7 Nonuniform learnability

PAC learnability has required that learnability was dependent on accuracy and confidence parameters. Nonuniform learnability is a more general notion and includes the hypothesis as a parameter of learnability. It will be shown that a sufficient condition for nonuniform learnability is that $\mathcal{H}$ is a countable union of hypothesis classes, each of which have the uniform convergence property. A new learning paradigm is introduced, Structured Risk Minimization (SRM) for countable hypothesis classes and the minimum descriptive length paradigm.

7.1. NONUNIFORM LEARNABILITY

We define that a hypothesis h is (ϵ, δ) competitive with hypothesis h' if, with probability higher than $1 - \delta$:

$$L_D(h) \leq L_D(h') + \epsilon$$

In PAC learnability, we are looking for a low risk hypothesis with respect to the minimal risk achieved by the hypothesis in our hypothesis class. The sample size in PAC depends only on accuracy and confidence, but in nonuniform learnability, we allow the sample size to be of $m_{\mathcal{H}}(\epsilon, \delta, h)$, making it also dependent on the hypothesis.

Definition 7.1. A hypothesis class $\mathcal{H}$ is *nonuniformly learnable* if there exists a learning algorithm A, function $m_{\mathcal{H}}(\epsilon, \delta, h)$ such that for every $\epsilon, \delta \in (0, 1)^2$ and for every $h \in \mathcal{H}$ if $m > m_{\mathcal{H}}(\epsilon, \delta, h)$ and every distribution D over the sample $S \sim D$ it holds that

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

The difference between nonuniform and agnostic PAC learnability is that we compare against the best one (in PAC), whereas here we compare against every hypothesis. This makes nonuniform learning a more general concept, implying that PAC learnable implies nonuniform learnable, but not vice versa.

7.1.1 Characterizing Nonuniform Learnability

We showed that a class of binary classifiers is PAC learnable if it's VC dimension is finite. Now, we characterize nonuniformly learnable classes.

Theorem 7.2. 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.

Theorem 7.3. Let $\mathcal{H}$ be a countable union of hypothesis classes $\mathcal{H}_n$, where each $\mathcal{H}_n$ has the uniform convergence property. Then, $\mathcal{H}$ is nonuniformly learnable.

Theorem 7.2. proof First direction. If a class is PAC agnostic learnable, then it is also nonuniformly learnable. Then, by Theorem 7.3. the countable union is nonuniformly learnable. Second direction: $\mathcal{H}$ is nonuniformly learnable using some algorithm A. Then, define $H_n = \{h \in \mathcal{H} : m(1/8, 1/7, h) \leq n\}$ So, now we have the sample size for which we can say that $L_D(A(S)) \leq 1/8$, which satisifes the no-free-lunch theorem, implying that the VC-dimension is finite, therefore H_n is agnostically learnable.

7.2. STRUCTURAL RISK MINIMIZATION

Another way to specify prior beliefs over classes is to put some weights over preffered ones. In Structural Risk Minimization (SRM) we define the hypothesis class as a union of n classes and then specify a weight function $w : \mathbb{N} \to [0, 1]$ assigning stronger weights to classes with stronger prior beliefs. Here we learn how to learn with such prior knowledge and talk about weighting schemes, inclucing Minimal Description Length.

Split the $\mathcal{H}$ into n countable unions. Then define a function $\epsilon_n(m, \delta) = \min\{\epsilon \in (0, 1) : m(\epsilon, \delta) \leq m\}$ which is the minimum ϵ for a fixed sample size, confidence and n, the lowest possible upper bound on the gap between empirical and true risk achievable by m samples. In every hypothesis class $\mathcal{H}_n$ the uniform convergence property holds. Now, define a weight function over n such that it sums up to 1. It reflects the confidence to each class or complexity of different hypothesis classes. When $\mathcal{H}$ is a countable union of infinite hypothesis classes, uniform weighting is approximated using some specific functions such as $\frac{6}{\pi^2 n^2}$ or 2^{-n}.

Theorem 7.4. Let w be a weight function, $\mathcal{H}$ a countable union of hypothesis classes, then it holds that

$$|L_D(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta)$$

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

Proof. Define $\delta_n = w(n)\delta$. Uniform convergence holds for all δ_n for all hypothesis classes $\mathcal{H}_n$. Applying the union bound over $n = 1, 2, \ldots$ we obtain this holds.

