Introduction to Machine Learning (67577)

Recitation 8
PAC II - The Fundamental Theorem of Statistical Learning

Second Semester, 2021

Contents

1 The Fundamental Theorem of Statistical Learning

Theorem 1.1 — The Fundamental Theorem of Statistical Learning. Let $\mathcal{H}$ be a hypothesis class of binary classifiers with VC-Dimension $d \leq \infty$. Then, $\mathcal{H}$ is PAC-learnable if and only if it is Agnostic-PAC learnable if and only if $d < \infty$. Moreover, if $d < \infty$ there there are absolute constants C_1, C_2 such that:

1. $\mathcal{H}$ is PAC learnable with sample complexity
$$C_1 \frac{d + \log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}$$
2. $\mathcal{H}$ is Agnostic-PAC learnable with sample complexity
$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}.$$
3. The upper bound on sample complexity is achieved by the ERM learner.

2 Agnostic PAC

The PAC Learnability Framework discussed previously gives us great power in determining **what** is learnable and **what** we need in order to PAC learn. However, in many real-world applications this framework is limitted for the following reasons:

1. **Noisy Responses**: The PAC framework assumes a probabiltiy distribution $\mathcal{D}$ over $\mathcal{X}$. In practice, we often observe some variability (noisiness) in the responses. This means that if we sample a specific $\mathbf{x} \in \mathcal{X}$ several times we might end up observing it sometimes as $(\mathbf{x}, +1)$ and sometimes as $(\mathbf{x}, -1)$
2. **Relax realizability**:As we do not know the true nature of our data (i.e. the true probability distribution $\mathcal{D}$) our choice of the hypothesis class might not enable us to include the true labeling function in it. For example, suppose we fitt a Decision Tree classifier to a dataset and limit its depth to be 5, while in reality the depth might by greater, or that the data was not even generated by a decision tree.
3. **Limited Loss**: The PAC framework is defined over the mis-classification loss. Often we would like to define other loss functions.

To expand our framework and lift the above limitations we define the Agnostic PAC Framework.

Definition 2.1 — Agnostic PAC learnability. A hypothesis class $\mathcal{H}$ is **Agnostic** PAC learnable with respect to loss $\ell : (\mathcal{X} \times \mathcal{Y}) \rightarrow [0, \infty)$ if there exists a function $m_{\mathcal{H}} : (0,1)^2 \rightarrow \mathbb{N}$ and a learning algorithm $\mathcal{A}$ such that:
- For every $\varepsilon, \delta \in (0,1)$
- For every distribution $\mathcal{D}$ **over** $\mathcal{X} \times \mathcal{Y}$

when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ *i.i.d* samples generated by $\mathcal{D}$, the algorithm returns a hypothesis $h_S := \mathcal{A}(S)$ such that:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(h_S) \leq \min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon\right] \geq 1-\delta.$$

Notice that if we would to assume realizability then the term $\min_{h'\in\mathcal{H}} L_{\mathcal{D}}(h')$ would be equal to zero. As we do not assume so we require that the learning algorithm will approximate the minimal **possible** loss by ε.

Exercise 2.1 Suppose we are given a noisy training set of m samples over $\mathbb{R}^d \times \{0,1\}$ and we use use the Perceptron algorithm to find a homogeneous halfspace that minimizes the empirical risk this data. How does the generalization error change decrease if we increase the size of the training set, assuming a constant confidence?

Solution: First, since we are dealing with noisy data we should use the Agnostic-PAC framework. Then, as we have previously seen, the VC-Dimension of homogeneous halfspaces over $\mathbb{R}^d$ is d. Let h_m be the hypothesis that our algorithm returned using a sample of size m ($i.i.d$ from unknown distribution $\mathcal{D}$). We want to describe $L_{\mathcal{D}}(h_m)$ as a function of m. We start with the following intuition: if we increase the sample size our generalization error should decrease, as from the law of large numbers we are expected to get closer towards $\mathcal{D}$.

