeed works out if $\frac{\log_2(\mathcal{H}[2m])}{m}$ converges to 0.

For the trivial bound $\mathcal{H}[m] \leq 2^m$, this is not true. For threshold classifiers on a line, we could show that $\mathcal{H}[m] \leq m+1$. This is sufficient. More generally, we ask: Is there a point after which $\mathcal{H}[m]$ grows subexponentially?

1 VC Dimension

Today, we will get to know the central notion of *VC dimension*. It was introduced by Vapnik and Chervonenkis in 1968. The VC dimension of a set of hypotheses $\mathcal{H}$ is roughly the point from which on $\mathcal{H}[m]$ is smaller than 2^m.

Definition 22.1. *A set of hypotheses $\mathcal{H}$ shatters a set $S \subseteq X$ if there are hypotheses in $\mathcal{H}$ that label S in all possible $2^{|S|}$ ways, that is, $\mathcal{H}[S] = 2^{|S|}$.*

Definition 22.2. *The* VC dimension *of a set of hypotheses $\mathcal{H}$ is the largest size of a set S that is shattered by $\mathcal{H}$, i.e., $\max\{|S| \mid \mathcal{H}[S] = 2^{|S|}\}$. If there are sets of unbounded sizes that are shattered then the VC dimension is infinite.*

Let us consider a few examples.

- For $X = \mathbb{R}$ and $\mathcal{H}$ being the class of functions of the form

$$h(x) = \begin{cases} -1 & \text{for } x \leq t \\ 1 & \text{otherwise} \end{cases}$$

  the VC dimension is 1. This is because any set $\{x\}$ is shattered because $h(x) = -1$ and $h'(x) = 1$ for suitable choices of h and h'. In contract, for any set of two points $x_1 \leq x_2 \in \mathbb{R}$, it is impossible that $h(x_1) = 1$ but $h(x_2) = -1$.

- If $\mathcal{H}$ is finite, then the VC dimension is at most $\log_2|\mathcal{H}|$.

- If X is infinite and $\mathcal{H}$ contains all functions $h\colon X \to \{-1, 1\}$, then the VC dimension is infinite.

2 Bounding the Growth Function by the VC Dimension

Theorem 22.3 (Sauer's Lemma)**.** *Let $\mathcal{H}$ be a hypothesis class of VC dimension d. Then for all $m \geq d$*

$$\mathcal{H}[m] \leq \sum_{i=0}^{d} \binom{m}{i} \ .$$

In order to prove Sauer's Lemma, the following lemma will turn out to be very helpful.

Lemma 22.4. *Consider a set of data points $S \subseteq X$ and let L be an arbitrary set of labelings $\ell\colon S \to \{-1, 1\}$. Then L shatters at least $|L|$ subsets of S. That is, there are at least $|L|$ distinct sets $S' \subseteq S$ such that S' can be labelled in all $2^{|S'|}$ different ways using functions from L.*

Proof. We prove the claim by induction on $|L|$. The base case is $|L| = 1$. In this case, the empty set is shattered.

For the induction step, consider that $|L| > 1$. In this case, there has to be some $x \in S$ such that $\ell(x) = -1$ for some $\ell \in L$ and $\ell'(x) = 1$ for some $\ell' \in L$. Let $L_- = \{\ell \in L \mid \ell(x) = -1\}$ and $L_+ = \{\ell \in L \mid \ell(x) = 1\}$. Now, apply the induction hypothesis on the sets L_- and L_+. Let $T_- \subseteq 2^S$ and $T_+ \subseteq 2^S$ denote the shattered sets respectively. By induction hypothesis, we have $|T_-| \geq L_-$ and $|T_+| \geq L_+$.

Note that there is no $S' \in T_-$ or $S' \in T_+$ with $x \in S'$ because the label of x is always fixed to -1 or 1.

All of $T_- \cup T_+$ is shattered by L. Additionally, if $S' \in T_- \cap T_+$, then $S' \cup \{x\}$ is also shattered by L because after assigning x an arbitrary label we can still assign all possible labels to the S' using a labelling in L. All sets constructed this way are not contained in T_- or T_+ because they always contain x.

Consequently, the number of shattered sets is at least

$$|T_- \cup T_+| + |T_- \cap T_+| = |T_-| + |T_+| - |T_- \cap T_+| + |T_- \cap T_+| = |T_-| + |T_+| \geq |L_-| + |L_+| = |L| \ . \ \square$$

Proof of Sauer's Lemma. Given any set $S \subseteq X$ of size m, we would like to bound $\mathcal{H}[S]$. To this end, let L be the set of possible labelings $\ell\colon S \to \{-1, 1\}$ applying different hypotheses from $\mathcal{H}$ on S. Formally, $L = \{h|_S \mid h \in \mathcal{H}\}$. By definition $\mathcal{H}[S] = |L|$.

Furthermore, let $T \subseteq 2^S$ be the family of subsets of S that are shattered by $\mathcal{H}$. Using Lemma 22.4, we know that $|T| \geq |L|$.

We also know that no set larger than d can be shattered, so T contains sets of size at most d. Therefore, the size of T is bounded by the number of such sets

$$|T| \leq \sum_{i=0}^{d} \binom{m}{i} \ .$$

In combination, $\mathcal{H}[S] = |L| \leq |T| \leq \sum_{i=0}^{d} \binom{m}{i}$. □

To simplify the expression in Sauer's Lemma, we can use the following bound on the binomial coefficients

$$\binom{m}{i} = \frac{m!}{(m-i)! \cdot i!} \leq \frac{m^i}{i!} = \left(\frac{m}{d}\right)^i \frac{d^i}{i!} \leq \left(\frac{m}{d}\right)^d \frac{d^i}{i!} \ .$$

Together with the definition of the exponetial function $\mathrm{e}^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$, we get

$$\sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \left(\frac{m}{d}\right)^d \frac{d^i}{i!} = \left(\frac{m}{d}\right)^d \sum_{i=0}^{d} \frac{d^i}{i!} \leq \left(\frac{m}{d}\right)^d \mathrm{e}^d \ .$$

This gives us the following corollary.

