CS6781 - Theoretical Foundations of Machine Learning

Lecture 4: Combinatorial Dimensions for Learning

January 30, 2020

Lecturer: Nika Haghtalab *Readings: Chp 6, UML*

1 Examples of PAC Learning

Can we PAC learn the following hypothesis classes?

Monotone Conjunctions. Yes! Last lecture we saw that we can efficiently learn monotone conjunctions in the consistency model. Note that there are at most 2^n monotone conjunctions. So, using Theorem 2.1, we can PAC learn monotone conjunctions with $m(\epsilon, \delta) = O\left(\epsilon^{-1}(n + \ln(1/\delta)\right)$.

DNFs. Is the set of all DNF functions efficiently (in the number of variables) learnable in the PAC model (by the class of DNFs)? In last lecture, we showed that DNFs are efficiently learnable in the consistency model. So, it remains to see whether the sample complexity of learning this class is polynomial in $1/\epsilon$, $1/\delta$, and the number of variables n.

Let us count the number of DNFs in existence. We noted that DNFs can express all possible functions on n variable. Note that there are $|\mathcal{X}| = 2^n$ instances in the domain, so there are $|\mathcal{C}| = 2^{|\mathcal{X}|}$ DNFs. Using Theorem 2.1 from the previous lecture gives us that

$$m_{\mathcal{C}}(\epsilon, \delta) = O\left(\frac{1}{\epsilon}(2^n + \ln(1/\delta))\right).$$

The exponential dependence of $m_{\mathcal{C}}(\epsilon, \delta)$ on n is not *sufficient* for showing that DNFs can be learned efficiently in n, $1/\epsilon$, $1/\delta$, in the PAC model We will see later that indeed this exponential dependence is necessary and DNFs cannot be learned efficiently in $\mathrm{poly}(n, 1/\epsilon, 1/\delta)$, in the PAC model.

Linear thresholds Are linear thresholds learnable in the PAC model? Previously, we saw that we can efficiently learn linear thresholds in the consistency model. But the set of all linear thresholds is infinitely large. Therefore, we cannot directly use Theorem 2.1 from the previous lecture to show that linear thresholds are PAC learnable. In the remainder of the lecture, we explore how we could reason about PAC learnability of infinite hypothesis classes by understanding their structure.

2 Effective Number of Hypotheses and VC Dimension

So far we have seen that when a concept class $\mathcal{C}$ is finite, with probability $1 - \delta$, any concept $c \in \mathcal{C}$ that is consistent with a sample set S of at least $\epsilon^{-1}(\ln(|\mathcal{C}|) + \ln(1/\delta))$ i.i.d. labeled instances has true error of $\mathrm{err}_{\mathcal{C}}(c) \leq \epsilon$. Unfortunately, when $\mathcal{C}$ is very large or infinite, as in many applications of machine learning, this sample complexity bound becomes less useful or even trivial. In this lecture, we see how we can prove non-trivial bounds on the sample complexity of $\mathcal{C}$ even if its infinitely large. We will introduce important concepts such as VC dimension and growth function and show how these combinatorial notions play an integral role in the theory of learnability.

Even when $\mathcal{C}$ is infinitely large it is possible that many of the hypothesis in $\mathcal{C}$ project to the "same type of behavior" on a sample set S. Formally, we define the set of all distinct labelings produced by a hypothesis class $\mathcal{C}$ on a set of (unlabeled) instances $S = \{x_1, \ldots, x_m\}$ by

$$\mathcal{C}[S] = \{(c(x_1), c(x_2), \ldots, c(x_m))\}_{c \in \mathcal{C}}.$$

That is, $\mathcal{C}[S]$ is the set of "projections" or "restrictions" of all concepts in $\mathcal{C}$ to the set of samples in S. At a high level, the larger $\mathcal{C}[S]$ can be made on a sample set S, the more expressive and complex $\mathcal{C}$ is. To capture this, we define the *growth function* as follows.

