Learning via Uniform Convergence

Prof. Dan A. Simovici

UMB
1. Uniform Convergence

2. Finite Classes are Agostically PAC-learnable
**Definition**

Let

- $\mathcal{H}$ be a hypothesis class,
- $Z$ a domain,
- $\ell$ a loss function,
- and $\mathcal{D}$ be a distribution.

A training set $S$ is $\epsilon$-representative with respect to the above elements, if for all $h \in \mathcal{H}$ we have:

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

Equivalently,

$$L_S(h) - \epsilon \leq L_\mathcal{D}(h) \leq L_S(h) + \epsilon.$$
Recall that $L_D(h) = E_{z \sim D}(\ell(h, z))$.
The next lemma stipulates that when the sample is $\frac{\epsilon}{2}$-representative, the ERM learning rule is guaranteed to return a good hypothesis.

**Lemma**

Assume that a training set $S$ is $\frac{\epsilon}{2}$-representative. Then, any output $h_S$ of $\text{ERM}_{\mathcal{H}}(S)$, namely, $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$ satisfies

$$L_D(h_S) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon.$$
Proof

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

$$L_D(h_S) \leq L_S(h_S) + \frac{\epsilon}{2}$$

(apply the $\frac{\epsilon}{2}$-representativeness of $S$ to $h_S$)

$$\leq L_S(h) + \frac{\epsilon}{2}$$

(because $h_S$ is an ERM predictor, hence $L_S(h_S) \leq L_S(h)$)

$$\leq L_D(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2}$$

(because $S$ is $\frac{\epsilon}{2}$-representative, so $L_S(h) \leq L_D(h) + \frac{\epsilon}{2}$)

$$\leq L_D(h) + \epsilon.$$
Definition

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

The term uniform refers to the fact that $m^{\text{UC}}(\epsilon, \delta)$ is the same for all hypotheses in $\mathcal{H}$ and all probability distributions $D$. 
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 $m_{\mathcal{H}}(\epsilon, \delta) \leq m^{UC}(\epsilon/2, \delta)$. 
Proof

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

\[
L_D(h) \leq L_S(h) + \epsilon/2,
\]

or

\[
L_D(h) \leq \min_{h' \in \mathcal{H}} L_D(h') + \epsilon/2 \leq \min_{h' \in \mathcal{H}} L_D(h') + \epsilon,
\]

hence \( \mathcal{H} \) is agnostically PAC-learnable with \( m_\mathcal{H}(\epsilon, \delta) = m^{\text{UC}}(\epsilon/2, \delta) \).
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 $D$ with probability at least $1 - \delta$ we have that for all $h \in \mathcal{H}$, $|L_S(h) - L_D(h)| < \epsilon$.
- Equivalently,

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

- Note that

$$\{S \mid \exists h \in \mathcal{G}, |L_S(h) - L_D(h)| > \epsilon\} = \bigcup_{h \in \mathcal{H}} \{S \mid |L_S(h) - L_D(h)| > \epsilon\}.$$
This implies

$$\mathcal{D}^m(\{S \mid \exists h \in \mathcal{G}, |L_S(h) - L_D(h)| > \epsilon\}) = \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S \mid |L_S(h) - L_D(h)| > \epsilon\})$$
Next phase:

- Let $\theta_i$ be the random variable $\theta_i = \ell(h, z_i)$. Since $h$ is fixed and and $z_1, \ldots, Z_m$ 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 $\ell$ is $[0, 1]$ and therefore, the range of $\theta_i$ is $[0, 1]$.
- Each term $\mathcal{D}^m(\{S \mid |L_S(h) - L_D(h)| > \epsilon\})$ is small enough for large $m$.
- We have:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \theta_i \text{ and } L_D(h) = \mu.$$
Finite Classes are Agostically PAC-learnable

By Hoeffding’s Inequality,

\[
\mathcal{D}^m(\{S \mid |L_S(h) - L_D(h)| > \epsilon\})
= P \left( \left| \frac{1}{m} \sum_{i=1}^{m} \theta_i - \mu \right| > \epsilon \right)
\leq \sum_{h \in \mathcal{H}} 2e^{-2m\epsilon^2}
\leq 2|\mathcal{H}|2e^{-2m\epsilon^2}.
\]

If we choose \( m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \), then

\[
\mathcal{D}^m(\{S \mid \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon\}) \leq \delta.
\]
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 $\ell : \mathcal{H} \times Z \rightarrow [0, 1]$ be a loss function. Then $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

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

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

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