COMP 585 **Spring 2024**

Probabilistic Toolkit for Learning and Computing *Instructor: Maryam Aliakbarpour*

March 7th, 2024 *Scriber: Jason S. Lenderman*

Lecture 16

1 Growth function, shattering and VC-dimension

Let $S = \{x_1, \dots, x_m\}$ be a set of m points from instance space $\mathcal{X}$. Let $\mathcal{C}$ be a concept class over instance space $\mathcal{X}$. The restriction of $\mathcal{C}$ to S, denoted by $\mathcal{C}_S$ is defined to be

$$\mathcal{C}_S = \{(c(x_1), \dots, c(x_m)) : c \in \mathcal{C}\}$$

where each function from S to $\{0,1\}$ is represented by an element of $\{0,1\}^{|S|}$ (or $\{0,1\}^m$.) While $\mathcal{C}$ might be infinite, it's "effective size" might be small in the sense that the number of functions induced by the restriction of $\mathcal{C}$ to any m points from $\mathcal{X}$ might be significantly smaller than 2^m (the obvious maximum number of such functions.) To formalize this we defined the growth function $\tau_{\mathcal{C}} : \mathbb{N} \to \mathbb{N}$ as follows

$$\tau_{\mathcal{C}}(m) = \sup_{S \subseteq \mathcal{X} : |S| = m} |\mathcal{C}_S| \, .$$

In words, $\tau_{\mathcal{C}}(m)$ is the maximum number of functions from $[m] to \{0,1\}$ that can be realized by considering restrictions of $\mathcal{C}$ to m-sets of points of belonging to $\mathcal{X}$.

As mentioned earlier we have the trivial upper bound $\tau_{\mathcal{C}}(m) \le 2^m$. However, under a simple condition we can get a much better upper bound on $\tau_{\mathcal{C}}(m)$. Take $S \subseteq \mathcal{X}$ with $|S| < \infty$. We say that $\mathcal{C}$ *shatters* finite S if $|\mathcal{C}_S| = 2^{|S|}$.

Example 1 (Axis-aligned rectangles)**.** To make the notion shattering more clear take $\mathcal{C}$ to be the collection of all axis-aligned rectangles in $\mathbb{R}^2$, i.e.

$$\mathcal{C} = \{[x_0, x_1] \times [y_0, y_1] : x_0, x_1, y_0, y_1 \text{ with } x_0 \le x_1 \text{ and } y_0 \le y_1\} \, .$$

It is easily verified that $|\mathcal{C}_S| = 2^{|S|}$ for any $S \subseteq \mathbb{R}^2$ with $|S| \in \{1,2\}$. So $\mathcal{C}$ shatters any 1 or 2-subset of $\mathbb{R}^2$. What about the 3-subset $\{(0,0),(1,1),(2,2)\}$ or the 4-subset $\{(0,-1),(0,1),(-1,0),(1,0)\}$. Can $\mathcal{C}$ shatter these sets?

We now define the VC-dimension of the concept class $\mathcal{C}$, denoted by $\mathrm{VC}_{\mathrm{dim}}(\mathcal{C})$, to be the largest $m \in \mathbb{N}$ such that there exists an $S \subseteq \mathcal{X}$ with $|S| = m$ that can be shattered by $\mathcal{C}$. If for each $m \in \mathbb{N}$ there exists an $S \subseteq \mathcal{X}$ with $|S| = m$ that can be shattered by $\mathcal{C}$ then we define $\mathrm{VC}_{\mathrm{dim}}(\mathcal{C}) = \infty$.

Example 2 (Axis-aligned rectangles (revisited))**.** Let us consider again the concept class $\mathcal{C}$ of axis-aligned rectangles in $\mathbb{R}^2$. Here we have $\mathrm{VC}_{\mathrm{dim}}(\mathcal{C}) = 4$. In order to establish this we need to show:

- There is a $S \subseteq \mathbb{R}^2$ that is shattered by $\mathcal{C}$. For example, $S = \{(0,-1),(0,1),(-1,0),(1,0)\}$.
- There is not subset S of $\mathbb{R}^2$ with $|S| \geq 5$ that is shattered by $\mathcal{C}$.

