Uniform Convergence and No Free Lunch

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 18

recall: PAC learning

A hypothesis class $\mathcal{H}$ is said to be **PAC learnable** if there exists a function $m_{\mathcal{H}}: (0,1)^2 \rightarrow \mathbb{N}$ and a learning algorithm with the following property:

- For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f: \mathcal{X} \rightarrow \{0,1\}$, if the realizability assumption holds w.r.t. $\mathcal{H}, \mathcal{D}, f$, then
- when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypotheses h such that,
- with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

sample complexity

- $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ is called the sample complexity
- In previous section, we have shown:
  - Every finite hypothesis class is PAC learnable with sample complexity
$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$$
  - We will soon see in this course that "finite"-ness is not essential here.

agnostic PAC learning

- In practice, the realizability assumption is too restrictive. We want to release this in our definition.

- Now consider the data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$: pairs of a data point and a label (instead of over the data point only)

$$L_{\mathcal{D}}(h) = P_{(x,y)\sim\mathcal{D}}[h(x) \neq y] =: \mathcal{D}(\{(x,y): h(x) \neq y\})$$

(compare with what we had before:

$$L_{\mathcal{D},f}(h) = P_{x\sim\mathcal{D}}[h(x) \neq f(x)] =: \mathcal{D}(\{x \in \mathcal{X}: h(x) \neq f(x)\})$$

)

agnostic PAC learning

A hypothesis class $\mathcal{H}$ is said to be **agnostic PAC learnable** if there exists a function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:

- For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, ~~and for every labeling function, if the realizability assumption holds,~~
- when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ ~~and labeled by the labeling function,~~ the algorithm returns a hypotheses h such that,
- with probability of at least $1 - \delta$ (over the choice of the examples),
  $\cancel{L_{(\mathcal{D},f)}(h) \leq \epsilon}$

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

agnostic PAC learning

A hypothesis class $\mathcal{H}$ is said to be **agnostic PAC learnable** if there exists a function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:

- For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,
- when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypotheses h such that,
- with probability of at least $1 - \delta$ (over the choice of the examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

example: Bayes optimal predictor

- Given any $\mathcal{D}$ over $\mathcal{X}\times\{0,1\}$, the best label predicting function will be (the naïve Bayes' classifier in Lecture 1)

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}[y = 1|x] \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

- We can prove (HW exercise) that this classifier is the best possible in the sense that $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$ for any other classifier g. That is,

$$f_{\mathcal{D}} = \operatorname*{argmin}_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$$

- However, since we don't know $\mathcal{D}$, we don't know $f_{\mathcal{D}}$. Therefore, we can only hope to be "approximately optimal":

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon = L_{\mathcal{D}}(f_{\mathcal{D}}) + \epsilon$$

beyond binary classification

- We can generalize the definition further by considering a general loss function: $\ell: \mathcal{H} \times Z \to \mathbb{R}_+$.
- This Z is a general set:
  - We used to take $Z = \mathcal{X} \times \mathcal{Y}$.
  - For instance, in unsupervised tasks, we could have $Z = \mathcal{X}$.
- Now we need to consider data distributions for this Z. Given this loss function ℓ, the error w.r.t. the data distribution $\mathcal{D}$ will be

$$L_{\mathcal{D}}(h) := \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$$

beyond binary classification

A hypothesis class $\mathcal{H}$ is said to be **agnostic PAC learnable w.r.t.** a set Z and a loss function $\ell: \mathcal{H} \times Z \to \mathbb{R}_+$, if there exists a function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property:

- For <u>every</u> $\epsilon, \delta \in (0,1)$, for <u>every</u> distribution $\mathcal{D}$ over Z, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypotheses h such that,
- with probability of at least $1 - \delta$ (over the choice of the examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

- where $L_{\mathcal{D}}(h) := \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$.

learning via uniform convergence

In this section, we give a sufficient condition for agnostic PAC learnable.

ϵ-representative sample

- A training set S is called **ϵ-representative** if for any $h \in \mathcal{H}$,

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$$

ϵ-representative sample

- A training set S is called **ϵ-representative** if for any $h \in \mathcal{H}$,

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$$

- Assume S is $\epsilon/2$-representative, then any output $h_S \in \mathrm{argmin}_{h\in\mathcal{H}} L_S(h)$ satisfies $L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$

Why? Let h^* be the argmin of $L_{\mathcal{D}}(h)$.

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h^*) + \frac{\varepsilon}{2} \leq L_{\mathcal{D}}(h^*) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$

$$= L_{\mathcal{D}}(h^*) + \varepsilon$$

uniform convergence property (UCP)

We say that a hypothesis class $\mathcal{H}$ has the **uniform convergence property (UCP)** if there exists a function $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ such that

- for every $\epsilon, \delta \in (0,1)$ and for every $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta)$ examples sampled i.i.d according to $\mathcal{D}$, then,
- with probability of at least $1 - \delta$, S is ϵ-representative.

uniform convergence property (UCP)

We say that a hypothesis class $\mathcal{H}$ has the **uniform convergence property (UCP)** if there exists a function $m_{\mathcal{H}}^{\text{UC}} : (0,1)^2 \to \mathbb{N}$ such that

- for <u>every</u> $\epsilon, \delta \in (0,1)$ and for <u>every</u> $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples sampled i.i.d according to $\mathcal{D}$, then,
- with probability of at least $1 - \delta$, S is ϵ-representative.

$$\boxed{\text{UCP} \Rightarrow \text{agnostic PAC}}$$

by taking $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\text{UC}}\left(\frac{\epsilon}{2}, \delta\right)$

finite classes are agnostic PAC learnable

(Assume the range of the loss function is [0,1])

- To show that finite classes are also agnostic PAC learnable, we will first use a union bound (as before) and then apply a concentration inequality.

