Learning via Uniform Convergence

Prof. Dan A. Simovici

UMB

Reminder

- For agnostic learning the generalization error is:

$$L_{\mathcal{D}}(h) = \mathcal{D}(\{(x, y) \mid h(x) \neq y\}).$$

- The empirical risk is:

$$L_S(h) = \frac{|\{i \mid h(x_i) \neq y_i \text{ for } 1 \leqslant i \leqslant m\}|}{m}.$$

Definition

Let $\mathcal{H}$ be a hypothesis class, and let $\mathcal{D}$ be a distribution. A training set S is ϵ-representative with respect to the above elements, if the absolute value of the difference between the empirical risk and the generalization error is less than ϵ,

$$|L_S(h) - L_{\mathcal{D}}(h)| \leqslant \epsilon.$$

for all $h \in \mathcal{H}$.

Equivalently,

$$L_S(h) - \epsilon \leqslant L_{\mathcal{D}}(h) \leqslant L_S(h) + \epsilon.$$

Definition

Let $\mathcal{H}$ be a class of hypotheses. A ERM predictor for $\mathcal{H}$ is a hypothsis g such that its empirical risk $L_S(g)$ is minimal, that is, $L_S(g) \leqslant L_S(h)$ for every sample S and hypothesis $h \in \mathcal{H}$.

The next lemma stipulates that when the sample is $\frac{\epsilon}{2}$-representative, the ERM learning rule applied to a sample S is guaranteed to return a good hypothesis h_S.

Lemma

Assume that a training set S is $\frac{\epsilon}{2}$-representative, that is, $|L_S(h) - L_{\mathcal{D}}(h)| \leqslant \frac{\epsilon}{2}$. Then, any h_S that minimizes the empirical risk

$$h_S \in argmin_{h \in \mathcal{H}} L_S(h)$$

satisfies

$$L_{\mathcal{D}}(h_S) \leqslant \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

Proof

For every $h \in \mathcal{H}$ we have

$$
\begin{aligned}
L_{\mathcal{D}}(h_S) &\leqslant L_S(h_S) + \frac{\epsilon}{2} \\
&\quad \text{(by the } \tfrac{\epsilon}{2}\text{-representativeness of } S \text{ to } h_S) \\
&\leqslant L_S(h) + \frac{\epsilon}{2} \\
&\quad \text{(because } h_S \text{ is an ERM predictor, hence } L_S(h_S) \leqslant L_S(h)) \\
&\leqslant L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&\quad \text{(because } S \text{ is } \tfrac{\epsilon}{2}\text{-representative, so } L_S(h) \leqslant L_{\mathcal{D}}(h) + \tfrac{\epsilon}{2}) \\
&\leqslant L_{\mathcal{D}}(h) + \epsilon.
\end{aligned}
$$

Thus, to ensure that the ERM rule is an agnostic PAC learner, it suffices to show that with probability of at least $1 - \delta$ over the random choice of a training set, it will be an ϵ-represeentative training set.

Generalized Loss Functions

Definition

Given a hypothesis set $\mathcal{H}$ and some domain Z let $\ell : \mathcal{H} \times Z \longrightarrow \mathbb{R}_{\geqslant 0}$ be a *loss function*.
The risk function is the expected loss of a classifier $h \in \mathcal{H}$ given by

$$L_{\mathcal{D}}(h) = E_{z \sim D}[\ell(h, z)].$$

The empirical risk for $S = \{s_1, \ldots, s_m\}$ is

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, s_i).$$

Definition

A hypothesis class $\mathcal{H}$ has the uniform convergence property (relative to a domain Z and a loss function ℓ) if there exists a function $m^{\mathsf{UC}} : (0,1)^2 \longrightarrow \mathbb{N}$ (the same for all hypotheses in $\mathcal{H}$ and all probability distributions $\mathcal{D}$) such that for every $\epsilon, \delta \in (0,1)$ if S is a sample of size m, where $m \geqslant m^{\mathsf{UC}}(\epsilon, \delta)$, then with probability at least $1 - \delta$, S is ϵ-representative.

The term *uniform* refers to the fact that $m^{\mathsf{UC}}(\epsilon, \delta)$ is the same for all hypotheses in $\mathcal{H}$ and all probability distributions $\mathcal{D}$.

