7 Nonuniform Learnability

7.1 NONUNIFORM LEARNABILITY

1. h is (ϵ,δ)-competitive with another hypothesis h' if $\mathbb{P}\{L_{\mathcal{D}}(h) \le L_{\mathcal{D}}(h') + \epsilon\} \ge 1-\delta$
2. **nonuniformly learnable** :

   $\exists A, m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}, \forall \epsilon, \delta \in (0,1), \forall h \in \mathcal{H} :$

   $\mathcal{D}^m\{S : L_{\mathcal{D}}(A(S)) \le L_{\mathcal{D}}(h) + \epsilon, |S| > m_{\mathcal{H}}^{NUL}(\epsilon,\delta,h)\} \ge 1-\delta$
3. The difference between aPAC and NL is the question of whether the sample size m may depend on h.
4. NL is a relaxation of aPAC.
5. **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.

   proof

   necessity: use following theorem;

   sufficiency: let $\mathcal{H}_n = \{h \in \mathcal{H} : m_{\mathcal{H}}^{NUL}(1/8, 1/7, h) \le n\}$. Then $\mathcal{H} = \cup_{n\in\mathbb{N}} \mathcal{H}_n$, using the fundamental of statistical learning, $VC(\mathcal{H}_n) < \infty$, and therefore $\mathcal{H}_n$ is agnostic PAC learnable.(If $VC(H_n) = \infty$, then do not exist $m_{\mathcal{H}}^{NUL}(1/8, 1/7, h) \le n$)
6. **theorem** Let $\mathcal{H}$ be a countable union of hypothesis class $\mathcal{H} = \cup_{n\in\mathcal{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ enjoys the uniform convergence property. Then, $\mathcal{H}$ is nonuniformly learnable. (The proof will be given in the next section)
7. Nonuniform learnability is a strict relaxation of agnostic PAC learnability.

7.2 STRUCTURAL RISK MINIMIZATION

1. **denote** $\epsilon_n(m,\epsilon) = min\{\epsilon \in (0,1) : m_{\mathcal{H}_n}^{UC}(\epsilon,\delta) \le m\}$
2. **weight function** : $\omega : \mathbb{N} \to [0,1], \sum_{n=1}^{\infty} \omega(n) \le 1$
3. **theorem** : $\mathcal{H} = \cup \mathcal{H}_n, \mathcal{H}_n\ has\ m_{\mathcal{H}_n}^{UC}. \forall \delta, \mathcal{D}, n, h$

   $\mathcal{D}^m\{S : |L_{\mathcal{D}}(h) - L_S(h)| \le \epsilon_n(m, \omega(n)\cdot\delta)\} \ge 1-\delta$

   proof :

   $\forall h \in \mathcal{H}_n, |L_{\mathcal{D}}(h) - L_S(h)| \le \epsilon_n(m,\delta_n)$

   $\forall h \in \mathcal{H}, \mathcal{D}^m\{S : |L_{\mathcal{D}}(h) - L_S(h)| \le \epsilon_n(m,\omega(n)\cdot\delta)\} \ge 1 - \sum \delta_n \ge 1-\delta$
4. **denote** $n(h) = min\{n : h \in \mathcal{H}_n\}$
5. $\mathcal{D}^m[L_{\mathcal{D}}(h) \le L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\cdot h)]ge1-\delta$, less constraints, higher probability.
6. **Structural Risk Minimizaiton(SRM)** :
   - **prior knowledge** : $\mathcal{H} = \cup_n \mathcal{H}_n, \mathcal{H}_n\ has\ m_{\mathcal{H}_n}^{UC}, \sum \omega(n) \le 1$
   - **input** : training set $S \sim \mathcal{D}^m$, confidence δ
   - **output** : $h \in argmin_{h\in\mathcal{H}}[L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\cdot\delta)]$
7. **theorem** $\omega(n) = \frac{6}{n^2\pi^2}, m_{\mathcal{H}}^{NUL}(\epsilon,\delta,h) \le m_{\mathcal{H}_{n(h)}}^{UC}(\epsilon/2, \frac{6\delta}{(\pi n(h))^2})$

   proof :

