ECE239AS: Foundations of Statistical Learning Spring 2018-19

Lecture 14: Structural Risk Minimization and non-uniform learnability

Lecturer: Suhas Diggavi *Scribe: Fraile, Rebanal, 06, 02, 2019*

14.1 Introduction

When $VCdim(H) = \infty$ the class is not PAC learnable, thus we require more relax, weaker notions of learnability. We will provide a characterization of nonuniform learnability and show that it is a strict relaxation of agnostic PAC learnability. We also show that H being a countable union of hypothesis classes, each of which enjoys the uniform convergence property, is a sufficient condition for nonuniform learnability. We will prove this results by introducing a new learning paradigm called Structural Risk Minimization (SRM).

14.2 Non-uniform learnability

Non-uniform learnability permits using nonuniform sampling sizes with respect to the different hypotheses with which the learner is competing. We will call a hypothesis h (ϵ, δ) -competitive with another hypothesis $h^{'}$ if with probability higher than $(1 - \delta)$,

$$L_P(h) \leq L_P(h^{'}) + \epsilon,$$

for $P : (X \times Y) \rightarrow [0, 1]$. Note that this is independent of dataset D. This idea of competitiveness is of no interest in PAC learnability as there one is looking for a hypothesis with an absolute low risk, or a low risk compared to the minimal risk achieved by hypotheses in H, the latter in the agnostic case. In effect, in PAC learning $\forall N \geq N_H^{ul}(\epsilon, \delta)$ with probability $1 - \delta$,

$$L_P(h_D) \leq min_{h^{'} \in H} L_P(h^{'}) + \epsilon.$$

In other words, ERM h_D is (ϵ, δ)-competitive to h^*, then

$$L_P(h_D) \leq L_P(h) + \epsilon \ \forall h \in H.$$

We use the idea of (ϵ, δ)-competitive hypotheses and let N depend on $h^{'} \in H$ that H is competitive with.

Definition: Non-uniform learnability *A hypothesis class H is nonuniformly learnable if there exist a learning algorithm, A, and a function $N_H^{NUL}(\epsilon, \delta, h) : (0, 1)^2 \times H \rightarrow N$ such that, for every $\epsilon, \delta \in (0, 1)$ and for every $h \in H$, if $N \geq N_H^{NUL}(\epsilon, \delta, h)$ then for every distribution P, with probability of at least $1 - \delta$, being h_D the output of $A(D)$, it holds that:*

$$L_P(h_D) \leq L_P(h) + \epsilon.$$

We will now use the following two equivalent theorems to characterize Non-uniform learnability, these two theorems are equivalent in the sense that one holds if and only if the other does too. We will first introduce both theorems, use **Theorem** 14.2 to prove **Theorem** 14.1 and subsequently prove **Theorem** 14.2.

Theorem 14.1. *An hypothesis class H is Non-uniformly learnable if and only if it is the countable union of agnostic PAC learnable cases, i.e., $H = \cup_{n\in\mathbb{N}} H_n$ with H_n agnostically PAC learnable.*

Theorem 14.2. *Let H be a hypothesis class that can be written as a countable union of hypothesis classes, $H = \cup_{n\in\mathbb{N}} H_n$, where each H_n has the uniform convergence property (i.e. it is agnostic PAC learnable). Then, H is nonuniformly learnable.*

***Proof: Theorem** 14.1.* First assume that $H = \cup_{n\in\mathbb{N}} H_n$ where each H_n is agnostic PAC learnable. Using the fundamental theorem of statistical learning, it follows that each H_n has the uniform convergence property. Therefore, using **Theorem** 14.2 we obtain that H is nonuniform learnable.

For the other direction, assume that H is nonuniform learnable using some algorithm A. For every $n \in N$, let $H_n = \{h \in H : N_H^{NUL}(\frac{1}{8}, \frac{1}{7}, h) \leq n\}$. Clearly, $H = \cup_{n\in\mathbb{N}} H_n$. In addition, using the definition of N_H^{NUL} we know that for any distribution D that satisfies the realizability assumption with respect to H_n, with probability of at least $6/7$ over $P \sim D^n$ we have that $L_D(A(P)) \leq \frac{1}{8}$. Using the fundamental theorem of statistical learning, this implies that the VC dimension of H_n must be finite, and therefore H_n is agnostic PAC learnable. □

14.2.1 Structural Risk Minimization

To prove **Theorem** 14.2 we will first introduce a new learning paradigm called **Structural Risk Minimization (SRM)** through the following two theorems:

