Machine Learning

CSE427

Mahbub Majumdar
Typeset by:
Syeda Ramisa Fariha

BRAC University
66 Mohakhali
Dhaka, Bangladesh

May 28, 2018

Superhumungous Thanks

These slides were typeset by Syeda Ramisa Fariha.

Without her tremendous dedication, these slides would not exist.

Table of Contents

Uniform Convergence Is Sufficient For Learnability

Finite Classes Are Agnostic PAC Learnable

Uniform Convergence Is Sufficient For Learnability

Goal

- Suppose $\mathcal{H}$ is not a bad hypothesis class. Then we want all the $h \in \mathcal{H}$ to have a low empirical risk $L_S(h)$.

- Another way of saying that is that we and the *empirical risk* to be close to the *true risk* for all hypothesis in $\mathcal{H}$.

$\Rightarrow$ This is the gist of *uniform convergence*

Uniform Convergence Is Sufficient For Learnability

Definition: ϵ - representative

A training set S, is called ϵ representative (wrt $\mathcal{H}, D, \ell, Z$) if

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

- When a sample is $\frac{\epsilon}{2}$-representative,

$\Rightarrow$ We will see that the ERM rule is guaranteed to give a *good hypothesis*

Uniform Convergence Is Sufficient For Learnability

$\frac{\epsilon}{2}$ Representative

Lemma: Assume that S is $\frac{\epsilon}{2}$ representative (wrt $\mathcal{H}$, D ,ℓ, Z). Then any $ERM_{\mathcal{H}}(S)$ output, $h_S \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} (L_S(h))$ satisfies

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

Proof:

$$\begin{aligned} L_D(h_S) &\leq L_S(h_S) + \frac{\epsilon}{2} && (1) \\ &\leq L_S(h) + \frac{\epsilon}{2} && (2) \\ &\leq L_D(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} && (3) \\ &= L_D(h) + \epsilon \end{aligned}$$

Lines 1 and 3 follow from S being $\frac{\epsilon}{2}$-representative. Line 2 follows from h_S being ERM.

Uniform Convergence Is Sufficient For Learnability

Definition: Uniform Convergence

$\mathcal{H}$ has the uniform convergence property (wrt Z, ℓ)

- if $\exists\ m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$
- such that $\forall\ \epsilon, \delta \in (0,1)$ and $\forall\ D$ over Z
- if S is a collection of $m \geq m_{\mathcal{H}}^{UC}$ samples drawn iid via D, then with probability of at least $1 - \delta$

$\Rightarrow$ S is ϵ-representative.

- $m_{\mathcal{H}}^{UC}$ is the minimum number of samples to obtain *uniform convergence*, i.e. to ensure with probability $1 - \delta$, that the sample is ϵ-representative
- Uniform refers to having a finite sample size that works for all $h \in \mathcal{H}$ and over all probability distributions D.

Uniform Convergence Is Sufficient For Learnability

Corollary

If a class $\mathcal{H}$ has the *uniform convergence property* with $m_{\mathcal{H}}^{UC}$, then the class is agnostically PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \geq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta)$$

In this case, the $ERM_{\mathcal{H}}$ learner is a successful agnostic PAC learner for $\mathcal{H}$

- From this we will prove that finite hypothesis classes are agnostic PAC learnable

Proof: Finite Classes Are Agnostic PAC Learnable

- First show that *uniform convergence* holds

PART A:

- Fix (ϵ, δ).
- Need to find an m guaranteeing that $\forall\ D$ with probability of at least $1 - \delta$,
- that for $S = \{Z_1, Z_2, \ldots, Z_m\}$ sampled iid from D,

Proof: Finite Classes Are Agnostic PAC Learnable

- and $\forall h \in \mathcal{H}$ that with high liklihood

$$|L_S(h) - L_D(h)| \leq \epsilon$$

- i.e.

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

- or equivalently,

$$D^m\Big(\big\{S \mid \exists h \in \mathcal{H},\ |L_S(h) - L_D(h)| > \epsilon\big\}\Big) <\ \delta.$$

Proof: Finite Classes Are Agnostic PAC Learnable

- Now,

$$\left\{S \;\middle|\; \exists h \in \mathcal{H},\ |L_S(h) - L_D(h)| > \epsilon\right\} = \bigcup_{h \in \mathcal{H}} \left\{S \;\middle|\; |L_S(h) - L_D(h)| > \epsilon\right\}$$

- Applying the *union bound*

$$D^m\Big(\{S \mid \exists h \in \mathcal{H},\ |L_S(h) - L_D(h)| > \epsilon\}\Big) \leq \sum_{h \in \mathcal{H}} D^m\Big(\{S \;\big|\; |L_S(h) - L_D(h)| > \epsilon\}\Big) \quad (4)$$

We Now argue that each right hand term in Equation 4 is small for large enough m.

