Uniform Convergence is Sufficient for Learnability

Minimizing empirical risk (on a sample S) one hopes that empirical risk is a good approximation of true risk.

DEFINITION 4.1. Sample is called ϵ representative if

$$\forall h \in \mathcal{H}, |L_D(h) - L_S(h)| < \epsilon$$

If S is $\frac{\epsilon}{2}$ representative, the ERM learning rule is possible to find h_S such that:

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

Proof. Start from definition 4.1. and break down the inequality for absolute values so:

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

$$L_S(h) - L_D(h) \geq -\frac{\epsilon}{2}$$

Take the second equation and write it as:

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

Now, remember that h_S is an ERM on S. This means it is also smaller or equal to any other hypothesis for the sample meaning we can upper bound it by any hypothesis $L_S(h)$. Finally, again, we use the substitution from the proof beginning, only apply it for $L_D(h)$ instead of $L_S(h)$.

- TODO: what changes if it's not an agnostic learner? Agnostic had zero loss on hypothesis.

DEFINITION 4.3. (*Uniform convergence*) A hypothesis class $\mathcal{H}$ has the uniform convergence property if exists a function $m_{\mathcal{H}} : (0,1)^2 -> \mathbb{N}$, a sample $S \sim Z$ of size $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ is ϵ-representative.

- TODO Why is it uniform: Uniform refers to the fact that this applies to $\forall h \in \mathcal{H}$ It is important that the loss of every h on a sample S is a good representative of the loss on the entire dataset.
- TODO what is the motivation for this lemma/definition in the grand scheme of things?

COROLLARY 4.4. If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}(\epsilon, \delta)$ then the class is agnostically PAC learnable with complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}(\epsilon/2, \delta)$

PROOF. Next, the goal is to prove that every finite hypothesis class with uniform convergence is PAC learnable We fix δ, ϵ, sample $S \sim D$, and use a universally convergent hypothesis class $\mathcal{H}$ such that:

$$D^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| \leq \epsilon\}) \geq 1 - \delta$$

The hypothesis is taken out, union bound is applied and transformed to a sum:

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

Now, we look at the uniform convergence, and how it behaves in large numbers. The loss of the sample is written down as $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$, and the true loss is an expectation over sample loss $L_D(h) = \mathcal{E}_{z \sim D}[l(h, z)]$ Hoeffding's inequality quantifies the gap between empirical averages and their expected value on a finite sample

- TODO: how exactly does the asymptotic rule not apply here?

The expecatation is approximated with μ, and the empirical average with the empirical loss $\frac{1}{m}\sum_{i=1}^{m} \theta_i$ Now inputting those into Hoeffding's inequality:

$$D^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} 2exp(-2m\epsilon^2)$$

$$D^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon\}) \leq 2|\mathcal{H}|exp(-2m\epsilon^2)$$

Finally, since:

$$D^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| \leq \epsilon\}) \leq \delta$$

once gets that

$$m \geq \frac{log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

Which means that the sample complexity is an upper bound integer

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

From a practical perspective, it makes sense to estimate a problem difficulty, by measuring it's sample complexity (how many samples do you need to guarantee agnostic PAC learning).