- TODO which union bound, how???

Second, define $n(h) = \min\{n : h \in \mathcal{H}_n\}$, which concludes the proof.

SRM gets as input a training set, hypothesis classes with weights. Outputs the hypothesis such that it minimizes the sample error in a weighted scheme. In

SRM, we are willing to tradeoff some risk towards having the smaller ϵ_n. SRM can be used only for classes which are countable unions of uniformly converging hypothesis classes.

Theorem 7.5. Let $\mathcal{H}$ be a hypothesis class which is a countable union of uniformly converging hypothesis classes with sample complexity $m_{\mathcal{H}_n}$. Weights w is defined over n: $w(n) = \frac{6}{\pi^2 n^2}$. Then, $\mathcal{H}$ is nonuniformly learnable using SRM with rate:

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

Proof Let A be an SRM algorithm with respect to w. First, we find m such that $m \geq m_{\mathcal{H}_{n(h)}}(\epsilon, w(n(h))\delta)$ and apply theorem 7.4. to get that

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

By definition of SRM, we want the minimum h so:

$$L_D(A(S)) \leq min[L_S(h') + \epsilon_{n(h')}(m, w(n(h'))\delta)] \leq L_S(h') + \epsilon_{n(h')}(m, w(n(h'))\delta)$$

If $m \geq m_{\mathcal{H}_{n(h)}}(\epsilon/2, w(n(h))\delta)$ then $\epsilon(m, w(n(h))\delta) \leq \epsilon/2$. Which concludes the proof. This shows that any countable union of finite VC-dimension classes is nonuniformly learnable.

- TODO how to apply theorem 7.4.???
- TODO class of all binary functions over X is not countable (why?)

When comparing nonuniform learning of entire class $\mathcal{H}$ versus PAC learning of a single subset class $\mathcal{H}_n$, there is less prior knowledge in nonuniform learning, resulting in larger complexity. For example, if $VCdim(\mathcal{H}_n) = n$ that makes $m_{\mathcal{H}_n} = C\frac{n+log(1/\delta)}{\epsilon^2}$ it can be proved that sample size differences are upper bounded by $4C\frac{2log(2n)}{\epsilon^2}$.

- TODO make calculation to find difference of sample sizes

7.3 MINIMUM DESCRIPTION LENGTH AND OCCAM'S RAZOR

Define $\mathcal{H}$ as a union of countable singleton classes $\mathcal{H}_n$. Each class has uniform convergence property with rate $m = \frac{2/\delta}{2\epsilon^2}$. Deriving expression for $\epsilon_n = \sqrt{\frac{log(2/\delta)}{2m}}$. Now, we derive the SRM rule using this accuracy and arrive at:

$$argmin[L_S(h) + \sqrt{\frac{-log(w(n)) + log(2/\delta)}{2m}}]$$

Prior knowledge of hypotheses is assigned through weights. Now, we will give different weights to hypotheses based on the length of the hypothesis (complexity). Take some (any) language with finite characters drawn from some fixed alphabet. $\Sigma = \{0, 1\}$ is the alphabet. $\rho = (0, 1, 1, 0, 1)$ is a finite sequence of chars. $|\rho|$ is the length of the sequence. Set of all finite strings is denoted as Σ^*. The description language for $\mathcal{H}$ is a function $\mathcal{H} \to \Sigma^2$ and maps each member of $h \in \mathcal{H}$ to a string $d(h)$. The language is prefix-free. Prefix free sequences enjoy Kraft's inequality $\sum_{\rho \in \Sigma} \frac{1}{2^{|\rho|}} \leq 1$. Therefore, this weighting scheme can be used for our hypotheses and their weighing.

Theorem 7.7 Let $\mathcal{H}$ be a hypothesis class and $d : \mathcal{H} \to 0, 1^*$ prefix-free description language for $\mathcal{H}$ and $|h|$ is the length of a sentence. Then, for all (things):

$$L_D(h) \leq L_S(h) + \sqrt{\frac{|h| + ln(2/\delta)}{2m}}$$

This theorem suggests that the loss depends also on the hypothesis size or length. This yields the Minimum Description Length paradigm.

7.3.1 Occam's Razor

Theorem 7.7. suggests when having two hypotheses with same risk, one with shorter description will be bounded by a lower value. However, this does not always stand to prefer the shorter hypotheses.