Non Uniform learnability and learning by Structural Risk Minimization.

Recap till now we have seen:

Notion of PAC learnability: $\mathcal{H}$ is agnostic PAC learnable if we can find A & $m_{\mathcal{H}}(\epsilon, \delta)$ s.t $\forall (\varepsilon, \delta) \in (0,1)^2$ and $\forall \mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ for $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left\{ L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon \right\} \geq 1 - \delta$$

Uniform convergence property: for $m \geq m_{\mathcal{H}}(\epsilon, \delta)$

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left\{ \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon \right\} \geq 1 - \delta$$

implies PAC learnability through Empirical Risk Minimization $A^{ERM}(S) = \arg\min L_S(h)$.

Finally we saw that unif conv prop holds iff VC dim ($\mathcal{H}$) is finite and also

$$m_{\mathcal{H}}(\varepsilon, \delta) \approx \frac{\text{VCdim}(\mathcal{H}) + \log 1/\delta}{\varepsilon^2}.$$

#

There are natural relaxed notions of learnability for infinite classes and associated criteria as well as algorithms.

Here we explore one such notion:

Notion of non-uniform learnability and associated Structural Risk Minimization algo.

Main idea of non-uniform learnability is to relax fact that $m_{\mathcal{H}}$ depends only on ϵ, δ.
We may make it h-dependent: indeed some hypothesis fcts might require more or less samples to be learned.

Definition. $\mathcal{H}$ is non-uniform learnable

if $\exists\, A(S)$ and $m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$ such that $\forall\, (\varepsilon, \delta) \in (0,1)^2$ and $\forall\, \mathcal{D}$ we have:

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left\{ L_{\mathcal{D}}(A(S)) \leq \min_{\substack{h \text{ s.t} \\ m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)}} L_{\mathcal{D}}(h) + \epsilon \right\} \geq 1 - \delta$$

Meaning: Fix $m = |S|$. Compare $L_{\mathcal{D}}(A(S))$ with best hypothesis in a subset of $\mathcal{H}$ namely such that $m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h) \leq m$ (instead of comparing $L_{\mathcal{D}}(A(S))$ with best hypothesis in all of $\mathcal{H}$).

Obviously we have:

$$L_{\mathcal{D}}(A(S)) \leq \min_{\substack{h \text{ s.t.} \\ m \geq m_{\mathcal{H}}^{NUL}(\epsilon,\delta,h)}} L_{\mathcal{D}}(h) + \varepsilon \Rightarrow L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$$

Thus Non Unif Learnability $\Rightarrow$ Agnostic PAC learnability.

Main Theorem:

An hypothesis class $\mathcal{H}$ is non-uniformly learnable if and only if it is a countable union $\mathcal{H} = \bigcup_{m \in \mathbb{N}} \mathcal{H}_m$ of agnostic PAC learnable classes $\mathcal{H}_m$.

- Proof of $\mathcal{H}$ NUL $\Rightarrow$ $\mathcal{H} = \bigcup \mathcal{H}_m$ with $\mathcal{H}_m$ learnable is easy and we do it first.
- Proof of converse $\mathcal{H} = \bigcup \mathcal{H}_m$, $\mathcal{H}_m$ learnable $\Rightarrow$ $\mathcal{H}$ NUL is more involved and requires notion of Structural Risk Min.

Proof of first direction:

[$\mathcal{H}$ NUL $\Rightarrow$ $\mathcal{H} = \bigcup_m \mathcal{H}_m$ with $\mathcal{H}_m$ learnable]

Assume $\mathcal{H}$ is NUL with some Algorithm $A(S)$.

Let $\mathcal{H}_m = \{ h \in \mathcal{H} \text{ s.t } m^{NUL}_{\mathcal{H}}(\varepsilon = \frac{1}{8}, \delta = \frac{1}{7}, h) \leq m \}$

Obviously $\bigcup_m \mathcal{H}_m = \mathcal{H}$.

By def of NUL:

$$\mathbb{P}_{S \sim \mathcal{D}^m} \{ L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}_m} L_{\mathcal{D}}(h) + \frac{1}{8} \} \geq \frac{6}{7}$$

Take $\mathcal{D}$ realizable s.t $\min_{h \in \mathcal{H}_m} L_{\mathcal{D}}(h) = 0$. Then for

these $\mathcal{D}$: $\mathbb{P}_{S \sim \mathcal{D}^m} \{ L_{\mathcal{D}}(A(S)) \leq \frac{1}{8} \} \geq \frac{6}{7}$

$$\Rightarrow \mathbb{P}_{S \sim \mathcal{D}^m} \{ L_{\mathcal{D}}(A(S)) \geq \frac{1}{8} \} \leq \frac{1}{7}$$

$\Rightarrow \text{VCdim}(\mathcal{H}_m) < +\infty$ (finite). Indeed otherwise the No Free Lunch Thm would hold and $\mathbb{P}_{S \sim \mathcal{D}^m} \{ L_{\mathcal{D}}(A(S)) \geq \frac{1}{8} \} \geq \frac{1}{7}$. $\Rightarrow$ $\boxed{\text{Hence } \mathcal{H}_m \text{ PAC learnable.}}$

Proof of converse direction.