Example 3 (Finite classes)**.** If the concept class $\mathcal{C}$ is finite then for any finite $S \subseteq \mathcal{X}$ we have clearly have

$$|\mathcal{C}_S| \leq |\mathcal{C}| = 2^{\log_2(|\mathcal{C}|)},$$

which implies that $\mathcal{C}$ cannot shatter any set with more than $\lfloor \log_2(|\mathcal{C}|) \rfloor$ points. Hence $\mathrm{VC}_{\mathrm{dim}}(\mathcal{C}) \leq \lfloor \log_2(|\mathcal{C}|) \rfloor$.

Note that if $\mathrm{VC}_{\mathrm{dim}}(\mathcal{C}) = d$ then $\tau_{\mathcal{C}}(m) = 2^m$ when $m \leq d$ and $\tau_{\mathcal{C}}(m) < 2^m$ for all $m > d$.

Why are we interested in VC-dimension? We have seen earlier that finite concept classes are PAC-learnable (via ERM.) But there are infinite concept classes that are also PAC-learnable, e.g. the concept class consisting of axis aligned rectangles. So if the cardinality does not distinguish learnable concept classes from non-learnable concept classes, then what does? It turns out the its the VC-dimension of $\mathcal{C}$ that determines learnability.

2 Introduction to the fundamental theorem of PAC-learning

For a concept class $\mathcal{C}$ the following are equivalent:

1. $\mathcal{C}$ has the uniform convergence property.
2. Any algorithm which selects a minimizer of the empirical risk achieves PAC-learning on $\mathcal{C}$.
3. $\mathcal{C}$ has finite VC-dimension.

We have already seen (1) $\implies$ (2), and we will soon see (3) $\implies$ (1). For (2) $\implies$ (3), note that we previously discussed a situation (see discussion on No-free-lunch theorem) where $\tau_{\mathcal{C}}(m) = 2^m$ for every $m \in \mathbb{N}$ (i.e. $\mathrm{VC}_{\mathrm{dim}}() = \infty$) and in which the ERM approach broke down because the vast majority of concepts minimizing the empirical risk had true errors which exceeded $\min_{c \in \mathcal{C}} \mathrm{err}(c)$ by a positive constant $\epsilon > 0$.

The proof of (3) $\implies$ (1) has two parts:

1. Sauer's Lemma: if $\mathrm{VC}_{\mathrm{dim}}(\mathcal{C}) \leq d$ then $\tau_{\mathcal{C}}(m) \leq m^d$, and

2. For an i.i.d. sample $(X_1, Y_1), \dots, (X_m, Y_m)$ with any distribution on $\mathcal{X} \times \{0,1\}$

$$\mathbf{E}\left[\sup_{c \in \mathcal{C}} |\widehat{\mathrm{err}}(c) - \mathrm{err}(c)|\right] \approx \sqrt{\frac{\log\left(\tau_{\mathcal{C}}(2m)\right)}{2m}}.$$

As we shall see later, together these will imply that if we take $m \approx d\epsilon^{-2}$ then we will get uniform convergence.

3 Sauer's Lemma

Lemma 3.1 (Sauer-Shelah-Perles)**.** *If $VC_{dim}(\mathcal{C}) \le d < \infty$ then for all $m \in \mathbb{N}$*

$$\tau_{\mathcal{C}}(m) \le \sum_{i=0}^{d} \binom{m}{i}, \tag{1}$$

and

$$\tau_{\mathcal{C}}(m) \le (em/d)^e \tag{2}$$

for all $m > d + 1$.

The Sauer-Shela-Perles lemma has two immediate interesting implications:

- it improves on what we can naïvely obtain from $\mathrm{VC}_{\mathrm{dim}}(\mathcal{C}) \le d$, namely $\tau_{\mathcal{C}}(m) < 2^m$ if $m > d$, and
- as the number of samples in S increases the size of the restriction $\mathcal{C}_S$ grows polynomially in $|\mathcal{S}|$ instead of exponentially in $|\mathcal{S}|$.

