COS-511: Learning Theory **Spring 2017**

Lecture 3: VC Dimension & The Fundamental Theorem

Lecturer: Roi Livni

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

3.1 Uniform Convergence and VC Dimension Cont.

So far we have defined the notion of learnability and the property of uniform convergence. Recall that a class of functions $\mathcal{C}$ are learnable if, given an IID **finite** sample we can find a hypothesis h that minimizes the *generalization error* w.r.t the best hypothesis in $\mathcal{C}$. (see Def. 1.1, for pac learnability.).

We discussed the notion of an ERM algorithm. An ERM algorithm is an algorithm that chooses a hypothesis which minimizes the *empirical error*. We've seen that finite class are learnable, via an ERM algorithm. The proof relied on showing that by seeing $O(\frac{1}{\epsilon^2}\log\frac{|C|}{\delta})$ example we can estimate the performance of *all* hypotheses in $\mathcal{C}$ uniformly.

The last result motivated the property of *uniform convergence*. A class has the uniform convergence property if, with enough samples, we can estimate uniformly the performance of all hypotheses in the class (see Def. 2.6). It is not hard to see that having the uniform convergence property means that an ERM algorithm will succeed in learning.

3.1.1 VC Dimension

Here we introduce the notion of VC-Dimension of a hypothesis class. The VC-Dimension of a hypothesis class is a strictly *combinatorial* property: namely, it is a property of the hypothesis class, and is completely independent of any distribution D on the domain or on the labels. Never the less we will see that this property of the hypothesis class is the main property that governs the learnability of the hypothesis class.

Definition 3.1. *For a sample S define $\mathcal{H}_S$ to be the restriction of $\mathcal{H}$ to S*

$$\mathcal{H}_S = \{h' : S \to \mathcal{Y} : \ h'(s) = h(s) \text{ for some } h \in \mathcal{H} \text{ and all } s \in S\}$$

Definition 3.2 (Shattered Set)**.** *Given a domain χ and a hypothesis class $\mathcal{H}$, a finite set $A \subseteq \chi$ is said to be shattered if for every subset $A' \subseteq A$ there is $h \in \mathcal{H}$ such that for $x \in A$, $h(x) = 1$ iff $x \in A'$. Formally A is shattered iff*

$$|\mathcal{H}_A| = 2^{|A|}$$

Definition 3.3 (VC dimension)**.** *The VC dimension of a hypothesis class, VC-dim($\mathcal{H}$), is defined as the maximal cardinality of a finite set A that is shattered.*

$$\textit{VC-dim}(\mathcal{H}) = \max\{|A| : \ A \text{ is shattered by } \mathcal{H}\}$$

3.1.1.1 Examples

Example 3.1 (Axis aligned rectangles)**.** *Consider Example. 2.1 of $\mathcal{H}$ consists of all target functions of the form:*

$$f_{z_1,z_2,z_3,z_4}(x_1,x_2) = \begin{cases} 1 & z_1 \le x_1 \le z_2,\ z_3 \le x_2 \le z_4 \\ 0 & \text{else} \end{cases}$$

We will show that VC-dim$(H) = 4$.

First we show that VC-dim$(H) \ge 4$ for that we need to show a set of size 4 that is shattered. Let us denote:

$$e_1 = (1,0) \quad e_2 = (0,1)$$

We wil show that the set $S = \{\pm e_1, \pm e_2\}$ is shattered. Choose any target function h over S, then we need to show that for some $f_{z_1,z_2,z_3,z_4}(\pm e_i) = h(\pm e_i)$. Now if $h(-e_1) = 1$ put $z_1 = -2$, else put $z_1 = 0$. Similarly if $h(e_1) = 1$ put $z_2 = 2$ and else put $z_2 = 0$. Similarly define z_3, z_4. if $h(-e_2) = 1$ put $z_3 = -2$, else put $z_3 = 0$, if $h(e_2) = 1$ put $z_4 = 2$ and else put $z_4 = 0$.

One can then check that $f_{z_1,z_2,z_3,z_4} = h$. Since that is for arbitrary h we've shown that $|\mathcal{H}_S| = 2^4$. In other words, all target functions are realizable.

Next we need to show that VC-dim$(H) < 5$. For that we need to show that *any* set of size 5 cannot be shattered. Choose a set $S = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots \mathbf{x}^{(5)}\}$. Now let A be a set of 4 points that are extermal points (i.e. all points have the maximal or the minimal value in some coordinate). Now since A is a set of just 4 points, there is $\mathbf{x}^{(i)} \notin A$, however if $f_{z_1,z_2,z_3,z_4}(\mathbf{x}) = 1$ for all $\mathbf{x} \in A$ we must have $f_{z_1,z_2,z_3,z_4}(\mathbf{x}^{(i)}) = 1$ also (since, for example $z_2 > \max\{x_1^{(j)}\} \ge x_1^{(i)}$ and so on...). thus we cannot realize $f(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $f(\mathbf{x}) = 0$ if $\mathbf{x} \notin A$.

