# Nonuniform Learnability

## Κεφάλαιο 7

Ιωσήφ Μουλίνος

7 Δεκεμβρίου 2018

# $(\epsilon, \delta)$ competitiveness - Nonuniform Learnability

We say that a hypothesis $h$ is $(\epsilon, \delta)$-competitive with another hypothesis $h'$ if, with probability higher than $(1 - \delta)$,

$$L_{\mathcal{D}}(h) \leq L_{\mathcal{D}}(h') + \epsilon$$

Nonuniform Learnability: Allow the sample size to depend on the hypothesis to which the learner is compared.

## Nonuniform learnability

A hypothesis class $\mathcal{H}$ is nonuniformly learnable if there exist a learning algorithm, A, and a function $m_{\mathcal{H}}^{NUL} : (0.1)^2 \times \mathcal{H} \to \mathbb{N}$ such that, for every $\epsilon, \delta \in (0, 1)$ and every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$ then for every distribution $\mathcal{D}$, with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$ it holds that

$$L_D(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

## Nonuniform learnability

A hypothesis class $\mathcal{H}$ is nonuniformly learnable if there exist a learning algorithm, A, and a function $m_{\mathcal{H}}^{NUL} : (0.1)^2 \times \mathcal{H} \to \mathbb{N}$ such that, for every $\epsilon, \delta \in (0, 1)$ and every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$ then for every distribution $\mathcal{D}$, with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$ it holds that

$$L_D(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

## Agnostic PAC learnability

A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a learning algorithm, A, and a function $m_{\mathcal{H}}^{NUL} : (0.1)^2 \times \mathcal{H} \to \mathbb{N}$ such that, for every $\epsilon, \delta \in (0, 1)$ and for every $\mathcal{D}$, if $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ then with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$ it holds that

$$L_D(A(S)) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

# Structural Risk Minimization (SRM) paradigm

So far we encoded prior knowledge by specifying a hypothesis class $\mathcal{H}$.

# Structural Risk Minimization (SRM) paradigm

So far we encoded prior knowledge by specifying a hypothesis class $\mathcal{H}$.
Express prior knowledge via specifying preferences over hypotheses within $\mathcal{H}$

- $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$, $\mathcal{H}_n$ has uniform convergence property.
- $w : \mathbb{N} \to [0, 1]$, $\sum_{n=1}^{\infty} w(n) \leq 1$
- $\epsilon_n : \mathbb{N} \times (0, 1) \to (0, 1)$, $\epsilon_n(m, \delta) = min\{\epsilon \in (0, 1) : m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) \leq m\}$

# Structural Risk Minimization (SRM) paradigm

So far we encoded prior knowledge by specifying a hypothesis class $\mathcal{H}$.
Express prior knowledge via specifying preferences over hypotheses within $\mathcal{H}$

- $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$, $\mathcal{H}_n$ has uniform convergence property.
- $w : \mathbb{N} \to [0,1]$, $\sum_{n=1}^{\infty} w(n) \leq 1$
- $\epsilon_n : \mathbb{N} \times (0,1) \to (0,1)$, $\epsilon_n(m,\delta) = min\{\epsilon \in (0,1) : m^{UC}_{\mathcal{H}_n}(\epsilon,\delta) \leq m\}$

  It follows that for every m and $\delta$, with probability of at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$ we have that

$$\forall h \in \mathcal{H}_n, |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m,\delta)$$

# Theorem: $\epsilon_n$-representativeness under SRM assumptions

Let $w : \mathbb{N} \to [0, 1]$ be a function such that $\sum_{n=1}^{\infty} w(n) \leq 1$. Let $\mathcal{H}$ be a hypothesis class that can be written as $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$ where for each n, $\mathcal{H}_n$ satisfies the uniform convergence property with a sample complexity function $m^{UC}_{\mathcal{H}_n}$. Then for every $\delta \in (0, 1)$ and distribution $\mathcal{D}$, with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$, the following bound holds (simultaneously) for every $n \in \mathbb{N}$ and $h \in \mathcal{H}_n$.

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta)$$

Therefore, for every $\delta \in (0, 1)$ and distribution $\mathcal{D}$, with probability of at least $1 - \delta$ it holds that

$$\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) \leq L_S(h) + \min_{n: h \in \mathcal{H}_n} \epsilon_n(m, w(n) \cdot \delta)$$

# Theorem: $\epsilon_n$-representativeness under SRM assumptions

Proof
$\forall n$ define $\delta_n = w(n)\delta$. Assuming uniform convergence for all n with rate

$$\forall h \in \mathcal{H}_n, |L_{\mathcal{D}}(h) - L_S(h) \leq \epsilon_n(m, \delta)$$

we get that if we fix n in advance, then with probability $\geq 1 - \delta_n$ over the choice of $S \sim \mathcal{D}^m$,

$$\forall h \in \mathcal{H}_n, |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, \delta_n)$$

Applying the union bound over $n = 1, 2, ...$ we obtain that with probability of at least $1 - \sum_n \delta_n = 1 - \delta \sum_n w(n) \geq 1 - \delta$, the preceding holds for all n.