[$\mathcal{H} = \bigcup_n \mathcal{H}_n$, $\mathcal{H}_n$ agnostic PAC learnable $\Rightarrow$ $\mathcal{H}$ NUL]

This requires a lemma (proved later on):

Lemma:
If $\mathcal{H} = \bigcup_n \mathcal{H}_n$ where $\mathcal{H}_n$ has unif convergence property then $\mathcal{H}$ is non-uniformly learnable.

Now assume $\mathcal{H} = \bigcup_n \mathcal{H}_n$, $\mathcal{H}_n$ agnostic PAC learnable. Then by Fund Thm of learning $\mathcal{H}_n$ must have the unif conv property. Thus by the lemma $\mathcal{H}$ is non-unif-learnable.

$\blacksquare$

The rest of this lecture is devoted to the tools allowing to prove the lemma.

Structural Risk Minimization.

Some intuitions first:

* $\mathcal{H} = \bigcup_n \mathcal{H}_n$ and we imagine that we have some prior confidence associated to each $\mathcal{H}_n$. So we associate some "weight", $0 \leq w_n \leq 1$, $\sum_n w_n = 1$, to each $\mathcal{H}_n$. The higher the weight the more we believe that the algorithm should belong to $\mathcal{H}_n$.

* in the lemma each class $\mathcal{H}_n$ has the unif conv property for some sample complexity $m^{UC}_{\mathcal{H}_n}(\epsilon, \delta)$.

* Given $m = |S|$ = sample size and confidence param δ we will consider the best "slack"

$$\mathcal{E}_n(m, \delta) = \min\left(\mathcal{E} \; ; \; m \geq m^{UC}_{\mathcal{H}_n}(\mathcal{E}, \delta)\right)$$

$$\left(\text{Recall } m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) \simeq \frac{\text{vcdim}\,\mathcal{H}_n + \log 1/\delta}{\mathcal{E}^2}\right).$$

Theorem :

Let $\mathcal{H} = \bigcup_n \mathcal{H}_n$ with $\mathcal{H}_n$ satisfying UC property with some $m^{UC}_{\mathcal{H}_n}(\epsilon, \delta)$.

Let $\varepsilon_n(m, \delta) = \min(\varepsilon : m \geq m^{UC}_{\mathcal{H}_n}(\varepsilon, \delta))$

Let $0 \leq w_n \leq 1$, $\sum_n w_n = 1$ any set of "weights" associated to $\mathcal{H}_n$ (set $w_0 \geq w_1 \geq w_2 \geq \ldots$).

Then

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left\{ \forall n,\ \sup_{h \in \mathcal{H}_n} \left| L_{\mathcal{D}}(h) - L_S(h) \right| \leq \varepsilon_n(m, w_n \delta) \right\} \geq 1 - \delta .$$

Proof of Theorem.

By UC for each $\mathcal{H}_n$ we have for $m \geq m^{UC}_{\mathcal{H}_n}(\varepsilon, \delta)$

$$\mathbb{P}\left\{ \sup_{h \in \mathcal{H}_n} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon \right\} \geq 1-\delta .$$

Here m and δ are fixed. So choose the best possible ε i.e:

$$\varepsilon \to \varepsilon_n(m, \delta) = \min\left(\varepsilon ; \; m^{UC}_{\mathcal{H}_n}(\varepsilon, \delta) \leq m\right).$$

$$\mathbb{P}\left\{ \sup_{h \in \mathcal{H}_n} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon_n(m, \delta)\right\} \geq 1-\delta$$

Apply this to $\delta \to w_n \delta$ for each $n \in \mathbb{N}$.

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left\{ \sup_{h \in \mathcal{H}_n} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon_n(m, w_n\delta)\right\} \geq 1-w_n\delta$$

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left\{ \sup_{h \in \mathcal{H}_n} |L_{\mathcal{D}}(h) - L_S(h)| \geq \varepsilon_n(m, w_n\delta)\right\} \leq w_n\delta$$

this is valid for all $n \in \mathbb{N}$.

Now we sum these inequalities over $m \in \mathbb{N}$:

$$\sum_{m \in \mathbb{N}} \mathbb{P}_{S \sim \mathcal{D}^m} \left\{ \sup_{h \in \mathcal{H}_m} |L_{\mathcal{D}}(h) - L_S(h)| \geq \varepsilon_m(m, w_m \delta) \right\} \leq \delta$$

By the union bound: $\mathbb{P}\{\exists m : \ldots\} \leq \sum_{m \in \mathbb{N}} \mathbb{P}\{\ldots\}$

Taking the converse probabilities:

$$\mathbb{P}\left\{ \forall m : \sup_{h \in \mathcal{H}_m} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon_m(m, w_m \delta) \right\} \geq 1 - \delta$$

$\blacksquare$

Let us discuss some implications of this thm. This will allow us to introduce the notion of structural risk.