Since our hypothesis class is PAC learnable it is also Agnostic-PAC learnable so:

$$L_{\mathcal{D}}(h_m) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$$

Using the Fundamental Theorem (1) we can bound the sample complexity by:

$$C_1 \frac{d+\log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon,\delta) \leq C_2 \frac{d+\log(1/\delta)}{\varepsilon^2}$$

Since we assume a specific constant confidence denote $C=\delta$ and therefore:

$$m \leq C_2 \frac{d+\log(1/C)}{\varepsilon^2} \implies m \leq \mathcal{O}\left(\frac{d}{\varepsilon^2}\right) \implies \varepsilon \leq \mathcal{O}\left(\sqrt{\frac{d}{m}}\right)$$

Put together, we we that we are able to upper bound the generalization error by:

$$L_{\mathcal{D}}(h_m) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \mathcal{O}\left(\sqrt{\frac{d}{m}}\right)$$

where the error decreases at the rate of $\sqrt{m}$. ■

3 Uniform Convergence

In order to show that Agnostic-PAC learnability implies PAC learnability, we need to first define the concept of Uniform Convergence (UC).

Definition 3.1 Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be a set of samples. S is called ε-representative for $\mathcal{D}, \mathcal{H}, \ell$ if

$$\forall h \in \mathcal{H} \quad |L_S(h) - L_\mathcal{D}(h)| < \varepsilon$$

Intuitively, this definition states that, no matter which hypothesis one chooses, the empirical- and generalization errors would be the same up to an ε difference.

Exercise 3.1 Let S be an $\varepsilon/2$-representative sample for $\mathcal{D}, \mathcal{H}, \ell$. Let h_S be any output of $ERM_\mathcal{H}(S)$, namely, $h_S \in argmin_{h \in \mathcal{H}} L_S(h)$. Then

$$L_\mathcal{D}(h_S) \leq \min_{h \in \mathcal{H}} L_\mathcal{D}(h) + \varepsilon$$

Solution: Let $h^* := argmin_{h \in \mathcal{H}} L_\mathcal{D}(h)$. Using the definition of h_S and 3.1 it holds that:

$$L_\mathcal{D}(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h^*) + \frac{\varepsilon}{2} \leq L_\mathcal{D}(h^*) + \varepsilon$$

■

Definition 3.2 Let $\mathcal{H}$ be a hypothesis class. We say that $\mathcal{H}$ has the **Uniform Convergence** property if there exists $m_\mathcal{H}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$:

$$\mathcal{D}^m(\{S \in (\mathcal{X} \times \mathcal{Y})^m \,|\, S \text{ is } \varepsilon\text{-representative}\}) \geq 1 - \delta$$

Put together, definitions 3.1, 3.2, mean that the hypothesis class has the Uniform Convergence properly if (for a sufficiently large m) is it "easy" enough to sample "good" representative datasets.

Exercise 3.2 Let $\mathcal{H}$ by a hypothesis class with the uniform convergence property ad $m_\mathcal{H}^{UC} : (0,1)^2 \to \mathbb{N}$. Show that $\mathcal{H}$ is Agnostic-PAC learnable with sample complexity $m_\mathcal{H}(\varepsilon, \delta) \leq m_\mathcal{H}^{UC}(\varepsilon/2, \delta)$.

Solution: To show that $\mathcal{H}$ is agnostic-PAC learnable we will show that for any $\mathcal{D}, \varepsilon, \delta$, given $m \geq m_\mathcal{H}^{UC}(\varepsilon/2, \delta)$ we have that

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[L_\mathcal{D}(h_S) \leq \min_{h' \in \mathcal{H}} L_\mathcal{D}(h') + \varepsilon\right] \geq 1 - \delta$$

Let $m \geq m_\mathcal{H}^{UC}(\varepsilon/2, \delta)$. As shown above (3.1), if S is $\varepsilon/2$-representative then

$$L_\mathcal{D}(h_S) \leq \min_{h \in \mathcal{H}} L_\mathcal{D}(h) + \varepsilon$$

From the uniform convergence property conclude that

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[L_\mathcal{D}(h_S) \leq \min_{h' \in \mathcal{H}} L_\mathcal{D}(h') + \varepsilon\right] \geq \mathcal{D}^m\left(\left\{S \in (\mathcal{X} \times \mathcal{Y})^m \,|\, S \text{ is } \frac{\varepsilon}{2}\text{-representative}\right\}\right) \geq 1 - \delta$$

■