Nonuniform Learnability
Nonuniform Learnability

Definition

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

$$L_D(A(S)) \leq L_D(h) + \epsilon.$$ 

Difference with PAC learning?

The sample size $m$ may depend on the hypothesis $h$ to which the error of $A(S)$ is compared.
Nonuniform Learnability: Characterization

When is an hypothesis class nonuniformly learnable?

**Theorem**

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.

**Note:**

- nonuniform learnability is a \textit{relaxation} of agnostic PAC learning;
- it is a \textit{strict relaxation}: there are hypothesis classes that are nonuniform learnable but are not agnostic PAC learnable.
In PAC learning: prior knowledge is encoded in choice of $\mathcal{H}$.

Different way to encode prior knowledge: assign a weight to various subsets of hypotheses, high weight = stronger preference for the subset.

Let

$$\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$$

and assume $\forall n \in \mathbb{N}$: $\mathcal{H}_n$ has the uniform convergence property, with sample complexity function $m^{\text{UC}}_{\mathcal{H}_n}(\epsilon, \delta)$
Define the function $\epsilon_n : \mathbb{N} \times (0, 1) \rightarrow (0, 1)$ as

$$\epsilon_n(m, \delta) = \min \left\{ \epsilon \in (0, 1) : m_{\mathcal{H}_n}^{\text{UC}}(\epsilon, \delta) \leq m \right\}$$

That is: for a fixed sample size $m$, we are interested in the lowest possible upper bound on the gap between empirical and true risk by using $m$ samples.
$w : \mathbb{N} \to [0, 1]$ be such that $\sum_{n=1}^{+\infty} w(n) \leq 1$.

$w$ is a weight function over $\mathcal{H}_1, \mathcal{H}_2, \ldots$

**Example:**

- if $\mathcal{H} = \bigcup_{n=1}^{N} \mathcal{H}_n \Rightarrow w(n) = \frac{1}{N}$ for $n = 1, \ldots, N$
- if $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$:
  - $w(n) = 2^{-n}$ or
  - $w(n) = \frac{6}{\pi^2 n^2}$
Weight Function and Generalization Error

**Theorem**

Let $w$ be a weight function. Let $\mathcal{H}$ be a hypothesis class with:

- $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$
- for all $n$, $\mathcal{H}_n$ satisfies uniform convergence property
- sample complexity function for $\mathcal{H}_n$ is $m^{UC}_{\mathcal{H}_n}$

Then, for every $\delta \in (0, 1)$ and distribution $\mathcal{D}$, with probability at least $1 - \delta$ (over the choice of $S \sim \mathcal{D}^m$), the following holds simultaneously for every $n \in \mathbb{N}$ and $h \in \mathcal{H}_n$:

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

Proof: on the board and in the book.
Corollary

Let $w$ be a weight function. Let $\mathcal{H}$ be a hypothesis class with:

- $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$
- for all $n$, $\mathcal{H}_n$ satisfies uniform convergence property
- sample complexity function for $\mathcal{H}_n$ is $m^{UC}_{\mathcal{H}_n}$.

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

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

Let $n(h) = \min\{n : h \in \mathcal{H}_n\}$. The above implies:

$$L_D(h) \leq L_S(h) + \epsilon_{n(h)}(m, w(n) \cdot \delta)$$
Structural Risk Minimization (SRM)

A paradigm for learning that is alternative to ERM

prior knowledge:

- $\mathcal{H} = \bigcup_n \mathcal{H}_n$ where $\mathcal{H}_n$ has uniform convergence with $m_{\mathcal{H}_n}^{\text{UC}}$
- $w : \mathbb{N} \rightarrow [0, 1]$ where $\sum_n w(n) \leq 1$

define:

- $\epsilon_n(m, \delta) = \min \{\epsilon \in (0, 1) : m_{\mathcal{H}_n}^{\text{UC}}(\epsilon, \delta) \leq m\}$
- $n(h) = \min \{n : h \in \mathcal{H}_n\}$

input: training set $S \sim D^m$, confidence $\delta$

output: $h \in \arg\min_{h \in \mathcal{H}} [L_S(h) + \epsilon_{n(h)}(m, w(n(h)))\delta]$
SRM and Nonuniform Learnability

**Theorem**

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

$$m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h) \leq m_{\mathcal{H}_n(h)}^{UC} \left(\frac{\epsilon}{2}, \frac{6\delta}{(\pi n(h))^2}\right)$$

SRM allows to learn hypothesis classes with infinite VC-dimension, even if for a weaker notion of learnability.
Description Language

\[ \Sigma = \text{finite set of symbols (e.g., } \Sigma = \{0, 1\}) \]

\[ \Sigma^* = \text{set of all finite length strings from } \Sigma \]

Let \( \sigma \in \Sigma^* : |\sigma| = \text{length of } \sigma \)

**Definition**

Let \( \mathcal{H} \) be a hypothesis class. A *description* language for \( \mathcal{H} \) is a function \( d : \mathcal{H} \rightarrow \Sigma^* \).

- \( d(h) \) is the *description* of \( h \);
- \( |h| = \text{length of } d(h) \)

**Requirement**: \( d \) must be *prefix-free*

- for every \( h, h' \) with \( h \neq h' \) we have \( d(h) \) is not a prefix of \( d(h') \)
Lemma (Kraft Inequality)

If $S \subseteq 0, 1^*$ is a prefix-free set of strings, then

$$\sum_{\sigma \in S} \frac{1}{2^{\|\sigma\|}} \leq 1$$

Proof.

Define probability distribution over $S$:

- keep tossing an unbiased coin with outcomes 0 and 1
- stop when outcome is a member of $S$

Let $\mathbb{P}[\sigma] = \text{probability } \sigma \in S \text{ is generated by process above}$.  

$S$ is prefix-free $\Rightarrow \mathbb{P} [\sigma] = \frac{1}{2^{\|\sigma\|}}$

Therefore

$$\sum_{\sigma \in S} \frac{1}{2^{\|\sigma\|}} = \sum_{\sigma \in S} \mathbb{P} [\sigma] \leq 1$$
Minimum Description Length

Weighting function: \( w(h) = \frac{1}{2^{|h|}} \)

**Theorem**

Let \( \mathcal{H} \) be a hypothesis class and let \( d: \mathcal{H} \to \{0, 1\}^* \) be a prefix-free description language for \( \mathcal{H} \). Then:

- for every sample size \( m \)
- for every \( \delta > 0 \)
- for every probability distribution \( D \)

with probability \( \geq 1 - \delta \) (over the choice of \( S \sim D^m \)) we have

\[
\forall h \in \mathcal{H} : L_D(h) \leq L_S(h) + \sqrt{\frac{|h| + \ln(2/\delta)}{2m}}
\]
Alternative learning paradigm for $\mathcal{H}$: Minimum Description Length (MDL)

prior knowledge:
- $\mathcal{H}$ is a countable hypothesis class
- $\mathcal{H}$ is described by a prefix-free language over $\{0, 1\}$
- for every $h \in \mathcal{H}$: $|h|$ = length of representation of $h$

input: training set $S \sim \mathcal{D}^m$, confidence $\delta$

output: $h \in \arg \min_{h \in \mathcal{H}} \left[ L_S(h) + \sqrt{\frac{|h| + \ln(2/\delta)}{2m}} \right]$
Occam’s Razor

“All things being equal, the simplest solution tends to be the best one.”

William of Ockham (1287-1347)

“A short explanation (that is, a hypothesis that a short length) tends to be more valid than a long explanation”

**Question**: What about the choice of the description language?!

**Answer**: As long as you choose *before* seeing the data, the bounds holds...
Bibliography

[UML] Chapter 7