Example 3.2 (Half Spaces with bias)**.** *Next we consider a class $\mathcal{H}$ similar to Example. 2.2 of halfspaces with bias:*

$$\mathcal{H} = \{f_{\mathbf{w}}(x) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$$

We will prove that VC-dim$(\mathcal{H}) = d + 1$.

First we show that VC-dim$(\mathcal{H}) \ge d + 1$ for that we need to show that there is a set of size $d + 1$ that is shattered.

Similar to before define $e_i = (0, 0, \dots, \underbrace{1}_{\text{i-thcoordinate}}, 0, \dots 0)$ and set $e_0 = 0$ then $S = \{e_0, e_1, \dots, e_d\}$ is a set of size $d+1$. for every function $h(\mathbf{x}) = \pm 1$ set $w_i = h(e_i)$ and set $b = h(e_0) \cdot \frac{1}{2}$. Finaly we set $\mathbf{w} = (w_1, \dots, w_d)$.

$$f_{\mathbf{w},b}(e_i) = \text{sign}(\mathbf{w} \cdot e_i + b) = \text{sign}(w_i + b) = \text{sign}(h(e_i) \pm \frac{1}{2}) = h(e_i)$$

Also for $f_{\mathbf{w},b}(0) = \text{sign}(\mathbf{w} \cdot 0 + b) = \text{sign}(b) = h(e_0)$. Thus, we realized an arbitrary target function over S.

Next we wish to show that VC-dim$(\mathcal{H}) < d + 1$. For the we use Radon's theorem. Recall that the convex hull of a set A, denoted $A^c = \{\sum_{i=1}^t \lambda_i x_i :\ \lambda > 0, \sum \lambda_i = 1, x_i \in A\}$.

Theorem 3.4 (Radon's Theorem)**.** *For every set $S \subseteq \mathbb{R}^d$ of size $d+2$, we can divide S into two disjoint sets whose convex hull intersect.*

Given a set S of size d we divide it into two disjoint set A_1, A_2 whose convex hull intersect. Now we show that we cannot have $f_{\mathbf{w},b}(\mathbf{x}) = 1$ if $\mathbf{x} \in A_1$ and $f_{\mathbf{w},b}(\mathbf{x}) = -1$ if $\mathbf{x} \in A_2$.

Indeed, suppose otherwise and let $a \in A_1^c \cap A_2^c$ then there are positive $\lambda_i^{(1)}, \lambda_i^{(2)}$ who sum to one and

$$a = \sum_{x_i \in A_1} \lambda_i^{(1)} x_i = \sum_{x_i \in A_2} \sum \lambda_i^{(2)} x_i$$

Next we show that $f_{\mathbf{w},b}(a) \geq 0$:

$$\begin{aligned} \mathbf{w} \cdot a + b = \mathbf{w} \cdot \left(\sum_{x_i \in A_1} \lambda_i^{(1)} x_i \right) + b &= \sum_{x_i \in A_1} \lambda_i^{(1)} \mathbf{w} \cdot x_i + \sum \lambda_i^{(1)} \cdot b \\ &= \sum_{x_i \in A_1} \lambda_i^{(1)} \left(\mathbf{w} \cdot x_i + b \right) \geq 0 \end{aligned}$$

Exactly the same way we show that $f_{\mathbf{w},b}(a) < 0$ which is a contradiction.

3.2 The Fundamental Theorem of Statistical Learning Theory

3.2.1 VC = ERM = Learnability

Theorem 3.5 (The Fundamental Theorem of Statistical Learning)**.** *Let $\mathcal{C}$ be a concept class of functions from a domain χ to $\{-1, 1\}$, and let the loss function ℓ be the $0-1$ loss. Then the following are equivalent*

1. *$\mathcal{C}$ is (agnostic) PAC learnable.*
2. *$\mathcal{C}$ is (realizable) PAC learnable.*
3. *$\mathcal{C}$ has finite VC dimension.*
4. *$\mathcal{C}$ has the uniform convergence property*
5. *$\mathcal{C}$ is learnable by a $ERM_{\mathcal{C}}$ algorithm.*

Further, if the VC-dimension of $\mathcal{H}$ is d then the sample complexity of the class (attained by an ERM) algorithm is given by

$$m(\epsilon, \delta) = O(\frac{d}{\epsilon} \log 1/\delta),$$

in the realizable model and

$$m(\epsilon, \delta) = O(\frac{d}{\epsilon^2} \log 1/\delta),$$

in the agnostic setting.

To summarize, the fundamental theorem states that that for the $0-1$ function the VC dimension completely characterizes the learnable classes, and as far as the PAC model goes, ERM algorithms are optimal.

The implications $1 \to 2$, $5 \to 1$ are trivial. The proof that $4 \to 5$ is essentialy the proof we gave for the special case fo finite hypotheses classes. (Cor. 2.5). We next prove $2 \to 3$. In the next lecture we will show $2 \to 3$ and $3 \to 4$.