# Structural Risk Minimization (SRM) complete

- prior knowledge: $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$, where $\mathcal{H}_n$ has uniform convergence with $m^{UC}_{\mathcal{H}_n}$
  $w : \mathbb{N} \to [0, 1]$, where $\sum_{n=1}^{\infty} w(n) \leq 1$
- define: $\epsilon_n(m, \delta) = min\{\epsilon \in (0, 1) : m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) \leq m\}$ and $n(h) = min\{n : h \in \mathcal{H}_n\}$
- input: training set $S \sim \mathcal{D}^m$, confidence $\delta$
- output: $h \in argmin_{h \in \mathcal{H}}[L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta)]$

SRM suitable for nonuniform learning of every class, which is countable union of uniformly converging hypothesis classes.

**Nonuniform learnability with SRM rule**

Let $\mathcal{H}$ be a hypothesis class such that $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ has the uniform convergence property with sample complexity $m^{UC}_{\mathcal{H}_n}$. Let $w : \mathbb{N} \to [0, 1]$ be such that $w(n) = \frac{6}{n^2 \pi^2}$. Then, $\mathcal{H}$ is nonuniformly learnable using the SRM rule with rate

$$m^{NUL}_{\mathcal{H}}(\epsilon, \delta, h) \leq m^{UC}_{\mathcal{H}_{n(h)}}(\frac{\epsilon}{2}, \frac{6\delta}{(\pi n(h))^2})$$

# proof

A: SRM algorithm with respect to w.
For every $h \in \mathcal{H}, \epsilon, \delta$ let $m \geq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2, w(n(h))\delta)$.
Using the previous theorem: with probability of at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$, we have for every $h' \in \mathcal{H}$,

$$L_{\mathcal{D}}(h') \leq L_S(h') + \epsilon_{n(h')}(m, w(n(h'))\delta)$$

Applying for A(S) returned from SRM rule and by definition of SRM we get

$$L_{\mathcal{D}}(A(S)) \leq min_{h'}[L_S(h') + \epsilon_{n(h')}(m, w(n(h'))\delta)] \leq L_S(h) + \epsilon_{n(h)}(m, w(n(h))\delta)$$

Finally if $m \geq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2, w(n(h))\delta)$ then $\epsilon_{n(h)}(m, w(n(h))\delta) \leq \epsilon/2$
UC $\Rightarrow$ with probability more than $1-\delta$, $L_S(h) \leq L_{\mathcal{D}}(h) + \epsilon/2$
Combining we obtain that $L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$

# Corollary

Let $\mathcal{H}$ be a hypothesis class that can be written as a countable union of hypothesis classes, $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ enjoys the uniform convergence property. Then $\mathcal{H}$ is nonuniformly learnable.

# Corollary

Let $\mathcal{H}$ be a hypothesis class that can be written as a countable union of hypothesis classes, $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ enjoys the uniform convergence property. Then $\mathcal{H}$ is nonuniformly learnable.

In chapter 4 we saw uniform convergence is sufficient for agnostic PAC learnability. The corollary generalizes this result to nonuniform learnability.

# Nonuniform learnability is a strict relaxation of agnostic PAC learnability

Example

- $\mathcal{X} = \mathbb{R}$
- for every $n \in \mathbb{N}$, $\mathcal{H}_n$ class of polynomial classifiers of degree $n$. $h(x) = sign(p(x))$, where $p : \mathbb{R} \to \mathbb{R}$ is a polynomial of degree n.
- let $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$. Therefore $\mathcal{H}$ is the class of all polynomial classifiers over $\mathbb{R}$.
- it easy to see that $VCdim(\mathcal{H}) = \infty$ while $VCdim(\mathcal{H}_n) = n + 1$

Hence, $\mathcal{H}$ is not PAC learnable, while according to previous corollary, $\mathcal{H}$ is nonuniformly learnable.

# Nonuniform learnability necessary and sufficient conditions

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable if and only if it is a countable union of agnostic PAC learnable hypothesis classes.

Proof

$\Rightarrow$

Assume $\mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$, $\mathcal{H}_n$ PAC learnable.
Using fundamental theorem of statistical learning, $\mathcal{H}_n$ uniform convergent.
Using the previous theorem we obtain that $\mathcal{H}$ is nonuniform learnable.

$\Leftarrow$

Assume that $\mathcal{H}$ is nonuniform learnable using some algorithm A. For every $n \in \mathbb{N}$, $\mathcal{H}_n = h \in \mathcal{H} : m_{\mathcal{H}}^{NUL}(1/8, 1/7, h) \leq n$. So $\mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$.
By the definition of $m_{\mathcal{H}}^{NUL}$, for any distribution $\mathcal{D}$ satisfying realizability assumption with respect to $\mathcal{H}_n$, with probability $\geq 6/7$ over $S \sim \mathcal{D}^n$, we have $L_{\mathcal{D}}(A(S)) \leq 1/8$.
Using F.T.S.L. the VC-dimension of $\mathcal{H}_n$ must be finite, $\mathcal{H}_n$ is agnostic PAC learnable.

# Uniform convergence

We say a hypothesis class $\mathcal{H}$ has the uniform convergence property (w.r.t. a domain Z and a loss function $\ell$) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, S is $\epsilon$-representative.

A training set S is called $\epsilon$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function $\ell$, and distribution $\mathcal{D}$) if

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