Corollary 22.5. *Let $\mathcal{H}$ be a hypothesis class of VC dimension d. Then for all $m \geq d$*

$$\mathcal{H}[m] \leq \left(\frac{\mathrm{e}m}{d}\right)^d \ .$$

Plugging this bound into Condition (1), we get that for a hypothesis class $\mathcal{H}$ of VC dimension d for all choices of $\epsilon > 0$, $\delta > 0$ if we draw m times independently from distribution $\mathcal{D}$ such that

$$m \geq \max\left\{\frac{8}{\epsilon}, \frac{2}{\epsilon}\log_2\left(\frac{2\left(\frac{2\mathrm{e}m}{d}\right)^d}{\delta}\right)\right\} = \max\left\{\frac{8}{\epsilon}, \frac{2d}{\epsilon}\log_2\left(\frac{2\mathrm{e}m}{d}\right) + \frac{2}{\epsilon}\log_2\left(\frac{2}{\delta}\right)\right\} \ ,$$

then with probability at least $1-\delta$, all $h \in \mathcal{H}$ with $\mathrm{err}_S(f) = 0$ have $\mathrm{err}_{\mathcal{D}}(h) < \epsilon$.

Corollary 22.6. *Any hypothesis class of finite VC dimension is PAC-learnable.*

3 Infinite VC Dimension

Not all hypothesis classes have a finite VC dimension. One example would be the set of all functions $X \to \{0,1\}$. As we will show, these hypothesis classes are not PAC-learnable.

Theorem 22.7. *Any hypothesis class of infinite VC dimension is not PAC-learnable.*

To show this theorem, we have to show that the function $m_{\mathcal{H}}$ in the definition of PAC-learnability does not exist. We will show the following.

Proposition 22.8. *Let $\mathcal{H}$ be a hypothesis class of VC dimension at least d. Then for every learning algorithm there exists a distribution such on that on a training set of size $\frac{d}{2}$ we have $\mathrm{err}_{\mathcal{D}}(h_S) \geq \frac{1}{8}$ with probability at least $\frac{1}{7}$.*

Proof. By definition $\mathcal{H}$ shatters a set of size d. So, let $T \subseteq X$, $|T| = d$, be such a set. By definition, any labeling $\ell\colon T \to \{-1, 1\}$ can be extended to a hypothesis $f \in \mathcal{H}$ such that $\ell(x) = f(x)$ for all $x \in T$. There are $k = 2^d$ such labelings. Let $f_1, \dots, f_k$, be the respective extended hypotheses. Each of them can be the ground truth. Let $\mathcal{D}_i$ denote the uniform distribution over pairs $(x, f_i(x))$ for $x \in T$.

Our learning algorithm will have to infer the correct i. The important observation is that any sample of size at most $\frac{d}{2}$ tells us the correct labels of only at most $\frac{d}{2}$ points in T. The others are still completely arbitrary.

Let h_S be the hypothesis computed by the learning algorithm on sample S. In principle, this may also be randomized. Our goal is to show that

$$\max_i \mathbf{Pr}\left[\mathrm{err}_{\mathcal{D}_i}(h_S) \geq \frac{1}{8}\right] \geq \frac{1}{7} \ .$$

We will apply Yao's principle: Draw I uniformly from $\{1, \dots, k\}$ and consider $\mathcal{D}_I$. This is potentially confusing: We first draw index I randomly and then we use probability distribution $\mathcal{D}_I$. Now, it suffices to show that

$$\mathbf{Pr}\left[\mathrm{err}_{\mathcal{D}_I}(h_S) \geq \frac{1}{8}\right] \geq \frac{1}{7} \ .$$

Fix any $x \in X$. We bound the probability that $h_S(x) \neq f_I(x)$. To this end, we think of the labels f_I being determined in a different way. First draw the sample S and determine the labels for the points in this set. Based on this, compute h_S. Only now determine the labels for the points not in this set. If x is not in the sample, then $h_S(x)$ is correct with probability $\frac{1}{2}$. It is not in the sample with probability at least $\frac{1}{2}$. Therefore

$$\mathbf{Pr}\left[h_S(x) \neq f_I(x)\right] \geq \frac{1}{4} \ .$$

This holds for all $x \in X$, therefore

$$\mathbf{E}\left[\mathrm{err}_{\mathcal{D}_I}(h_S)\right] \geq \frac{1}{4} \ .$$

Now, we can apply Markov's inequality to get

$$\mathbf{Pr}\left[\mathrm{err}_{\mathcal{D}_I}(h_S) < \frac{1}{8}\right] = \mathbf{Pr}\left[1 - \mathrm{err}_{\mathcal{D}_I}(h_S) > \frac{7}{8}\right] \leq \frac{1}{\frac{7}{8}}\mathbf{E}\left[1 - \mathrm{err}_{\mathcal{D}_I}(h_S)\right] \leq \frac{3}{4} \cdot \frac{8}{7} = \frac{6}{7} \ .$$

This proves the claim. □

References and Further Reading

These notes are based on notes and lectures by Anna Karlin `https://courses.cs.washington.edu/courses/cse522/17sp/` and Avrim Blum `http://www.cs.cmu.edu/~avrim/ML14/`. Also see the references therein.