Thm says that with high prob we have

$$\forall m ; \sup_{h \in \mathcal{H}_m} | L_{\mathcal{D}}(h) - L_S(h) | \leq \varepsilon_m (m, w_m \delta)$$

this means also that

$$\forall m , \forall h \in \mathcal{H}_m : | L_{\mathcal{D}}(h) - L_S(h) | \leq \varepsilon_m (m, w_m \delta)$$

Now we can fix $h \in \mathcal{H}$. We have

$$| L_{\mathcal{D}}(h) - L_S(h) | \leq \varepsilon_m (m, w_m \delta) ; \forall m \text{ s.t } h \in \mathcal{H}_m .$$

$$\Rightarrow | L_{\mathcal{D}}(h) - L_S(h) | \leq \min_{m : h \in \mathcal{H}_m} \varepsilon_m (m, w_m \delta)$$

$$\leq \varepsilon_{m(h)} (m, w_{m(h)} \delta)$$

we choose $m(h) = \min (m : h \in \mathcal{H}_m)$ First $\mathcal{H}_m$ where h occurs

Thus the theorem states that with high probability we have:

$$L_{\mathcal{D}}(h) \leq L_S(h) + \varepsilon_{n(h)}(m, w_{n(h)}\delta)$$

This suggests that to make $L_{\mathcal{D}}(h)$ small a good idea is to use the

Structural risk minimization rule or algo:

$$\boxed{A^{SRM}(S) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \left(L_S(h) + \varepsilon_{n(h)}(m, w_{n(h)}\delta) \right)}$$

This means we penalize the empirical risk with an additional term $\varepsilon_{n(h)}(m, w_{n(h)}\delta)$. This term is constructed from the decomposition $\mathcal{H} = \bigcup_n \mathcal{H}_n$ with $m^{UC}_{\mathcal{H}_n}(\varepsilon, \delta)$ and $\varepsilon_n(m, \delta) = \min(\varepsilon ; m \geq m^{UC}_{\mathcal{H}_n}(\varepsilon, \delta))$ $n(h) = \min(n : h \in \mathcal{H}_n)$ and a set of weights $0 \leq w_n \leq 1$.

Recall that this construction was motivated by the lemma:

<u>Lemma</u> Let $\mathcal{H} = \bigcup_n \mathcal{H}_n$ with each $\mathcal{H}_n$ satisfying the UC property. Then $\mathcal{H}$ is non uniformly learnable with

$$m_{\mathcal{H}}^{NUL}(\varepsilon, \delta, h) \leq m_{\mathcal{H}_{n(h)}}^{UC}\left(\frac{\varepsilon}{2}, w_{n(h)}\delta\right).$$

<u>Proof</u>:

By definition of $A^{SRM}(S)$ we have

$$L_{\mathcal{D}}(A^{SRM}(S)) = \min_{h\in\mathcal{H}}\left(L_S(h) + \varepsilon_{n(h)}(m, w_{n(h)}\delta)\right)$$

$$\leq L_S(h) + \varepsilon_{n(h)}(m, w_{n(h)}\delta)$$

Take $m \geq m^{UC}_{\mathcal{H}_{n(h)}}\left(\frac{\varepsilon}{2}, w_{n(h)}\delta\right)$.

From the above definitions it can then be seen that

$$\varepsilon_{m(h)}(m, w_{m(h)}\delta) \leq \frac{\varepsilon}{2},$$

Thus for $m \geq m^{UC}_{\mathcal{H}_{m(h)}}(\frac{\varepsilon}{2}, w_{m(h)}\delta)$ we have

$$L_{\mathcal{D}}(A^{SRM}(S)) \leq L_S(h) + \frac{\varepsilon}{2}$$

Now $\mathcal{H}_{m(h)}$ (recall $h \in \mathcal{H}_{m(h)}$ by def of $m(h)$) satisfies UC property. So with prob at least $1-\delta$ we have $L_S(h) \leq L_{\mathcal{D}}(h) + \frac{\varepsilon}{2}$.

Therefore we conclude that with prob $\geq 1-\delta$

$$L_{\mathcal{D}}(A^{SRM}(S)) \leq L_{\mathcal{D}}(h) + \varepsilon$$

for $h \in \mathcal{H}_{m(h)}$ and $m \geq m^{UC}_{\mathcal{H}_{m(h)}}(\frac{\varepsilon}{2}, w_{m(h)}\delta)$.

Now take any $m^{NUL}_{\mathcal{H}}(\epsilon, \delta, h) = m^{UC}_{\mathcal{H}_{n(h)}}\left(\frac{\epsilon}{2}, w_{n(h)}\, \delta\right)$

The above statement is equivalent to

$$\mathbb{P}\left(L_{\mathcal{D}}\left(A^{SRM}(S)\right) \le \min_{\substack{h \text{ such that} \\ m^{NUL}_{\mathcal{H}}(\epsilon,\delta,h) \le m}} L_{\mathcal{D}}(h) + \epsilon \right) \geq 1 - \delta$$

This is the def of Non Unif Learnability. $\blacksquare$