Non-uniform learning

- If $VC(H)$ is finite, $\widehat{err}_S$ converges uniformly to err_D.

That is, with probability at least $1-\delta$,

$$\sup_{h \in H} \left| \widehat{err}_S(h) - err_D(h) \right| \leq \varepsilon$$

assuming $|S| \geq \widetilde{\Omega}\left(\frac{VC(H) + \log(1/\delta)}{\varepsilon^2} \right)$

- Note that the guarantee

$$\left| \widehat{err}_S(h) - err_D(h) \right|$$

does <u>NOT</u> depend on h.

- That is, the convergence is uniform over all $h \in H$.

- We now consider non-uniform convergence.

Definition: A hypothesis class $H \subseteq Y^X$ is non-uniformly learnable iff there exists a learning algorithm A such that for all $\varepsilon, \delta \in (0,1)$, $h \in H$ there exists positive integer m such that for every distribution D over $X \times Y$ if S is an i.i.d. sample from D of size m

then

$$err_D(A(S)) \leq err_D(h) + \varepsilon.$$

---

1) Note the order of quantifiers:

- $\forall \varepsilon \forall \delta \forall h \in H \; \exists m \; \forall D \; \ldots$

Compare that with agnostic PAC:

- $\forall \varepsilon \forall \delta \; \exists m \; \forall D \forall h \; \ldots$

2) Note that if H is agnostically PAC learnable then it is non-uniformly learnable.

3) The reverse implication is false.

4) Given a learning algorithm A, we define $m_A(\varepsilon, \delta, h)$ to be the smallest sample size such that if $|S| \geq m_A(\varepsilon, \delta, h)$ then with probability at least $1-\delta$,

$$err_D(A(S)) \leq err_D(h) + \varepsilon.$$

Lemma: If H is non-uniformly learnable the H is a countable union of classes with finite VC-dimension. That is,

$$H = \bigcup_{n=1}^{\infty} H_n$$

where $VC(H_n) < +\infty$ for all n.

Proof: • Let A be a learning algorithm for non-uniform learning of H.

• Let $m_A : (0,1)^2 \times H \to \mathbb{N}$ be the sample size function.

• Define

$$H_n = \{ h \in H : m_A(1/8, 1/7, h) \leq n \}$$

• Clearly

$$H = \bigcup_{n=1}^{\infty} H_n$$

- H_n is PAC learnable using A since $m_A(1/\varepsilon, 1/7, h) \leq n$.

- Thus H_n has finite VC dimension.

Example:

$$H_n = \{ \text{sign}(p(x)) : p(x) = a_n x^n + \dots + a_1 x + a_0 \}$$

$$H = \{ \text{sign}(p(x)) : p(x) \text{ is a polynomial} \}$$

- H_n is (agnostically) PAC learnable

- H is NOT (agnostically) PAC learn.

- We will see later that H is non-uniformly learnable.

Structural risk minimization

- Suppose $H = \bigcup_{n=1}^{\infty} H_n$ where

$H_1, H_2 \ldots$ have finite VC-dim.

- Each H_n satisfies w.p. $1-\delta$,

$$\sup_{h \in H_n} \left| \widehat{err}_S(h) - err_D(h) \right| \leq \varepsilon$$

if $|S| \geq m_{UC}(\varepsilon, \delta, VC(H_n))$

- Consider $w : \mathbb{N} \to [0,1]$ such that

$$\sum_{n=1}^{\infty} w(n) = 1$$

- w can be thought of as
  a) weight function
  b) probability distribution
  c) prior distribution

- For example

$$w(n) = 2^{-n} \quad \text{or} \quad w(n) = \frac{6}{\pi^2 n^2}$$

- If H is finite union $H_1 \cup \dots \cup H_N$

$$w(n) = \frac{1}{N}$$

- Let

$$\varepsilon_n(m, \delta) = \min\left\{\varepsilon \in (0,1) : m_{UC}(\varepsilon, \delta, VC(H_n)) \leq m\right\}$$

Theorem:

Let D be any distribution over $X \times Y$. Let $H = \bigcup_{n=1}^{\infty} H_n$ where $VC(H_n) < +\infty$. Let $w : \mathbb{N} \to [0,1]$ s.t. $\sum_{n=1}^{\infty} w(n) = 1$. Let m be a positive integer. Let $\delta \in (0,1]$. Let S be i.i.d. sample from D of size m. Then, with

probability at least $1-\delta$,
for all $n \geq 1$ and all $h \in H_n$,

$$|\widehat{err}_S(h) - err_D(h)| \leq \varepsilon_n(m, w(n)\cdot\delta).$$

Also, w.p. $1-\delta$, for all $h \in H$,

$$err_D(h) \leq \widehat{err}_S(h) + \min_{n : h \in H_n} \varepsilon_n(m, w(n)\cdot\delta)$$

Proof:

- W.p. $1 - \delta \cdot w(n)$

$$\sup_{h \in H_n} \left| err_D(h) - \widehat{err}_S(h) \right| \leq \varepsilon_n\left(m, w(n)\cdot\delta\right)$$

- Union bound over all $n = 1, 2, \ldots$ implies that w.p. $1-\delta$ for all $n \geq 1$,

$$\sup_{h \in H_n} \left| err_D(h) - \widehat{err}_S(h) \right| \leq \varepsilon_n\left(m, w(n) \cdot \delta\right)$$

- The second part is trivial consequence of the first one. ∎

- Let $n(h) = \min\{n \in \mathbb{N} : h \in H_n\}$

- Theorem implies that

$$err(h) \leq \widehat{err}_S(h) + \varepsilon_{n(h)}\left(m, w(n(h)) \cdot \delta\right)$$

$err_D(h) = \underbrace{\quad\quad\quad\quad\quad\quad}$

Can be computed from S only !!!

SRM algorithm

Parameters: $\delta \in (0,1)$, decomposition $H = \bigcup_{n=1}^{\infty} H_n$, $w : \mathbb{N} \to [0,1]$ s.t. $\sum_{n=1}^{\infty} w(n) = 1$

Input: Labeled sample S

Output:

$$\hat{h} = \underset{h \in H}{\operatorname{argmin}} \left(\widehat{err}_S(h) + \varepsilon_{n(h)}\left(m, w(n(h)) \cdot \delta\right) \right)$$

Lemma:

Let $H_1, H_2, \ldots$ be classes with finite VC-dimension. Let

$H = \bigcup_{n=1}^{\infty} H_n$. Then H is non-uniformly learnable using SRM with, say, $w(n) = 2^{-n}$ and, say, $\delta = 1/|S|$.

Proof:

- Let $\epsilon, \delta \in (0,1)$ and $h \in H$.
- Consider smallest n such that $h \in H_n$.
- Let $m \geq \max\left\{\frac{1}{\delta}, m_{UC}\left(\epsilon/2, \frac{\delta}{2^n}, VC(H_n)\right)\right\}$
- Let S be i.i.d. sample of size m.

By previous theorem $\forall h' \in H$

- By previous [illegible]

$$\left|err_D(h') - \widehat{err}_S(h')\right| \le \varepsilon_{n(h')}\left(m, w(n(h')) \cdot \delta\right)$$

w.p. at least $1-\delta$

- Let $\hat{h} = SRM(S)$

- By definition of SRM

$$\widehat{err}_S(\hat{h}) + \varepsilon_{n(\hat{h})}\left(m, w(n(\hat{h})) \cdot \frac{1}{m}\right) \le$$

$$\widehat{err}_S(h) + \varepsilon_{n(h)}\left(m, w(n(h)) \cdot \frac{1}{m}\right).$$