REMINDER: Agnostic PAC Learning

- The realizability assumption (the existence of a hypothesis $h^* \in \mathcal{H}$ such that $P_{x \sim \mathcal{D}}(h^*(x) = f(x)) = 1$) is not realistic in many cases.
- Agnostic learning replaces the realizability assumption and the targeted labeling function f, with a distribution $\mathcal{D}$ defined on pairs (data, labels), that is, with a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$.
- Since $\mathcal{D}$ is defined over $\mathcal{X} \times \mathcal{Y}$, the the generalization error is

$$L_{\mathcal{D}}(h) = \mathcal{D}(\{(x, y) \mid h(x) \neq y\}).$$

Theorem

If a class $\mathcal{H}$ has the uniform convergence property with a function m^{UC}, then the class $\mathcal{H}$ is agnostically PAC learnable with the sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leqslant m^{UC}(\epsilon/2, \delta).$$

Furthermore, in this case, the $ERM_{\mathcal{H}}$ paradigm is a successful agnostic learner for $\mathcal{H}$.

Proof

Suppose that $\mathcal{H}$ has the uniform convergence property with a function m^{UC}.
For every $\epsilon, \delta \in (0, 1)$ if S is a sample of size m, where $m \geqslant m^{\mathrm{UC}}(\epsilon/2, \delta)$, then with probability at least $1 - \delta$, S is $\epsilon/2$-representative, which means that for all $h \in \mathcal{H}$ we have:

$$L_{\mathcal{D}}(h) \leqslant L_S(h) + \epsilon/2.$$

By the definition of h_S we have:

$$\begin{aligned} L_{\mathcal{D}}(h_S) &\leqslant \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon/2 \\ &\leqslant \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon, \end{aligned}$$

hence $\mathcal{H}$ is agnostically PAC-learnable with $m_{\mathcal{H}}(\epsilon, \delta) = m^{\mathrm{UC}}(\epsilon/2, \delta)$.

Theorem

Uniform convergence holds for a finite hypothesis class.

Proof: Fix $\epsilon, \delta \in (0, 1)$.

- We need a sample $S = \{s_1, \ldots, s_m\}$ of size m that guarantees that for any $\mathcal{D}$ with probability at least $1 - \delta$ we have that

$$|L_S(h) - L_{\mathcal{D}}(h)| \leqslant \epsilon$$

  for all $h \in \mathcal{H}$ (that is, S is a representative sample).

- Equivalently,

$$\mathcal{D}^m(\{S \mid \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta.$$

- Note that:

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

This implies

$$\begin{aligned}&\mathcal{D}^m(\{S \mid \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\ &= \sum_{h\in\mathcal{H}} \mathcal{D}^m(\{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}).\end{aligned}$$

Next phase: each term of the right side of previous inequality $\sum_{h\in\mathcal{H}} \mathcal{D}^m(\{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$ is small enough (for large m).

- Let θ_i be the random variable $\theta_i = \ell(h, s_i)$. Since h is fixed and and $s_1, \ldots, s_m$ are iid random variables, it follows that $\theta_1, \ldots, \theta_m$ are also iid random variables.
- $E(\theta_1) = \cdots = E(\theta_m) = \mu$.
- Range of ℓ is $[0, 1]$ and therefore, the range of θ_i is $[0, 1]$.
- Each term $\mathcal{D}^m(\{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$ is small enough for large m.
- We have:

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \theta_i \text{ and } L_{\mathcal{D}}(h) = \mu.$$

By Hoeffding's Inequality,

$$\begin{aligned}
&\mathcal{D}^m(\{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\
&= \quad P\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right) \\
&\leqslant \quad \sum_{h\in\mathcal{H}} 2e^{-2m\epsilon^2} \\
&\leqslant \quad 2|\mathcal{H}|e^{-2m\epsilon^2}.
\end{aligned}$$

To have $\mathcal{D}^m(\{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leqslant \delta$ we need $2|\mathcal{H}|e^{-2m\epsilon^2} \leqslant \delta$, which is equivalent to

$$m \geqslant \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}.$$

A Corollary

Recall that the ERM algorithm returns a hypothesis h such that for which $L_S(h)$ is minimal.

Corollary

Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and $\ell : \mathcal{H} \times Z \longrightarrow [0,1]$ be a loss function. Then $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) = \left\lceil \frac{\log \frac{2|\mathcal{H}|}{\delta}}{2\epsilon^2} \right\rceil.$$

Furthermore, the class is agnostically PAC learnable using the ERM algorithm with sample complexity;

$$m_{\mathcal{H}}(\epsilon, \delta) \leqslant m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leqslant \left\lceil \frac{2 \log \frac{2|\mathcal{H}|}{\delta}}{\epsilon^2} \right\rceil.$$