Proof of Lemma 3.1. Here we focus on the proof on the inequality at (1), but we note in passing that the bound at (2) maybe be established by the inequality at (1) and induction on d. Now to establish the bound at (1) it suffices to establish

$$|\mathcal{C}_S| \le |\{T \subseteq S : \mathcal{C} \text{ shatters } T\}| \tag{3}$$

for any finite $S \subseteq \mathcal{X}$. To see this note by the definition of VC-dimension that $\mathcal{C}$ does not shatter any S with $|S| > d$ and that S has $\sum_{i=0}^{d} \binom{|S|}{i}$ subsets of size not exceeding d. Now using the bound at (3) it follows that $\tau_{\mathcal{C}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$. We now focus on proving (3) by using an inductive argument on $|S|$. For the base case, i.e. when $|S| = 1$, S has two possible subsets, namely $\emptyset$ and S itself. If $|\mathcal{C}_S| = 2$ then $\emptyset$ (which is trivially always shattered) and S are both shattered, hence we have

$$|\mathcal{C}_S| = 2 = |\{T \subseteq S : \mathcal{C} \text{ shatters } T\}$$

as desired. If $|\mathcal{C}_S| = 1$ then $\emptyset$ is shattered, but S is not shattered. Hence

$$|\mathcal{C}_S| = 1 = |\{T \subseteq S : \mathcal{C} \text{ shatters } T\}$$

as desired. This establishes the base case. Now for the inductive step assume (3) holds for any $S \subseteq \mathcal{X}$ with $|S| < m$. Now consider $S = \{x_1, \ldots, x_m\}$ and let $S' = \{x_2, \ldots, x_m\}$. Define

$$Y_0 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{C}_S \text{ and } (1, y_2, \ldots, y_m) \in \mathcal{C}_S\}$$

and

$$Y_1 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{C}_S \text{ or } (1, y_2, \ldots, y_m) \in \mathcal{C}_S\},$$

and observe that $|\mathcal{C}_S| = |Y_0| + |Y_1$. We will now relate $|Y_0|$ and $|Y_1|$ to the number of subsets of S that $\mathcal{C}$ can shatter. By the induction assumption we have

$$\begin{aligned}|Y_1| = |\mathcal{C}_{S'}| &\leq |\{T \subseteq S' : \mathcal{C} \text{ shatters } T\}| \\ &= |\{T \subseteq S : x_1 \notin T \text{ and } \mathcal{C} \text{ shatters } T\}|.\end{aligned}$$

Now by the defintion of Y_0 we have for every $(y_2, \ldots, y_m) \in Y_0$ that there exists concepts $c_1, c_2 \in \mathcal{C}$ such that $c_1(x_1) = 0$, $c_2(x_1) = 1$, and $c_1(x_i) = c_2(x_i) = y_i$ for $i = 2, \ldots, m$. Let $\mathcal{C}'$ consist of the pairs of such concepts, i.e. c_1 and c_2, as $(y_2, \ldots, y_m)$ ranges over Y_0. Then

$$|Y_0| = |\mathcal{C}'_{S'}| \leq |\{T \subseteq S' : \mathcal{C}' \text{ shatters } T\}|.$$

But by the construction of $\mathcal{C}'$ we have $\mathcal{C}'$ shatters $T \subseteq S'$ implies $\mathcal{C}'$ shatters $\{x_1\} \cup T$. Hence

$$|Y_0| \leq |\{T \subseteq S : x_1 \in T \text{ and } \mathcal{C}' \text{ shatters } T\}| \leq |\{T \subseteq S : x_1 \in T \text{ and } \mathcal{C} \text{ shatters } T\}|.$$

So we have

$$\begin{aligned}|\mathcal{C}_S| &= |Y_0| + |Y_1| \\ &\leq |\{T \subseteq S : x_1 \in T \text{ and } \mathcal{C} \text{ shatters } T\}| + |\{T \subseteq S : x_1 \notin T \text{ and } \mathcal{C} \text{ shatters } T\}| \\ &= |\{T \subseteq S : \mathcal{C} \text{ shatters } T\}|\end{aligned}$$

as desired. □