Chapter 1

Learning via uniform convergence

In this Chapter we will develop a more general tool, namely the *uniform convergence*, and apply it to show that any finite class is learnable in the agnostic PAC model with general loss functions, as long as the range loss function is bounded.

1.1 Uniform convergence is sufficient for learnability

Given a hypothesis class $\mathcal{H}$, the ERM learning algorithm:

- Receives a training sample S.
- The learner evaluates the risk of each $h \in \mathcal{H}$ on the given sample S.
- The learner outputs a member h^* of $\mathcal{H}$ that minimizes this empirical risk.
- The hope is that a h that minimizes the empirical risk with respect to S is a risk minimizer with respect to the true data probability distribution as well. For that, it sufficies to ensure that the empirical risks of all members of $\mathcal{H}$ are good approximations of their true risk.

In other words, we need that uniformly over all hypoteses in the hypotesis class, the empirical risk will be close to the true risk. This is formalized in the following definition. Then, a lemma introduces us to a first result.

Definition 1.1.1 (ε-representative sample)**.** A training set S is called ε-representative (with respect to domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if:

$$\forall h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon \tag{1.1}$$

Lemma 1.1.1. *Assume that a training set S is $\frac{\varepsilon}{2}$-representative (with respect to domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$). Then any output of $ERM_{\mathcal{H}}(S)$, namely, any $h_S \in \operatorname{argmin}_{h\in\mathcal{H}} L_S(h)$, satisfies:*

$$L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon \tag{1.2}$$

Proof. For every $h \in \mathcal{H}$:

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

□

The consequence of this lemma is that, if with probability $1 - \delta$, a random training set S is ε-representative, then the ERM rule is an agnostic PAC learner. The uniform convergence condition formalizes this requirement.

Definition 1.1.2 (Uniform convergence)**.** A hypothesis calss $\mathcal{H}$ has the *uniform convergence property* (with respect to a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1-\delta$, S is ε-representative.

Corollary 1.1.1.1. *If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$, then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\varepsilon}{2}, \delta)$. Furthermore, in that case, the $ERM_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.*

1.2 Finite classes are agnostic PAC learnable

Let's start with a lemma that will be useful to give a proof of a subsequent theorem.

Lemma 1.2.1 (Hoeffding's inequality)**.** *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that $\forall i,\ \mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then for any $\varepsilon > 0$:*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right] \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}} \tag{1.3}$$

Theorem 1.2.2. *Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0,1]$ be a loss function. Then:*

- *$\mathcal{H}$ enjoys the uniform convergence property with sample complexity:*

$$m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq \left\lceil \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\varepsilon^2} \right\rceil \tag{1.4}$$

- *$\mathcal{H}$ is agnostically PAC learnable using the ERM algorithm with sample complexity:*

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2}, \delta\right) \leq \left\lceil \frac{2\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{\varepsilon^2} \right\rceil \tag{1.5}$$

Proof. Let's give the proof of the previous theorem by steps.

▷ $\mathcal{D}^m(\{S : \forall h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon\}) \geq 1 - \delta$

▷ $\mathbb{P}_{\text{bad}} = \mathcal{D}^m(\{s : \exists h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq \delta$

▷ $\{S : \exists h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} = \bigcup_{h \in \mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}$

▷ Union bound:
$\mathcal{D}^m(\{S : \exists h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\})$

▷ We now demonstrate that for any fixed h, the difference $|L_S(h) - L_{\mathcal{D}}(h)|$ is small enough.

▷ Recall that:

$$L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(h,z)]$$
$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)$$

By the linearity of expectation, it follows that $L_{\mathcal{D}}(h)$ is also the expected value of $L_S(h)$. Hence, the quantity $|L_S(h) - L_{\mathcal{D}}(h)|$ is the deviation of the random variable $L_S(h)$ from its expectation.
We need to show that the measure of $L_S(h)$ is concentrated around its expectation value.

▷ Law of large numbers: if m is large, the average converges to the expectation.

▷ From Lemma 1.2.1 it follows:

$$\mathcal{D}^m(\{S \ : \ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) = \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right] \leq 2e^{-2m\varepsilon^2}$$

▷ $\displaystyle\sum_{h\in\mathcal{H}}\mathcal{D}^m(\{S \ : \ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq |\mathcal{H}|2e^{-2m\varepsilon^2} = \sum_{h\in\mathcal{H}}2e^{-2m\varepsilon^2}$

▷ We impose:

$$|\mathcal{H}|2e^{-2m\varepsilon^2} \overset{!}{\leq} \delta \Longrightarrow m \geq \log\left(\frac{2|\mathcal{H}|}{\delta}\right)\frac{1}{2\varepsilon^2}$$
$$\Longrightarrow \sum_{h\in\mathcal{H}}\mathcal{D}^m(\{S \ : \ |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq \delta$$

□

Remark 1 (The discretization trick). There is a simple trick that allows us to get a good estimate of the practical sample complexity of infinite hypothesis class:

- Consider a hypothesis class determined by d real number parameters.
- In principle, we have a hypothesis class of infinite size.
- In practice, real numbers are represented with 64 bits double precision variables.
- For d parameters: $|\mathcal{H}| = 2^{64d}$, so $|\mathcal{H}|$ is larger but finite.
- We have:
$$m_{\mathcal{H}}(\varepsilon,\delta) \leq m_{\mathcal{H}}^{\mathrm{UC}}\left(\frac{\varepsilon}{2}\right) \leq \frac{2\log\left(2\frac{2^{64d}}{\delta}\right)}{\varepsilon^2}$$
- So the bound depends on the chosen number representation.