Theorem 14.3. *Let the hypothesis class $\mathcal{H}$ be a union of hypothesis classes $\mathcal{H}_n$ that each satisfy the uniform convergence property with sample complexity function $N_{\mathcal{H}_n}^{UC}(\epsilon, \delta)$, i.e. $\mathcal{H} = \bigcup_{n\in\mathbb{N}} \mathcal{H}_n$. Let the weight function $w(n) \in [0, 1]$ and $\sum_{n\in\mathbb{N}} w(n) \leq 1$. Let the function $\epsilon_n(N, \delta) = \min\{\epsilon \in (0, 1) : N_{\mathcal{H}_n}^{UC}(\epsilon, \delta) \leq N\}$. Then, for every $\delta \in (0, 1)$ with probability at least $(1 - \delta)$ over a distribution $\mathcal{D} \sim \mathcal{P}^N$ for all $h \in \mathcal{H}_n$, for all $n \in \mathbb{N}$:*

$$|L_{\mathcal{P}}(h) - \hat{L}_{\mathcal{D}}(h)| \leq \epsilon_n(N, w(n) \cdot \delta) \tag{14.1}$$

This implies that, for every $\delta \in (0, 1)$ and distribution $\mathcal{D}$, with probability at least $(1 - \delta)$,

$$\forall h \in \mathcal{H}, |L_{\mathcal{P}}(h) - \hat{L}_{\mathcal{D}}(h)| \leq \min_{n:h\in\mathcal{H}} \epsilon_n(N, w(n) \cdot \delta) \tag{14.2}$$

.

Proof. To begin, $\forall n$ define $\delta_n = w(n) \cdot \delta$. Then, by applying the assumption that uniform convergence holds for all n with the rate given by **Theorem** 14.2 and by fixing n, with probability of at least $(1 - \delta)$ and for a distribution $\mathcal{D} \sim \mathcal{P}^N$, we have $\forall h \in \mathcal{H}_n$

$$|L_{\mathcal{P}}(h) - \hat{L}_{\mathcal{D}}(h)| \leq \epsilon_n(N, \delta_n). \tag{14.3}$$

By considering the complementary event of (14.3) can be used to start a sequence of upper bounds that concludes with proving the above **Theorem** 14.3:

$$\mathbb{P}_{\mathcal{D}\sim\mathcal{P}^n}(\bigcup_{n\in\mathbb{N}}\{\exists\ h\in\mathcal{H}_n : |L_{\mathcal{P}}(h)-\hat{L}_{\mathcal{D}}(h)|>\epsilon_n(N,\delta_n)\})$$

$$\begin{aligned}
&\leq^{(a)} \sum_{n\in\mathbb{N}} \mathbb{P}_{\mathcal{D}\sim\mathcal{P}^n}(|L_{\mathcal{P}}(h)-\hat{L}_{\mathcal{D}}(h)|>\epsilon_n(N,\delta_n)\\
&\leq^{(b)} \sum_{n\in\mathbb{N}} \delta_n\\
&=^{(c)} \sum_{n\in\mathbb{N}} w(n)\cdot\delta\\
&=^{(d)} \delta\sum_{n\in\mathbb{N}} w(n)\\
&\leq^{(e)} \delta
\end{aligned}$$

The complementary event of (14.3) is first upper bounded using the union bound in (a).

Now if we assume (14.3), then the probability of this complementary event must be δ_n in (b).

Decomposing δ_n to $w(n)\cdot\delta$ in (c) facilitates clarity in the following steps. Step (d) is simply pulling out δ from the sum since δ no longer depends on n.

Finally, by using the assumption in (14.1) that $\sum_{n\in\mathbb{N}} w(n)\leq 1$, (e) arrives at the upper bound of the probability of (14.3)'s complementary event. Therefore, the probability of (14.3) occurring given its assumptions must be at least $(1-\delta)$, which is exactly what is claimed in (14.1). □

Using this result, we can begin to define the Structural Risk Minimization paradigm. But first, we define a new variable

$$n_h = \min\{n : h\in\mathcal{H}_n\}. \tag{14.4}$$

By combining (14.4) and (14.1), we obtain

$$\forall h\in\mathcal{H}, |L_{\mathcal{P}}(h)-\hat{L}_{\mathcal{D}}(h)|\leq \min_{n:h\in\mathcal{H}_n}\{\epsilon_n(N,w(n)\cdot\delta)\}\leq \epsilon_{n_h}(N,w(n_h)\cdot\delta). \tag{14.5}$$

Following directly from this, we can rearrange the inequality in (14.5) to not be in terms of $L_{\mathcal{P}}(h)$ since we can not directly control this loss value as it is dependent on an unknown distribution $\mathcal{P}$. The result is the Structural Risk Minimization (SRM) paradigm that is used for finding a structural risk minimizer h:

$$h_{\mathcal{D}}^{SRM} = arg\,min_{h\in\mathcal{H}}\{\hat{L}_{\mathcal{D}}(h)+\epsilon_{n_h}(N,w(n_h)\cdot\delta\} \tag{14.6}$$

Compared to the Empirical Risk Minimization paradigm, SRM spreads its bias across both low empirical error $\hat{L}_{\mathcal{D}}(h)$ and classes with smaller differences $\epsilon_{n_h}(N,w(n_h)\cdot\delta$ between the empirical and true error. The latter term of the h^{SRM} paradigm acts like a regularization term.

