Lecture 2, October 18, 2016

Machine Learning Theory

No-free-lunch

Ruth Urner

In the last class, we saw that, for binary hypothesis classes, bounded VC-dimension implies learnability (in the sense of Definition 1 in Lecture 2). This lecture will show that the reverse is also true: Binary hypothesis classes are *only learnable* if their VC-dimension is bounded. Most of this lecture will be devoted to the so called no-free-lunch theorem. The equivalence of learnability and bounded VC-dimension will follow as a corollary to the proof of this theorem.

The no-free-lunch result formalizes an important insight in statistical learning, namely, that there is no single learning algorithm that will be successful on all tasks. Or, in other words, for a given machine learning task, it is essential to choose a learner (that may mean choosing a hypothesis class that the learner would output) suitable for the given problem. This is a task that is up to the user of machine learning methods. A user needs to incorporate their knowledge about the application to ensure that, for their application, learning with the chosen hypothesis class can succeed (that is, that there is function in the class with reasonably small loss). The no-free-lunch result formally shows us that this step is inevitable. We should not expect that there is a single "magically" strong learning method that will do well no matter which task we throw at it.

1 There is no free lunch

A bit more formally, the no free lunch says that for every learning algorithm and sample size, there exists a task (that is, a distribution) on which this learner will fail when given samples of the specified size.

Theorem 1 (No-free-lunch)**.** *Let $\mathcal{X}$ be an infinite domain and let $\mathcal{A}$ be a learner. Let m be some sample size. Then, there exists a distribution P over $\mathcal{X} \times \{0,1\}$ such that*

- *there is a function $f : \mathcal{X} \to \{0,1\}$ such that $L_P(f) = 0$*
- $\mathbb{P}_{S^m}[L_P(\mathcal{A}(S)) \geq 1/8] > 1/7$.

Let us contrast this (rather pessimistic) statement with the notion of learnability, which we know is achievable for classes of bounded VC-dimension. For a class of bounded VC-dimension $\mathcal{H}$, we know that there is a learning algorithm (for example any ERM learning algorithm) such that, for every error parameter ϵ and confidence parameter δ, there is a sample size $m(\epsilon, \delta)$, such that, when the learner sees samples of at least this size, it will output a function of loss at most $\inf_{h \in H} L_P(h) + \epsilon$ (with high probability $> 1 - \delta$).

It is important to note that this is only a statement about how close the learner gets to $\inf_{h \in H} L_P(h)$ (a quantity which is also called the *approximation error* of the class). This quantity can still be rather large for some P.

Now the no-free-lunch result tells us that we can not get around this quantity and get a stronger guarantee. It is not possible, for any learner, to guarantee small total error (say, $L_P(\mathcal{A}(S)) \leq \epsilon$) for arbitrarily small ϵ from large sample sizes on, if we would like these sample sizes to depend only on ϵ and δ, but not on the distribution (since the data generation is something that we do not control). The above theorem tells us, for any learner and any sample size there are values for ϵ and δ, namely $1/8$ and $1/7$ respectively, and there is a distribution such that with probability at least $1/7$ over samples *i.i.d.* from P, we have $L_P(\mathcal{A}(S)) \geq 1/8$.

Proof of the no-free-lunch theorem. Let $\mathcal{A}$ be a learner and let m be a sample size. Let $C \subseteq \mathcal{X}$ with $|C| = 2m$. There are $T = 2^m$ functions from C to $\{0, 1\}$. We denote them by $f_1, \ldots, f_T$. For each f_i, we define a distribution P_i over $C \times \{0, 1\}$ as follows:

$$P_i((x, y)) \begin{cases} \frac{1}{|C|} \text{ if } y = f_i(x) \\ 0 \text{ else} \end{cases}$$

Now we have, by construction, $L_{P_i}(f_i) = 0$. That is, for these distributions, the first requirement in the theorem is fulfilled. We will show that one of these distributions will actually force large loss on $\mathcal{A}$, as in the second part of the statement. We will show that

$$\max_{i \in [T]} \mathbb{E}_{S \sim P_i^m}[L_{P_i}(\mathcal{A}(S))] \geq 1/4 \tag{1}$$

Note that, stating that the max in the above is larger than $1/4$ is equivalent to saying that "there exists an $i \in [T]$" such that the quantity is larger than $1/4$. It is easy to show that the above statement then implies the statement of the theorem.

By an application of Markov's inequality (see Lemma B.1 in the Book [1]), for random variables taking values in $[0, 1]$, we have that for all $a \in (0, 1)$

$$\mathbb{P}[Z > a] \geq \frac{\mathbb{E}[Z]}{1 - a}$$

Thus, Equation (1) implies that for there exists an i such that

$$\mathbb{P}_{S \sim P_i^m}[L_{P_i}(\mathcal{A}(S)) > 1/8] \geq \frac{1/4 - 1/8}{1 - 1/8} = \frac{1}{7}.$$

We now proceed to proving Equation (1). There are $k = (2m)^m$ possible sequences of m examples from C. Denote these sequences by $S_1, \ldots, S_k$. Also, if $S_j = (x_1, \ldots, x_m)$ we denote by S_j^i the sequence containing the instances in S_j labeled by the function f_i, namely, $S_j^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m)))$.

