The VC dimension

Previously, the error has been split into approximation (bias) and estimation error. The definition of PAC learnability requires that the estimation error is bounded over all distributions. Next, the goal is to find out which classes of $\mathcal{H}$ are PAC-learnable, and so far finite hypothesis classes have only proven to be PAC learnable. The class of all hypothesis functions over a domain has shown **not** to be PAC learnable. In this chapter, it is shown that infinite clases can be PAC learnable. The set of learnable classes is defined nicely in the setting of binary classification and 0-1 loss by Vapnik and Chervonenkis (1970) as the VC-dimension.

6.1 Infinite-size classes can be learnable

Finite size classes are learnable with their complexity bounded by the log of the their size. But, the size is not the right characterization of complexity.

Example 6.1 Define a threshold function $h_a(x) = 1_{[x<a]}$. Let $\mathcal{H}$ be the set of all hypotheses with this function over a domain which is clearly infinite.

Lemma 6.1 Using the ERM rule, $\mathcal{H}$ is PAC learnable with sample complexity $m_h(\epsilon, \delta) \leq log(2/\delta)/\epsilon$

Proof If there is an $a*$ such that the error is zero using that $a*$ and set some $a_0 < a* < a_1$ such that the probability of getting x less or more than $a*$ is exactly ϵ. Now, we define a training set S, with bounds $b_0 \leq b_S(ERM) \leq b_1$. A sufficient condition for $L_D(h_S) \leq \epsilon$ is that both $b_0 \geq a_0$ and $b_1 leq a_1$.

- TODO: question on clarity here (shouldn't the condition be inverted ?)

$$P_{S\sim D^m}[L_D(h_S) > \epsilon] \leq P_{S\sim D^m}[b_0 < a_0] + P_{S\sim D^m}[b_1 > a_1]$$

Now, we take each probability, and assume they happen if all training examples are not in $(a_0, a*)$ and $(a*, a_1$ for b_0 and b_1 respectively. Looking back at the starting definition of probablity ϵ, this probability is exactly $(1 - \epsilon)$, and since the training set size is of m, we have

$$P_{S\sim D^m}[b_0 < a_0] = P_{S\sim D^m}[\forall(x, y) \in S, x \notin (a_0, a*)] = (1 - \epsilon)^m \leq e^{-\epsilon m}$$
$$P_{S\sim D^m}[L_D(h_S) > \epsilon] \leq 2e^{-\epsilon m}$$

$$\delta \leq 2e^{-\epsilon m}$$

This proves that infinite hypothesis classes can indeed be learnable, which we show using the threshold function.

6.2 The VC dimension

In the no-free-lunch theorem, it has been shown that without restricting the hypothesis class and *evil* adversary can construct a distribution which for the learning algorithm will perform poorly, while another learner will succeed. The adversary used a finite subset of the domain $C \subset \mathcal{X}$ and considered a distribution focused around C, deriving the distribution from C to $0, 1$, using the fact that the powerset $2^{|C|}$ of functions was available –> **very important**.

Definition 6.2 Next, we define restrictions of $\mathcal{H}$ to C. $\mathcal{H}$ is the class of functions from $\mathcal{X}$ to $0, 1$. We say that $\mathcal{H}$ to C is a restriction, set of functions from C to $0, 1$ which can be derived from $\mathcal{H}$. If the restriction from $\mathcal{H}$ to C involves all possible functions from C to $0, 1$, then $\mathcal{H}$ shatters C ($|\mathcal{H}_C| = 2^{|C|}$).

Consider the thresholding function as $\mathcal{H}$ and a single datapoint in C. $\mathcal{H}$ shatters C since all functions from C to $0, 1$ can be derived from the threshold function. If we include another datapoint, this is no longer valid with the threshold function (we might need something like a range function).

Whenever some set C is shattered by $\mathcal{H}$, the adversary is not restricted by $\mathcal{H}$ and the no-free-lunch theorem applies.

Corollary 6.4 Extending the no-free-lunch theorem, we state that if 2m is the size of C and if $\mathcal{H}$ shatters C, then there exists a distribution over $\mathcal{X} \times 0, 1$ and a predictor $h \in \mathcal{H}$ such that $L_D(h) = 0$, but with probability of at least $1/7$, $L_D(A(S)) \geq 1/8$.