- We need to show that most of time, our choice of training sample is lucky: 😁

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

- Equivalently, we need to show that rarely, out choice of training sample is unlucky: 😰

$$\mathcal{D}^m(\{S: \text{there exists } h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta$$

finite classes are agnostic PAC learnable

- Note that $\mathcal{D}^m(\{S: \text{there exists } h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$

$$= \mathcal{D}^m\left(\bigcup_{h \in \mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right)$$

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

- This concludes the first step (union bound).

finite classes are agnostic PAC learnable

- The second step requires us to bound $\mathcal{D}^m(\{S: |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$.
- In other words, the two quantities $\frac{1}{m}\sum_{i=1}^{m}[\ell(h, z_i)]$ and $\mathbb{E}_{z\sim\mathcal{D}}[\ell(h, z)]$ should not be far from each other.

finite classes are agnostic PAC learnable

- The second step requires us to bound $\mathcal{D}^m(\{S: |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$.
- In other words, the two quantities $\frac{1}{m}\sum_{i=1}^{m}[\ell(h, z_i)]$ and $\mathbb{E}_{z\sim\mathcal{D}}[\ell(h, z)]$ should not be far from each other.

LEMMA (Hoeffding's Inequality) *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all i, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon > 0$*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2\exp\left(-2\,m\,\epsilon^2/(b-a)^2\right).$$

finite classes are agnostic PAC learnable

- The second step requires us to bound $\mathcal{D}^m(\{S: |L_S(h) - L_\mathcal{D}(h)| > \epsilon\})$.
- In other words, the two quantities $\frac{1}{m}\sum_{i=1}^{m}[\ell(h, z_i)]$ and $\mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$ should not be far from each other.

LEMMA (Hoeffding's Inequality) *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all i, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon > 0$*

Let $\theta_i = \ell(h, z_i)$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2\exp\left(-2\,m\,\epsilon^2/(b-a)^2\right).$$

finite classes are agnostic PAC learnable

- Therefore, combining the first step

$$\mathcal{D}^m(\{S: \text{there exists } h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \le \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S: |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})$$

with the second step

$$\mathcal{D}^m(\{S: |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \le 2\exp(-2m\epsilon^2)$$

yields?

$$\le |\mathcal{H}|\, 2\exp(-2m\varepsilon^2) \overset{\text{set}}{\le} \delta$$

That is, $-2m\varepsilon^2 \le \log\left(\frac{\delta}{2|\mathcal{H}|}\right)$

That is, $m \ge \frac{1}{2\varepsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$

finite classes are agnostic PAC learnable

Corollary. Let $\mathcal{H}$ be a finite hypothesis class, and let $\ell: \mathcal{H} \times Z \to [0,1]$ be a loss function. Then $\mathcal{H}$ enjoys UCP with sample complexity

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

Furthermore, the class is agnostic PAC learnable using the ERM algorithm with sample complexity

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

no free lunch

universal learner

- In previous discussion, we assume that there is a hypothesis class $\mathcal{H}$ which serves as the search space for our model h.
- We then find the ERM $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$
- $\mathcal{H}$ is a prior belief, determined by the task.
- Is this prior belief necessary? Is it possible to have a universal learner that works for any task? Specifically, is there an algorithm that outputs a low-risk h as long as it receives a large number of training data?

universal learner

More specifically, does there exist a learning algorithm A and a training set size m, such that:

- for every distribution $\mathcal{D}$, if A receives m i.i.d. examples from $\mathcal{D}$, there is a high chance it outputs a predictor h with a low risk?

This is impossible ☹

no free lunch (NFL)

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification w.r.t. the 0-1 loss over a domain $\mathcal{X}$. Let m, the size of the training set, be any number with $m < |\mathcal{X}|/2$. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1. There exists a function $f: \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$;
2. With probability of at least $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(h) \geq \frac{1}{8}$ where $h = A(S)$ is the output of the algorithm.

TLDR version: “Any algorithm will fail for some reasonable data distribution.”

no free lunch (NFL)

Theorem (No-Free-Lunch)

Let A be any learning algorithm for the task of binary classification w.r.t. the 0-1 loss over a domain $\mathcal{X}$. Let m, the size of the training set, be any number with $m < |\mathcal{X}|/2$. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1. There exists a function $f: \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$;
2. With probability of at least $\frac{1}{7}$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(h) \geq \frac{1}{8}$ where $h = A(S)$ is the output of the algorithm.

Wordier version: “Every learner fails on some task, though the task can be successfully learned by another learner.”

Questions?

Reference

- *PAC learning :*
  - *[S-S] Ch 2.1-2.3, 3.1*
- *Agnostic PAC learning:*
  - *[S-S] Ch 3.2-3.3, 4.1-4.3*
- *No Free Lunch:*
  - *[S-S] Ch 5.1*