Finite Classes Are Agnostic PAC Learnable

Proof: Finite Classes Are Agnostic PAC Learnable

PART B:

- Recall that,

$$L_D(h) = \underset{Z \sim D}{\mathbb{E}} \left[\ell(Z, D) \right]$$

and

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, Z_i)$$

- Each Z_i is sampled iid from D

Finite Classes Are Agnostic PAC Learnable

Proof: Finite Classes Are Agnostic PAC Learnable

$\Rightarrow$ Thus

$$\mathbb{E}(\ell(h, Z_i)) \;=\; L_D(h)$$

- Linearity of expectation

$$\begin{aligned}
\mathbb{E}(L_S(h)) &= \frac{1}{m} \sum \mathbb{E}(\ell(h, Z_i)) \\
&= \frac{1}{m} \times m \; L_D(h) \\
&= L_D(h)
\end{aligned}$$

Finite Classes Are Agnostic PAC Learnable

Proof: Finite Classes Are Agnostic PAC Learnable

- Thus, the *deviation* of $L_S(h)$ from its mean is

$$\Delta L_S(h) \equiv \left| L_S(h) - \mathbb{E}(L_S(h)) \right| = \left| L_D(h) - L_S(h) \right|$$

- We want to show the probability of having a significant deviation $\Delta L_S(h)$, is small.
- **Law of large numbers:** As $m \to \infty$, empirical average converges to the true average.
- For finite m, use *Hoeffding's inequality*

Finite Classes Are Agnostic PAC Learnable

Lemma: Hoeffding's inequality

- Let $\theta_1, \ldots, \theta_m$ be iid random variables and assume $\forall i$

$$\begin{aligned} \mathbb{E}(\theta_i) &= \mu \\ \mathbb{P}[a \leq \theta_i \leq b] &= 1 \end{aligned}$$

- then for $\epsilon > 0$,

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

Finite Classes Are Agnostic PAC Learnable

Using Hoeffding's inequality

- Let $\theta_i \equiv \ell(h, Z_i)$
- Since h is fixed and Z_i are iid, $\implies$ θ_i are iid.
- $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \theta_i$
- $L_D(h) = \mu$
- Assume the range of the loss functions, $\ell \in [0, 1]$ and $\theta_i \in [0, 1]$

Finite Classes Are Agnostic PAC Learnable

Using Hoeffding's inequality

- Putting everything together,

$$
\begin{aligned}
&D^m\Big(\Big\{S \;\Big|\; |L_S(h) - L_D(h)| > \epsilon\Big\}\Big) && (5)\\
&= \mathbb{P}\left[\frac{1}{m}\left|\sum_{i=1}^{m}(\theta_i - \mu)\right| > \epsilon\right] && (6)\\
&\leq 2e^{-2m\epsilon^2} && (7)
\end{aligned}
$$

Finite Classes Are Agnostic PAC Learnable

Sample Complexity for APAC

- Inserting Equation 7 into Equation 4

$$
\begin{aligned}
D^m\left(\left\{S \middle| \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon\right\}\right) &\leq \sum_{h \in \mathcal{H}} 2e^{-2m\epsilon^2} \\
&= 2|\mathcal{H}|e^{-2m\epsilon^2}
\end{aligned}
$$

- We want

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

- Thus,

$$m \geq \frac{1}{2\epsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

Finite Classes Are Agnostic PAC Learnable

Comments

- For PAC Learning

$$m \sim \mathcal{O}(\epsilon^{-1})$$

- But for Agnostic PAC learning

$$m \sim \mathcal{O}(\epsilon^{-2})$$

- Thus for APAC learning, need many more examples.
- For example, if $\epsilon = 0.01$, then for APAC, you need 100 times more samples than for PAC Learning.

Finite Classes Are Agnostic PAC Learnable

S is $\epsilon/2$-representative

- Since uniform convergence implies APAC Learnability for S being $\frac{\epsilon}{2}$-representative, we should replace $\epsilon \to \epsilon/2$.
- Then, the sample complexity is

$$m \geq \frac{2}{\epsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

- Then $\mathcal{H}$ is agnostically PAC learnable with this sample complexity.