For distribution P_i, the possible samples are now $S_1^i \ldots S_k^i$. Note that these samples all have the sample probability to occur. Thus,

$$\mathbb{E}_{S \sim P_i^m}[L_{P_i}(\mathcal{A}(S))] \;=\; \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}_{P_i}(\mathcal{A}(S_j^i)) \;. \tag{2}$$

We get

$$\begin{aligned}\max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{P_i}(\mathcal{A}(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{P_i}(\mathcal{A}(S_j^i)) \\ &= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{P_i}(\mathcal{A}(S_j^i)) \\ &\geq \min_{j\in[k]}\frac{1}{T}\sum_{i=1}^{T} L_{P_i}(\mathcal{A}(S_j^i))\ , \end{aligned} \tag{3}$$

where we used that "maximum" is larger than "average" and that "average" is larger than "minimum". Now, fix some $j \in [k]$. Let $S_j = \{x_1, \ldots, x_m\}$ and $\{v_1, \ldots v_p\} = C \setminus S_j$. Note that $p \geq m$. Therefore, for every function $h : C \to \{0, 1\}$ and every i we have

$$\begin{aligned} L_{P_i}(h) &= \frac{1}{2m}\sum_{x\in C}\mathbf{1}[h(x) \neq f_i(x)] \\ &\geq \frac{1}{2m}\sum_{r=1}^{p}\mathbf{1}[h(v_r) \neq f_i(v_r)] \\ &\geq \frac{1}{2p}\sum_{r=1}^{p}\mathbf{1}[h(v_r) \neq f_i(v_r)]\ . \end{aligned} \tag{4}$$

Thus,

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T} L_{P_i}(\mathcal{A}(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}\mathbf{1}\big[\mathcal{A}(S_j^i)(v_r) \neq f_i(v_r)\big] \\ &= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}\big[\mathcal{A}(S_j^i)(v_r) \neq f_i(v_r)\big] \\ &\geq \frac{1}{2}\cdot\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}\big[\mathcal{A}(S_j^i)(v_r) \neq f_i(v_r)\big]\ . \end{aligned} \tag{5}$$

Next, we fix some $r \in [p]$. We can partition all the functions in $f_1, \ldots, f_T$ into $T/2$ disjoint pairs, where for a pair $(f_i, f_{i'})$ differs exactly on v_r. Since for such a pair we must have $S_j^i = S_j^{i'}$, we get

$$\mathbf{1}\big[\mathcal{A}(S_j^i)(v_r) \neq f_i(v_r)\big] + \mathbf{1}\Big[\mathcal{A}(S_j^{i'})(v_r) \neq f_{i'}(v_r)\Big] = 1\ .$$

This yields

$$\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}\big[\mathcal{A}(S_j^i)(v_r) \neq f_i(v_r)\big] = \frac{1}{2}\ .$$

Overall, we get

$$\begin{aligned}\max_{i\in[T]} \mathbb{E}_{S\sim P_i^m}[L_{P_i}(\mathcal{A}(S))] &\geq \max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{P_i}(\mathcal{A}(S_j^i)) \\ &\geq \min_{j\in[k]} \frac{1}{k}\sum_{j=1}^{k} L_{P_i}(\mathcal{A}(S_j^i)) \\ &\geq \min_{j\in[k]} \frac{1}{2} \min_{r\in[p]} \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\big[\mathcal{A}(S_j^i)(v_r) \neq f_i(v_r)\big] = \frac{1}{4}.\end{aligned}$$

□

Let us see how the construction in the above proof shows that classes $\mathcal{H}$ of infinite VC-dimension are not learnable. We need to show that for any learner there exist ϵ and δ such that for all m there is a P such that

$$\mathbb{P}[L_P(\mathcal{A}(S)) > \inf_{h\in H} L_P(h) + \epsilon] > \delta.$$

This statement is simply the contrapositive of the learnability requirement. To show this statement, we repeat the construction in the proof of the no-free-lunch theorem. We choose C to be $2m$ points that are shattered by $\mathcal{H}$. Since $\mathcal{H}$ has infinite VC-dimension, we can always find such a set C. Note that for the distributions that we construct in the proof, we now have $\inf_{h\in H} L_P(h) = 0$, since $\mathcal{H}$ shatters these points. Thus, choosing $\epsilon = 1/8$ and $\delta = 1/7$ yields the non-learnability claim.

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning.* Cambridge University Press, 2014.