   $\mathbb{P}\{L_{\mathcal{D}}(h) \le L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\cdot h)\} \ge 1-\delta$

   if $m \ge m_{\mathcal{H}_{n(h)}}^{UC}(\epsilon/2, \omega(n(h))\delta)$, then $\epsilon_{n(h)}(m, \omega(n(h))\cdot h) \le \epsilon/2$

Uniform convergence $L_S(h) \leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2}$

$L_{\mathcal{D}}(A(S)) \leq L_S(h) + \epsilon_{n(h)}(m, \omega(n(h)) \cdot h) \leq L_{\mathcal{D}}(h) + \epsilon$

8. *No-Free-Lunch-for-Nonuniform-Learnability*

   $\forall\{\mathcal{X}, |\mathcal{X}| = \infty\}$, the class of all binary valued functions over $\mathcal{X}$ is not a countable union of classes of finite VC-dimension.(Exercise 7.5)

   proof : If $\mathcal{H}$ shatters an infinite set . Then, for any sequence of classes $\mathcal{H}_n : n \in \mathbb{N}$ such that $\mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$, there exists some n for which $VCdim(\mathcal{H}_n) = \infty$.

   1. Assume $\exists\{\mathcal{H}_n\}, \mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$, and $\forall\mathcal{H}_n, VCdim(\mathcal{H}_n) < \infty$.
   2. Subproblem: For $K \subseteq \mathcal{X}, |K| = \infty$ we can always construct subsets sequence $\{K_n\}, K_n \subseteq K; \forall K_n, |K_n| > VCdim(\mathcal{H}_n); \forall n \neq m, K_n \cap K_m = \emptyset$.

      subproof :

      First, let $K_1 \subseteq K, |K_1| = VDdim(\mathcal{H}_1) + 1$.

      Second, suppose that $K_1, \ldots, K_{r-1}$ has been chosen, because $|K| = \infty$, we always can choose $K_r \subseteq K \backslash (\cup_{i=1}^{r-1} K_i)$ such that $|K_r| = VCdim(\mathcal{H}_r) + 1$.
   3. The subproblem implies that $\forall n \in \mathbb{N}, \exists f_n \notin \mathcal{H}_n$.
   4. $\exists f = [f_1, f_2, \ldots, f_n], f \in \mathcal{H}, but\ \forall n : f_n \notin \mathcal{H}_n$, which makes a contradiction.

9. $\forall\{\mathcal{X}, |\mathcal{X}| = \infty\}$, there exists no nonuniform learner w.r.t. the class of all deterministic binary classifiers.

10. Assume $VCdim(\mathcal{H}_n) = n$, then $m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) = C\frac{n+log(1/\delta)}{\epsilon^2}$ (Ch6)

    If $\omega(n) = \frac{6}{n^2\pi^2}$, then $m^{NUL}_{\mathcal{H}}(\epsilon, \delta, h) - m^{UC}_{\mathcal{H}_n}(\epsilon/2, \delta) \leq 4C\frac{2log(2n)}{\epsilon^2}$

    The gap between $m^{NUL}_{\mathcal{H}}$ and $m^{UC}_{\mathcal{H}_n}$ increases with the index of the class, which reflecting the value of knowing a good priority order on the hypotheses in $\mathcal{H}$.