Definition 2.1. The *growth function* of $\mathcal{C}$ is the largest cardinality of $\mathcal{C}[S]$ as a function of the size of S, i.e.,

$$\Pi_{\mathcal{C}}(m) = \max_{S \in \mathcal{X}^m} |\mathcal{C}[S]| \,.$$

Fact 2.2. $\Pi_{\mathcal{C}}(m) \leq 2^m$ *for any concept class with binary labels.*

As seen below, the importance of growth function is that it determines the sample complexity of PAC learning. That is, the size of $\mathcal{C}$ does not matter. Rather, it is the growth function of $\mathcal{C}$ that matters.

Theorem 2.3 (PAC Learnability of Infinite Concept Classes)**.** *Let $\mathcal{A}$ be an algorithm that learns a concept class $\mathcal{C}$ in the consistency model. Then, $\mathcal{A}$ learns the concept class $\mathcal{C}$ in the PAC learning model using a number of samples that satisfies*

$$m \geq \frac{c_0}{\epsilon}\left(\ln(\Pi_{\mathcal{C}}(m)) + \ln(\frac{1}{\delta})\right), \tag{1}$$

for a fixed constant c_0.

Interpreting the above theorem may look challenging at first sight because the dependence on m appears on both sides of the inequality. For example, when $\Pi_{\mathcal{C}}(m) = 2^m$, Equation 1 cannot be satisfied and the guarantees of Theorem 2.3 falls apart.

Next we define an important combinatorial notion that helps us bound and interpret $\Pi_{\mathcal{C}}(m)$ and simplify the sample complexity bound of Theorem 2.3.

Definition 2.4. We say that a set of unlabeled instances S is *shattered* by $\mathcal{C}$, if $|\mathcal{C}[S]| = 2^{|S|}$. That is, $\mathcal{C}$ labels S in every way possible.

Definition 2.5. The Vapnik-Chervonenkis (VC) dimension of $\mathcal{C}$, denoted by $\mathrm{VCDim}(\mathcal{C})$ is the size of the largest set S that is shattered by $\mathcal{C}$.

Note that to show that VC dimension of a class $\mathcal{C}$ is d, we must show two things:

- There *exists* an instance set S of size d that is shattered by $\mathcal{C}$.
- *No* set of size $d + 1$ exists that can be shattered by $\mathcal{C}$.

3 Examples

As we see next, in many infinite concept class $\Pi_{\mathcal{C}}(m) \ll 2^m$.

1-D thresholds. Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{C} = \{h_a : a \in \mathbb{R}\}$ be the set of 1-dimensional (positive halfspace) thresholds, where

$$h_a(x) = 1(x \geq a).$$

What is the growth function of the class of 1-D thresholds? Take an arbitrary set of m instances in $\mathcal{X}$ and for ease of representation assume that they are in increasing order $S = \{x_1, \dots, x_m\}$. Note that the labeling each 1-D threshold h_a produces can be fully determined by the the smallest x_i labeled positive, i.e., smallest $x_i \geq a$. There are at most $\binom{m}{0} + \binom{m}{1} = m + 1$ ways to choose which zero to one of the instances in S is the minimum instance being labeled positive. Therefore, $\Pi_{\mathcal{C}}(m) \leq m + 1$. Alternatively, all thresholds h_a for $a \in [x_i, x_{i+1})$ label all instances similarly. There are at most $m + 1$ intervals, namely $(-\infty, x_1), [x_1, x_2), \dots, [x_{m-1}, x_m), [x_m, \infty)$ that lead to unique labelings of the set of instance S. Therefore, $\Pi_{\mathcal{C}}(m) \leq m + 1$.

What is the VC dimension of this class? Clearly, we can shatter one instance. However, no set of two instances $\{x_1, x_2\}$ can be shattered. This is because, assuming that $x_1 \leq x_2$, we cannot label x_1 as positive and x_2 as negative. So the VC dimension of 1-D thresholds is 1.