An important result that follows is that the SRM paradigm can be used for nonuniform learning (NUL) of every class, which can be expressed as a countable union of uniformly converging (UC) hypothesis classes.

Theorem 14.4. *Let $\mathcal{H}$ be a hypothesis class that can be expressed as a union of uniformly converging hypothesis classes $\mathcal{H}_n$ each with sample complexity $N^{UC}_{\mathcal{H}_n}(\epsilon, \delta)$ (defined as in **Theorem** 14.3), i.e. $\mathcal{H} = \bigcup_{n\in\mathbb{N}} \mathcal{H}_n$. Let $w(n) = \frac{6}{n^2\pi^2}$. Then, we get that $\mathcal{H}$ is nonuniformly learnable using the SRM paradigm (14.6) with sample complexity:*

$$N^{NUL}_{\mathcal{H}}(\epsilon, \delta, h) \leq N^{UC}_{\mathcal{H}_{n_h}}(\frac{\epsilon}{2}, \frac{6\delta}{(\pi n_h)^2}). \tag{14.7}$$

It is then clear that **Theorem** 14.4 also happens to prove **Theorem** 14.2.

Proof. Let $N \geq N^{UC}_{\mathcal{H}_{n_h}}(\epsilon, w(n)\cdot\delta)$. Then, using **Theorem** 14.3 and its assumption that $\sum_n w(n) \leq 1$, $\forall h' \in \mathcal{H}$,

$$L_{\mathcal{P}}(h') \leq \hat{L}_{\mathcal{D}}(h') + \epsilon_{n_{h'}}(N, w(n_{h'}) \cdot \delta) \tag{14.8}$$

with probability at least $(1-\delta)$ for $\delta \in (0,1)$ and for a dataset $\mathcal{D} \sim \mathcal{P}^N$.

In particular, (14.8) holds for the structural risk minimizer,

$$h^{SRM}_{\mathcal{D}} = arg\, min_{h\in\mathcal{H}}\{\hat{L}_{\mathcal{D}}(h) + \epsilon_{n_h}(N, w(n_h)\cdot\delta)\} \tag{14.9}$$

By combining (14.8) and (14.9), $\forall h \in \mathcal{H}$,

$$\begin{aligned} L_{\mathcal{P}}(h^{SRM}_{\mathcal{D}}) \leq& \hat{L}_{\mathcal{D}}(h^{SRM}_{\mathcal{D}}) + \epsilon_{n_{h^{SRM}_{\mathcal{D}}}}(N, w(n_{h^{SRM}_{\mathcal{D}}})\cdot\delta) \\ \leq& \hat{L}_{\mathcal{D}}(h) + \epsilon_{n_h}(N, w(n_h)\cdot\delta) \end{aligned} \tag{14.10}$$

Next, if $N \geq N^{UC}_{\mathcal{H}_{n_h}}(\frac{\epsilon}{2}, w(n_h)\cdot\delta)$, then

$$\epsilon_{n_h}(N, w(n_h) \leq \epsilon/2. \tag{14.11}$$

By using (14.11) and (14.10), we obtain

$$L_{\mathcal{P}}(h^{SRM}_{\mathcal{P}}) \leq L_{\mathcal{D}}(h) + \epsilon/2. \tag{14.12}$$

Further, because each $\mathcal{H}_n$ has the uniform convergence property, for $N \geq N^{UC}_{\mathcal{H}_{n_h}}(\frac{\epsilon}{2}, w(n_h)\cdot\delta)$ and with probability at least $(1-\delta)$,

$$
\begin{aligned}
|L_{\mathcal{P}}(h) - \hat{L}_{\mathcal{D}}(h)| \leq & \epsilon/2 \\
\Rightarrow \hat{L}_{\mathcal{D}}(h) \leq & L_{\mathcal{P}}(h) + \epsilon/2
\end{aligned} \tag{14.13}
$$

Finally, by using (14.13) and (14.12), we get if $N \geq N^{UC}_{\mathcal{H}_{n_h}}(\frac{\epsilon}{2}, w(n_h) \cdot \delta)$, then

$$
L_{\mathcal{P}}(h^{SRM}_{\mathcal{P}}) \leq L_{\mathcal{P}}(h) + \epsilon \tag{14.14}
$$

with probability at least $(1 - \delta)$.

□

Minimum Description Length

As before, let $\mathcal{H}$ be a countable hypothesis class expressable as a countable union of singleton classes h_n, i.e. $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \{h_n\}$. Using Hoeffding's inequality, each h_n has the uniform convergence property with rate $N^{UC}(\epsilon, \delta) = \frac{\log(2/\delta)}{2\epsilon^2}$.

