Algorithms and Uncertainty, Winter 2017/18 Lecture 15 (4 pages)

VC Dimension

Instructor: Thomas Kesselheim

Recall our setting from last time. We can to classify data points from a set X using hypothesis $h\colon X \to \{0,1\}$. The class of all hypotheses is called $\mathcal{H}$. There is a ground truth $f\colon X \to \{0,1\}$ 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 $x_1, \ldots, x_m$ with labels $y_1, \ldots, y_m$, the *training error* of a hypothesis is

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

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

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

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

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

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

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

The question that we are interested in is if the generalization 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 15.1. *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 15.2. *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

$$h(x) = \begin{cases} 0 & \text{for } x \leq t \\ 1 & \text{otherwise} \end{cases}$$

  the VC dimension is 1. This is because any set $\{x\}$ is shattered because $h(x) = 0$ 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 $h(x_2) = 0$.

- 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 $h \colon X \to \{0,1\}$, then the VC dimension is infinite.

2 Bounding the Growth Function by the VC Dimension

Theorem 15.3 (Sauer's Lemma). *Let $\mathcal{H}$ be a hypothesis class of VC dimension d. Then for all $m \geq d$*

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

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

Lemma 15.4. *Consider a set of data points $S \subseteq X$ and let L be an arbitrary set of labellings $\ell \colon S \to \{0,1\}$. 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 $\ell(x) = 0$ for some $\ell \in L$ and $\ell'(x) = 1$ for some $\ell' \in L$. For $i \in \{0,1\}$, let $L_i = \{\ell \in L \mid \ell(x) = i\}$. Now, apply the induction hypothesis on the sets L_0 and L_1. Let $T_0 \subseteq 2^S$ and $T_1 \subseteq 2^S$ denote the shattered sets respectively. By induction hypothesis, we have $|T_0| \geq L_0$ and $|T_1| \geq L_1$.

Note that there is no $S' \in T_i$ with $x \in S'$ because the label of x is always fixed to i.

All of $T_0 \cup T_1$ is shattered by L. Additionally, if $S' \in T_0 \cap T_1$, 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_0 or T_1 because they always contain x.

Consequently, the number of shattered sets is at least

$$|T_0 \cup T_1| + |T_0 \cap T_1| = |T_0| + |T_1| - |T_0 \cap T_1| + |T_0 \cap T_1| = |T_0| + |T_1| \geq |L_0| + |L_1| = |L| \enspace . \quad \square$$

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 labellings $\ell \colon S \to \{0,1\}$ 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 15.4, 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 15.5. *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 15.6. *Any hypothesis class of finite VC dimension is PAC-learnable.*

3 The Unrealizable Case

In our results so far, we only considered the "realizable case". That is, there is a ground truth $f\colon X \to \{0,1\}$ and $f \in \mathcal{H}$. Actually, in any machine learning setting, this is too strong an assumption. Usually, the features do not describe a concept entirely. Coming back to our original example of spam classification, typical features might be word counts, sender IP addresses, header data, and so on. Of course, based on only this information, it is impossible to fully correctly classify all e-mails. Even if it was possible, we might choose only a smaller hypothesis class $\mathcal{H}$ to allow efficient learning.

In the unrealizable case, one therefore asks how many sample are necessary to be able to correctly estimate the generalization error from the training error. We ask that with probability at least $1-\delta$

$$|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| < \epsilon \qquad \text{for all } h \in \mathcal{H}.$$

If this condition is fulfilled then choosing a hypothesis with small generalization error is approximately the same as choosing one with small training error.

For this *uniform convergence*, a similar theory exists. Instead of Condition (1), it is now sufficient if

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

So, most importantly, there is such a choice of m whenever the VC dimension is finite.

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.