7.3 MINIMUM DESCRIPTION LENGTH AND OCCAM'S RAZOR

1. Let $\mathcal{H}$ be a countable hypothesis class. Then $\mathcal{H}$ can be rewrited as $\mathcal{H} = \cup_{n\in\mathbb{N}}\{h_n\}$, each singleton classes has the uniform convergence property with rate $m^{UC}(\epsilon, \delta) = \frac{log(2/\delta)}{2\epsilon^2}$, and $\epsilon_n(m, \delta) = \sqrt{\frac{log(2/\delta)}{2m}}$. SRM rule becomes:
   - $argmin_{h_n\in\mathcal{H}}[L_S(h) + \sqrt{\frac{-log(\omega(n))+log(2/\delta)}{2m}}]$
   - $argmin_{h_n\in\mathcal{H}}[L_S(h) + \sqrt{\frac{-log(\omega(h))+log(2/\delta)}{2m}}]$
2. **the description of h** : Fix some finite set Σ of symbols, the description function $d = \mathcal{H} \rightarrow \Sigma^* \subseteq \Sigma$, its length is denoted by $|h|$.
   - σ is always used to represent $d(h)$
   - prefix-free
3. **Kraft Inequality** : If $S \subseteq \{0, 1\}^*$ is a prefix-free set of strings, then $\sum\limits_{\sigma\in S}\frac{1}{2^{|\sigma|}} \leq 1$
4. $\omega(h) = \frac{1}{2^{|h|}}$
5. **theorem** : $\mathcal{H}$, prefix-free description language $d : \mathcal{H} \rightarrow \{0, 1\}^*$, then

   $\mathcal{D}^m\{\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{|h|+ln(2/\delta)}{2m}}\} \geq 1 - \delta$
6. **Minimum Description Length (MDL)**
   - **prior knowledge** :
     - $\mathcal{H}$ is a countable bypothesis class

        - $\mathcal{H}$ is described by a prefix-free language over $\{0,1\}$
        - For every $h \in \mathcal{H}$, $|h|$ is the length of the representation of h
    - **input** : A training set $S \sim D^m$, confidence δ
    - **output** : $h \in argmin_{h\in\mathcal{H}}[L_S(h) + \sqrt{\frac{|h|+ln(2/\delta)}{2m}}]$
7. Pre theorem conveys a philosophical message : A short explanation (that is, a hypothesis that has a short length) tends to be more valid than a long explanation.
8. The more complex a hypothesis h is , the larger the sample size it has to fit to guarantee that it has a small true risk $L_{\mathcal{D}}(h)$.
9. Choosing a description language (or, equivalently, some weighting of hypotheses) is a weak form of committing to a hypothesis.
10. Rather than committing to a single hypothesis, we spread out our commitment among many.
11. As long as it is done independently of the training sample, our generalization bound holds.
12. Just as the choice of a single hypothesis to be evaluated by a sample can be arbitrary, so is the choice of description language.

7.4 OTHER NOTIONS OF LEARNABLITY-CONSISTENCY

1. Weak consistency : convergence in probablity
2. Strong consistency: sure convergence
3. **Definition** (Consistency) : A learning rule A is consistent w.r.t. $\mathcal{H}$ and $\mathcal{P}$

   domain set Z, probability distributions set $\mathcal{P}$,

   $\exists m_{\mathcal{H}}^{CON} : (0,1)^2 \times \mathcal{H} \times \mathcal{P} \to \mathbb{N}$ such that, $\forall \epsilon, \delta \in (0,1), \forall h \in \mathcal{H}, \forall \mathcal{D} \in \mathcal{P}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h, \mathcal{D})$ then $\mathcal{D}^m\{S|L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon\} \geq 1 - \delta$
4. If $\mathcal{P}$ is the set of all distributions, then A is universally consistent w.r.t. $\mathcal{H}$.
5. **Memory algorithm** : memorize the training examples, and, given a test point x, it predicts the majority label among all labeled instances of x that exist in the training sample.

   Not nonuniformly learnable, but universally consistent for every countable domain $\mathcal{X}$ and a finite label set $\mathcal{Y}$. (exercise 7.6)

   proof :

   1. Let $\{x_i : i \in \mathbb{N}\}$ be an enumeration of the elements of $\mathcal{X}$, and $i \leq j \Rightarrow \mathcal{D}(x_i) \geq \mathcal{D}(x_j)$.
   2. It's easy to verify $\lim\limits_{n\to\infty} \sum\limits_{i\geq n}^{\infty} \mathcal{D}(x_i) = 0(S_n \to 1 \Rightarrow S - S_n \to 0)$.
   3. $\forall \eta > 0, \exists N \in \mathbb{N}$ such that $\forall i > N, \mathcal{D}(\{x_i\}) < \eta$, then