In other words, if $\mathcal{H}$ shatters C of size $2m$, then one cannot learn C using m examples.

Definition 6.5. VC-dimension The VC dimension of a hypothesis class is a maximum size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. One uses the maximum, since removing from the set C makes it easier to shatter.

Theorem 6.6 If VC-dimension is infinite, then $\mathcal{H}$ is **not** PAC learnable.

6.3 Examples

Now we wish to show examples of calculating VC dimension on several hypothesis classes by showing that $VCdim(\mathcal{H}) = d$ if there exists a set C of size d shattered by $\mathcal{H}$ and that every set of size $d + 1$ is **not** shattered by $\mathcal{H}$.

6.3.1 Threshold functions

We first try shattering a set of $C = \{c_1\}$ using threshold functions. As all possible labeling are possible, C is shattered. Growing $C = \{c_1, c_2\}$ by one more does not allow all possible labelings as $(0, 1)$ can't be achieved. Therefore $VCdim(\mathcal{H}) = 1$.

6.3.2 Intervals 6.3.3. Axis aligned rectangles

Similarly, as with thresholds we attempt different labellings with intervals ($VCdim(\mathcal{H}) = 2$) and axis aligned intervals ($VCdim(\mathcal{H}) = 4$).

6.3.4 Finite classes

We have seen that infinite classes can have finite VC dimension. For finite classes, we know their size $|\mathcal{H}|$ and we know the size of $|C|$, therefore we can state that $\mathcal{H}$ shatters C if $|\mathcal{H}_C| \leq |\mathcal{H}|$ and $|\mathcal{H}| \geq 2^{|C|}$. So, the upper bound of $VCdim$ of finite classes is $log_2(|\mathcal{H}|)$.

6.3.5 VC-Dimension and the Number of Parameters

VC-dimension equalled number of parameters (size of |C|) defining the hypothesis class. This is not always true.

6.4 The fundemental theorem of PAC learning

Theorem 6.7 Fundemental Theorem of Statistical Learning \ Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0, 1\}$. with $0 - 1$ loss. Then, the following claims are equivalent:

1. $\mathcal{H}$ has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC leaner for $\mathcal{H}$
3. $\mathcal{H}$ is agnostic PAC learnable
4. $\mathcal{H}$ is PAC learnable
5. Any ERM rule is a successful PAC learner for $\mathcal{H}$
6. $\mathcal{H}$ has a finite VC-dimension

- FUN FACT: which definition came first? VC then, PAC (1989), then proof VC – PAC (1989)

Theorem 6.8. Fundemental theorem of Statistical Learning – Quantitative Version. Same conditions for 6.7. apply, but also VC dimension is finite $d < \infty$ There exist absolute constants C_1, C_2 such that

1. $\mathcal{H}$ has the uniform convergence property with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

2. $\mathcal{H}$ is agnostic PAC learnable with sample complexity (same as above)

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

3. $\mathcal{H}$ is PAC learnable with sample complexity 3. $\mathcal{H}$ is PAC learnable with sample complexity 3. $\mathcal{H}$ is PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

The theorem holds for binary classification tasks, but not for all tasks. Also, if ERM fails, learning is possible through other methods.

6.5 Proof of theorem 6.7

It is difficult to show 6 -> 1. We use two claims:

1. When VCdim is finite, the size of the restricted set C grows polynomially, rather than exponentially with $|C|$. This claim is referred to as *Sauer's lemma.*
2. Uniform convergence has been proven on finite hypothesis classes. It will be shown that uniform convergence holds for "small effective size" classes (that is classes from 1.)

6.5.1 Sauer's Lemma and the Growth Function

The growth function measures the "effective" size of $\mathcal{H}$ on a set of m examples. Formally defined, the growth function is the number of different functions from a set $|C|$ of size m to $\{0, 1\}$ that can be obtained by restricting $\mathcal{H}$ to C :

$$\tau_{\mathcal{H}} = \max_{C \in \mathcal{X}, |C|=m} |\mathcal{H}_C|$$

If $VCdim(\mathcal{H}) = d$, then for any $m \leq d$ we have that $\tau_{\mathcal{H}} = 2^m$.

6.5.2 Uniform Convergence for Classes of Small Effective Size