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

Lecture 4: No Free Lunch & Sauer's Lemma

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.*

In previous lecture we defined the notion of VC-dimension. We stated the fundemental theorem which roughly says that the following are equivalent Learnability = finite VC dimension= Uniform Convergence = learnable through ERM. In this lecture we will prove that learnability implies finite VC dimension and take a first step in prove that finite VC implies uniform convergence. Namely, we will prove Sauer's Lemma.

4.0.1 Infinite VC Dimension implies in-learnability

Suppose that $\mathcal{H}$ has an infinite VC-dimension, and assume it is learnable. Let $2m$ be such that for a sample of size m for every distribution D $|\mathrm{err}(h_S) - \mathrm{err}(h^*)| < \frac{1}{4}$ with probability at least $\frac{1}{8}$.

Let D be a uniform distribution, supported on a set $X = (x_1, \ldots, x_{2m})$ that shatters $\mathcal{H}$, then for every $\mathbf{y} = \{0,1\}^{2m}$ there is $h_\mathbf{y} \in \mathcal{H}$ such that $h_\mathbf{y}(x_i) = y_i$. Suppose we choose a subset $S' = s_1, \ldots, s_m \subseteq X$ of size m and we randomly choose a hypothesis $h_\mathbf{y}$ (where we pick $\mathbf{y}$ uniformly at random) and present to the algortihm a sample $S = (s_1, h_\mathbf{y}(s_1), \ldots, (s_m, h_\mathbf{y}(s_m)))$. The clearly h_S is independent of any labelling of elements outside of S and we obtain that

$$\mathop{\mathbb{E}}_{\mathbf{y}\sim Y}\left[\frac{1}{m}\sum_{x\notin S'} \ell_{0,1}(h_S(x), h_\mathbf{y}(x))|S'\right] = \frac{1}{2}$$

Since $h_{S'}$ is accurate on S' we obtain that

$$\mathop{\mathbb{E}}_{\mathbf{y}\sim Y}\left[\frac{1}{2m}\sum_{i=1}^{m} \ell_{0,1}(h_{S'}(x_i), h_\mathbf{y}(x_i))|S'\right] = \frac{1}{4}$$

Taking expectation over S' and employing Fubini (i.e. $\mathbb{E}_{S'}\mathbb{E}_\mathbf{y} = \mathbb{E}_\mathbf{y}\mathbb{E}_{S'}$) We have that

$$\mathop{\mathbb{E}}_{\mathbf{y}\sim Y}\mathop{\mathbb{E}}_{(x,y)\sim D_\mathbf{y}}\left[\frac{1}{2m}\sum_{i=1}^{m} \ell_{0,1}(h_{S'}(x_i), h_\mathbf{y}(x_i))\right] = \frac{1}{4}$$

Since this is true by expectation, we obtain that for *some* $h_\mathbf{y}$ we have that

$$\mathop{\mathbb{E}}_{(x,y)\sim D_\mathbf{y}}\left[\frac{1}{2m}\sum_{i=1}^{m} \ell_{0,1}(h_{S'}(x_i), h_\mathbf{y}(x_i))\right] \geq \frac{1}{4}$$

By Markov's inequality we obtain that at least with probability $\frac{1}{8}$ we have that:

$$\mathop{\mathbb{E}}_{S\sim D_\mathbf{y}}[\mathrm{err}(h_S)] \geq \frac{1}{8}$$

Corollary 4.1 (No Free Lunch Theorem)**.** *Consider any $m \in \mathbb{N}$, any domain χ of size $|\mathcal{X}| = 2m$, and any algorithm A which outputs a hypothesis $h \in \mathcal{H}$ given a sample S. Then there exists a concept $h : \mathcal{X} \to \{0,1\}$ and a distribution $\mathcal{D}$ such that:*

- *The error* $err(f) = 0$
- *With probability at least* $\frac{1}{10}$, $err(h_S) \geq \frac{1}{10}$.

4.1 Sauer's Lemma

We are left with proving that finite VC dimension implies uniform convergence. As a prerequisite we are going to prove Sauer's lemma. Define the growth function

$$\tau_{\mathcal{H}}(m) = \max_{S \subseteq X, |S|=m} \{|\mathcal{H}_S|\}$$

If $d = \text{VC-dim}(\mathcal{H})$ then we have $\tau_{\mathcal{H}}(m) = 2^m$ for all $m \leq d$, we next prove Sauer's Lemma that $\tau_{\mathcal{H}}(m) = O(m^d)$

Lemma 4.2 (Sauer's Lemma)**.** *Let* $\mathcal{H}$ *be a class with VC-dim*$(H) = d$*, then:*

$$\tau_{\mathcal{H}}(m) \leq \sum_{t=0}^{d} \binom{m}{t} = O(m^d)$$

Proof. We use induction over $m + d$. For the base case, if $m + d = 0$, if $|\mathcal{H} > 1$, there exists $x \in \chi$ and $h_1, h_2 \in \mathcal{H}$ such that $h_1(x) \neq h_2(x)$ and $\{x\}$ is shattered, contradiction to the fact that $d = 0$.

Next, we assume that statement is true for $m + d = k$ and set out to prove it for $m + d = k + 1$. Let $S = \{x_1, \ldots, x_m\}$ be a set of sample such that $\tau_{\mathcal{H}}(m) = |\mathcal{H}_S|$. and for every $h \in \mathcal{H}_{\mathcal{S}}$ let, $h_{|m}$ be the restriction of h to $S/\{x_m\}$. We next define to hypothesis classes:

$$\mathcal{H}_1 = \left\{h_{|m} : h \in \mathcal{H}_S\right\}$$
$$\mathcal{H}_2 = \{h \in \mathcal{H}_S : h(x_m) = 1 \text{ and } \exists h' \in \mathcal{H}_S \text{ s.t.} h'(x_m) = 0\}$$

We first claim that

$$|\mathcal{H}_S| = |\mathcal{H}_1| + |\mathcal{H}_2|$$

Indeed, for any $h \in \mathcal{H}_S$ assume that $h(x_m)$ is unique, i.e. for any $h' \in \mathcal{H}_S$ we have $h(x_m) \neq h'(x_m)$. Then h is counted once by $\mathcal{H}_1$ (it is not in $\mathcal{H}_2$). On the other case, if $h(x_m)$ is not unique, then their common restriction is counted, once by $\mathcal{H}_1$ and h or its counterpart is counted once by $\mathcal{H}_2$.

Next, by definition and induction hypothesis we have that

$$|\mathcal{H}_1| \leq \tau_{\mathcal{H}}(m-1) = \sum_{t=0}^{d} \binom{m-1}{t}.$$

On the other hand, we have that $\text{VC-dim}(\mathcal{H}_2) \leq d-1$. Indeed, if there exists a set $z_1, \ldots, z_d$ that is shattered by $\mathcal{H}_2$, then the set $z_1, \ldots, z_d, x_m$ is also shattered by $\mathcal{H}$ (since for every hypothesis in $\mathcal{H}_2$ there are two hypothesis with different assignments in $\mathcal{H}$ over x_m). Thus, again by induction hypothesis we have that

$$|\mathcal{H}_2| \leq \sum_{t=0}^{d-1} \binom{m-1}{t}.$$

Taken together we have that

$$|\mathcal{H}_S = |\mathcal{H}_1| + |\mathcal{H}_2| \leq 1 + \sum_{t=0}^{d-1} \binom{m-1}{t} + \binom{m-1}{t+1} = 1 + \sum_{t=0}^{d-1} \binom{m}{t+1} = \sum_{t=0}^{d} \binom{m}{t} = O(m^d)$$

□