      $\mathbb{P}_{S\sim\mathcal{D}^m}[\exists x_i : \mathcal{D}(\{x_i\}) > \eta, x_i \notin S] \leq \sum\limits_{i=1}^{N} \mathbb{P}[x_i \notin S] \leq N(1-\eta)^m \leq Ne^{-\eta m}$
   4. $\forall \epsilon > 0, \exists N \in \mathbb{N}$ such that $\sum_{n\geq N} \mathcal{D}(\{x_n\}) < \epsilon$, which also means that $\forall n > N, \mathcal{D}(\{x_n\}) < \epsilon$.

      Let $\eta = \mathcal{D}(\{x_n\})$, then, $\forall k \in [N], \mathcal{D}(\{x_k\}) \geq \eta$.

      $\mathbb{P}_{S\sim\mathcal{D}^m}[\mathcal{D}(\{\exists x_i : x_i \notin S\}) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[\exists x_i \in [N] : x_i \notin S] \leq Ne^{-\eta m}$

7.5 DISCUSSING THE DIFFERENT NOTIONS OF LEARNABLITY

1. What Is The Risk of the Learned Hypothesis?

- PAC learning and nonuniform learning gives us an upper bound on the true risk of the learned hypothesis based on its empirical risk.
- Consistency guarantees do not provide such a bound, but estimate the risk of the output predictor using a validation set.(Ch11)

2. How Many Examples Are Required to Be as Good as the Best Hypothesis in $\mathcal{H}$?
   - PAC learning gives a crisp answer
   - nonuniform learning this number depends on the best hypothesis in $\mathcal{H}$
   - consistency it also depends on the underlying distribution
   - In this sense, PAC learning is the only useful definition of learnability
   - If $\mathcal{H}$ has a large approximation error, PAC's risk may still be large. This reflects the fact that the usefulness of PAC learning relies on the quality of our prior knowledge.
   - If PAC fails, we change the $\mathcal{H}$.
   - If nonuniform algorithm fails, we change a different weighting function.
3. How to Learn? How to Express Prior Knowledge?
   - The definition of PAC learning yields the limitation of learning(via the No-Free-Lunch theorem) and the necessity of prior knowledge.
     - Choose $\mathcal{H}$ by prior knowledge.
     - $ERM_{\mathcal{H}}$
   - nonuniform learnability
     - Encode prior knowledge by specifying weights over(subsets of) hypothesis of $\mathcal{H}$.
     - SRM (pays *estimation error* and *do not know the low bound of m*).
   - consistent algorithm
     - Does not yield a natural learning paradigm or a way to encode prior knowledge.
     - In fact, in many cases there is no need for prior knowledge at all.(*Memorize algorithm*).
     - Weak requirement
4. Which Learning Algorithm Should We Prefer?
   - w.r.t. the set of all functions from $\mathcal{X} \to \mathcal{Y}$, which gives us a guarantee that for enough training examples, we will always be as good as the Bayes Optimal predictor.
   - problems:
     - the sample complexity of the consistent algorithm, non enough examples
     - it's not very hard to make any PAC or nonuniform learner consistent:

       Firstly, we run nonuniform learned predictor, obtain the bound on the true risk;

       Then, if the bound is small enough we are done, otherwise, we revert to Memorize algorithm.
5. The "contradiction" between "No-Free-Lunch" and "Memory algorithm"
   - **No-Free-Lunch** : Let $\mathcal{X}$ be a countable infinite domain and let $\mathcal{Y} = \{\pm 1\}$, then for $\forall A$, and a training set size, m, $\exists \mathcal{D}, h^* : \mathcal{X} \to \mathcal{Y}$, A is likely to return a classifier with a large error.
   - **Memorize algorithm** : $\forall \mathcal{D}, h^* : \mathcal{X} \to \mathcal{Y}, then \ \exists m$, memorize algorithm is likely to return a classifier with a small error.