1-D intervals Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{C} = \{h_{a,b} : a, b \in \mathbb{R}\}$ be the set of 1-dimensional intervals, where $h_{a,b}(x) = 1(a \leq x \leq b)$.

What is the growth function of the class of 1-D intervals? Take an arbitrary set of m instances in $\mathcal{X}$ and for ease of representation assume that they are in increasing order $S = \{x_1, \dots, x_m\}$. So, each labeling produced by a hypothesis can be uniquely determined by the smallest and largest instances in S that are labeled as positives. In more detail, when at least 2 instances in S are labeled as positives, the largest and smallest positive instance are distinct. When only 1 instance in S is labeled as positives, then the smallest and largest positive instance is one and the same. And lastly, when 0 instance is labeled as positive, then the largest and smallest positive instances do not exist in S. All together, there are at most

$$\binom{m}{0} + \binom{m}{1} + \binom{m}{2} = \frac{m^2 + m + 2}{2}$$

labelings that can be produced by $\mathcal{C}$ on S.

What is the VC dimension of the class of 1-D intervals? It's not hard to see that we can shatter a set of 2 points. However, we cannot shatter any set of three points, because given $x_1 \leq x_2 \leq x_3$, we cannot label x_2 as negative and x_1, x_3 as positive. So the VC dimension of 1-D intervals is 2.

Axis-aligned rectangles in $\mathbb{R}^2$. We defined this class in the previous lectures. *What is the VC dimension of this class?* It's not hard to see that a set of 4 instances placed on the mid points of edges of a rectangle is shattered by axis-aligned rectangles. But, no set of 5 instances can be shattered. This is because, given any 5 points at least one of them is in the interior of the bounding box of the other four. Therefore, we cannot label the interior instance as negative while labeling the rest of the instances as positives. Therefore, VC dimension of this class is 4.

What is the growth function of this class? Following the same idea as intervals in 1-dimension, we note that each axis-aligned rectangle is an intersection of one horizontal and one vertical interval in one dimension. So the number of labels that can be produced by the set of rectangles is at most the product of the number of labelings that can be produced by considering 1-dimensional intervals on the two coordinates individually. Therefore, $\Pi_{\mathcal{C}}(m) \leq \left(\frac{m^2+m+2}{2}\right)^2$.

4 Relating VC Dimension and Growth Function

In the next lecture, we prove the following important relationship between the growth function of a hypothesis class and its VC dimension.

Lemma 4.1 (Sauer's Lemma)**.** *Consider any hypothesis class $\mathcal{C}$ and let $d = \mathrm{VCDim}(\mathcal{C})$. For all m,*

$$\Pi_{\mathcal{C}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Lemma 4.2. *For $m > d$,*

$$\sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d \in O(m^d).$$

Proof. We have

$$\begin{aligned}
\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} &\leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} && \text{Since } d/m \leq 1\\
&\leq \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i} && \text{Adding non-negative terms}\\
&\leq \left(1+\frac{d}{m}\right)^m && \text{The Binomial theorem}\\
&\leq e^d.
\end{aligned}$$

□

So one way to interpret VC dimension is that it's the value at which the growth function $\Pi_{\mathcal{C}}(m)$ stops growing exponentially — i.e., 2^m because the set is being shattered — and become a polynomial, i.e. $O(m^d)$. In other words, $\ln(\Pi_{\mathcal{C}}(m))$ stops growing as $O(m)$ and starts to grow as $\log(m)$. Note that it is precisely this logarithmic dependence on m that makes the bound in Theorem 2.3 meaningful. Using the above bound on the growth function in Theorem 2.3,

$$m_{\mathcal{C}}(\epsilon,\delta) \in O\left(\frac{1}{\epsilon}\left(\mathrm{VCDim}(\mathcal{C})\ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})\right)\right).$$