- Therefore, w.p. $1-\delta$,

$$err_D(\hat{h}) \le \widehat{err}_S(h) + \varepsilon_{n(h)}\left(m, w(n(h)) \cdot \frac{1}{m}\right)$$

$$\le \widehat{err}_S(h) + \varepsilon_{n(h)}\left(m, w(n(h)) \cdot \delta\right)$$

$$\leq err_D(h) + 2\varepsilon_{n(h)}\left(m, w(n(h)) \cdot \delta\right)$$

- Since $m \geq m_{UC}\left(\varepsilon/2, w(n(h)) \cdot \delta, VC(H_n)\right)$

$$\varepsilon_{n(h)}\left(m, w(n(h)) \cdot \delta\right) \leq \varepsilon/2 .$$

- So

$$err_D(\hat{h}) \leq err_D(h) + \varepsilon .$$

Corollary:

H is non-uniformly learnable if and only if H can be expressed as a countable union $H = \bigcup_{n=1}^{\infty} H_n$ where

$H_1, H_2, \ldots$ have finite VC-dim.

Minimum description length

- Suppose H is countable (or finite)
- Let $w: H \to [0,1)$ s.t. $\sum_{h \in H} w(h) = 1$
- For a fixed $h \in H$, Hoeffding's inequality states

$$\left| err_D(h) - \widehat{err}_S(h) \right| \leq \sqrt{\frac{\log(w(h)) + \log(2/\delta)}{2m}}$$

w.p. $1 - w(h) \cdot \delta$

- By union bound over all $h \in H$,

$$\forall h: \left| err_D(h) - \widehat{err}_S(h) \right| \leq \sqrt{\frac{\log(w(h)) + \log(2/\delta)}{2m}}$$

$\forall h \in H \quad |err_D(h) - \hat{err}_S(h)| \leq \sqrt{\frac{\log(w(h)) + \log(1/\delta)}{2m}}$

w.p. $1-\delta$.

- SRM rule becomes

$$\hat{h} = \underset{h \in H}{\operatorname{argmin}} \; \widehat{err}_S(h) + \sqrt{\frac{\log(w(h)) + \log(1/\delta)}{2m}}.$$

Prefix-free codes

Definition: Function $c: H \to \{0,1\}^*$ is called a prefix-free code iff $c(h)$ is NOT a prefix of $c(h')$ for any $h, h' \in H$ $h \neq h'$.

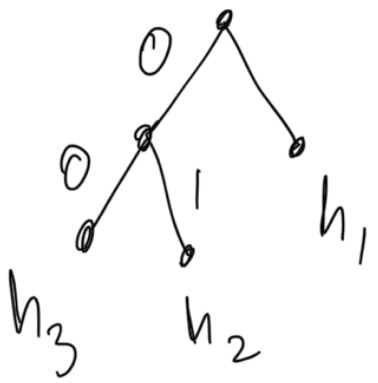
Example:

$H = \{h_1, h_2, h_3\}$

$c(h_1) = 1$

$c(h_2) = 01$

$c(h_3) = 00$



C is prefix-free

Example:

$H = \{h_1, h_2, h_3\}$

$c(h_1) = 0$

$c(h_2) = 01$

$c(h_3) = 1$

is NOT prefix-free

since $c(h_1)$ is prefix of $c(h_2)$.

Theorem (Kraft's inequality)

If $c: H \to \{0,1\}^*$ is prefix-free

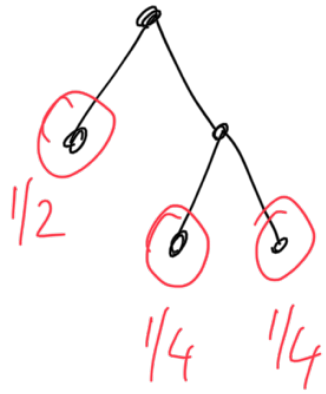
then

$$\sum_{h \in H} 2^{-|c(h)|} \leq 1.$$

Proof:

- **Fact:** In a binary tree

$$\sum_{v \in \text{leaf}(T)} 2^{-\text{depth}(v)} \leq 1.$$



- Prefix-free code corresponds to leaves of a tree.
- $\text{depth}(h) = |c(h)|$

- In MDL we can use

$$w(h) = 2^{-|c(h)|}$$

This is referred to as

- This is related to

Occam's razor

- Occam in 13-th century suggested that "simple explanations are better".