Hoeffding's inequality. *Let Θ_1, ..., Θ_m be a sequence of i.i.d. random variables and assume that for all i, $\mathcal{E}[\Theta_i] = \mu$ and $\mathcal{P}[a \leq \Theta_i \leq b] = 1$. Then, for any $\epsilon > 0$,*

$$
\mathcal{P}[|\frac{1}{m}\sum_{i=1}^{m} \Theta_i - \mu| > \epsilon] \leq 2\exp(-2m\epsilon^2/(b-a)^2). \tag{14.15}
$$

Hence, $\epsilon_n = \sqrt{\frac{\log(2/\delta)}{2N}}$. Using the SRM rule with $\delta_n = w(n) \cdot \delta$ gives

$$
\begin{aligned}
h^{SRM}_{\mathcal{D}} = & arg\, min_{h_n \in \mathcal{H}} \{\hat{L}_{\mathcal{D}}(h_n) + \sqrt{\frac{\log(2/\delta_n)}{2N}}\} \\
= & arg\, min_{h_n \in \mathcal{H}} \{\hat{L}_{\mathcal{D}}(h_n) + \sqrt{\frac{-\log(w(n)) + \log(2/\delta)}{2N}}\}
\end{aligned} \tag{14.16}
$$

This formulation of the structural risk minimizer $h^{SRM}_{\mathcal{D}}$ prefers hypotheses h_n with higher weights $w(n)$. A natural train of thought to follow is deciding how to design the weights in such a way that hypotheses we believe are more likely to be correct are assigned higher weights.

We turn to understanding minimum description lengths of hypotheses as an approach to assigning weights to hypotheses. Begin by letting $\mathcal{H}$ be the hypothesis class we want to describe. Then, fix a finite set Σ to represent an alphabet, the elements of which are characters. For example, in forming binary strings, the corresponding alphabet $\Sigma = \{0, 1\}$ and a string could be formed as $\sigma = (1, 1, 0, 0, 0, 1, 0)$ which corresponding length $|\sigma| = 7$.

Let Σ^* denote the set of all finite length strings for some alphabet Σ. Then a description language for $\mathcal{H}$ is a function $\mathcal{C} : \mathcal{H} \to \Sigma^*$. In other words, a description of $\mathcal{H}$ that maps each member h of $\mathcal{H}$ to a string $\mathcal{C}(h)$ is a way to describe different hypotheses using the same set of elements.

An important, desirable property of such description languages d is that they are prefix-free. That is, for every distinct hypothesis h and h' in $\mathcal{H}$, $\mathcal{C}(h)$ is not a prefix of $\mathcal{C}(h')$. Such description languages enjoy unique, lossless decodability of each description $\mathcal{C}(h)$. Further, prefix-free collections of strings enjoy the following combinatorial property:

Kraft's inequality. *Let Σ^* be the set of all finite length strings with alphabet $\{0, 1\}$. If we let $\mathcal{S} \subseteq \Sigma^*$ and σ be a string from $\mathcal{S}$, then*

$$\sum_{\sigma \in \mathcal{S}} \frac{1}{2^{|\sigma|}} \leq 1. \tag{14.17}$$

Proof. Define a probability distribution over the elements of $\mathcal{S}$ as follows: Repeatedly toss an unbiased coin, with faces labeled 0 and 1, until the sequence of outcomes is a member of $\mathcal{S}$, then stop. For each string $\sigma \in \mathcal{S}$, let $P(\sigma)$ be the probability that this process generates the string σ. Note that since $\mathcal{S}$ is prefix-free, for every string $\sigma \in \mathcal{S}$, if the coin toss outcomes follow the bits (outcomes) of σ then we will stop only once the sequence of outcomes equals σ.

We therefore obtain that for every string $\sigma \in \mathcal{S}, P(\sigma) = \frac{1}{2^{|\sigma|}}$. Finally, these probabilities add up to at most 1. □

If we now apply (14.17) to our definition of the weights, we have

$$\begin{aligned} w(h) =& \frac{1}{2^{|\mathcal{C}(h)|}} \\ \Rightarrow \sum_h w(h) =& \sum_h \frac{1}{2^{|\mathcal{C}(h)|}} \leq 1. \end{aligned} \tag{14.18}$$

The final result being the Structural Risk Minimization paradigm that prioritizes hypotheses with minimum description length:

$$h_{\mathcal{D}}^{SRM} = arg\, min_{h \in \mathcal{H}} \{\hat{L}_{\mathcal{D}}(h) + \sqrt{\frac{|\mathcal{C}(h)| + \log(2/\delta)}{2N}}\} \tag{14.19}$$