Learning Theory

Lecture 1

Ruth Urner
York University, Toronto, Canada

SIMLES online summer school

August 17, 2020

Course on Machine Learning Theory

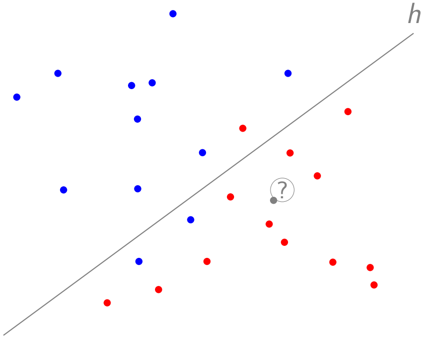
Lecture 1, Monday Introduction to learning theory, notation, warm-up exercises, small historical tour of ML theory

Lecture 2, Tuesday Major concepts and techniques in statistical learning theory: VC-theory and other techniques

Lecture 3, Wednesday Research topics

Lecture 4, Thursday Fireside chat

Machine Learning:



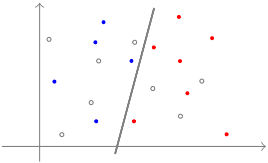
Task:

Given Data S

We fit a function h

Question:
How well will h predict the class of new, unseen datapoints?

Formal framework of learning theory



Instance space: $\mathcal{X} \subseteq \mathbb{R}^d$
Label set: $\mathcal{Y} = \{0, 1\}$
Data: $S = ((x_1, y_1), (x_2, y_2), \dots (x_n, y_n))$
Predictor: $h : \mathcal{X} \to \mathcal{Y}$
Loss function: $\ell(h, x, y) \;=\; \mathbf{1}[h(x) \neq y]$
Empirical risk: $\mathcal{L}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, x_i, y_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[h(x_i) \neq y_i]$
Data generating distribution: P over $\mathcal{X} \times \mathcal{Y}$
Assumption: S is an i.i.d. sample from P^n
Learning algorithm: $\mathcal{A} : S \mapsto h$

Goal: Find a predictor $h = \mathcal{A}(S)$ with small true (i.e. expected) loss over the data generating distribution!

$$\mathcal{L}_P(h) \;=\; \mathop{\mathbb{E}}_{(x,y)\sim P}[\ell(h, x, y)] \;=\; \mathop{\mathbb{E}}_{(x,y)\sim P}[\ell(h, x, y)]$$

Challenge: We can only observe the empirical loss..

Quiz 1

Claim: Let $h : \mathcal{X} \to \{0, 1\}$ be a predictor and let $n \in \mathbb{N}$ denote a sample size. Then, for every data-generating distribution P, we have:

$$\mathop{\mathbb{E}}_{S \sim P^n} [\mathcal{L}_S(h)] \; = \; \mathcal{L}_P(h)$$

(The expectation of the empirical loss is the true loss of h.)

This is..

(a) true for every fixed sample size n

(b) not true for any n

(c) true for large enough n (depending on P)

(d) true for large enough n (depending on h)

(e) only true "in the limit" as $n \to \infty$

Quiz 1–Solution

$$\forall P \underset{S\sim P^n}{\mathbb{E}}[\mathcal{L}_S(h)] \;=\; \mathcal{L}_P(h)$$

Answer: (a) true for every fixed sample size n!

$$\begin{aligned}
\underset{S\sim P^n}{\mathbb{E}}[\mathcal{L}_S(h)] &= \underset{S\sim P^n}{\mathbb{E}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[h(x_i)\neq y_i]\right]\\
&= \frac{1}{n}\sum_{i=1}^{n}\underset{S\sim P^n}{\mathbb{E}}[\mathbf{1}[h(x_i)\neq y_i]]\\
&= \frac{1}{n}\sum_{i=1}^{n}\underset{(x,y)\sim P}{\mathbb{E}}[\mathbf{1}[h(x)\neq y]]\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_P(h)\\
&= \mathcal{L}_P(h)
\end{aligned}$$

Quiz 2

New claim: Let n be a sample size. Then, for all learning algorithms $\mathcal{A} : S \mapsto h$ and for all distributions P, we have

$$\mathop{\mathbb{E}}_{S \sim P^n}[\mathcal{L}_S(\mathcal{A}(S))] = \mathcal{L}_P(\mathcal{A}(S))$$

This is..

(a) true for every fixed sample size n

(b) not true for any n

(c) true for large enough n (depending on P)

(d) true for large enough n (depending on $\mathcal{A}$)

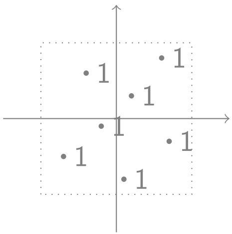
(e) only true "in the limit" as $n \to \infty$

$$\forall \mathcal{A} \forall P \underset{S \sim P^n}{\mathbb{E}} [\mathcal{L}_S(\mathcal{A}(S))] \; = \; \mathcal{L}_P(\mathcal{A}(S))$$

Answer: (b) not true for any n!

In fact, there exists a learner $\mathcal{A}$ and a distribution P, such that for all samples S:

$$|\mathcal{L}_S(\mathcal{A}(S)) \; - \; \mathcal{L}_P(\mathcal{A}(S))| \; = \; 1$$



- $\mathcal{X} = [-1, 1]^2$
- P uniform over $\mathcal{X} \times \{1\}$
- $\Rightarrow$ all samples of the form:
  $S = ((x_1, 1), (x_2, 1) \ldots (x_n, 1))$
- $[\mathcal{A}(S)](x) = \begin{cases} 1 \text{ if } (x, 1) \in S \\ 0 \text{ otherwise} \end{cases}$

Contradiction?

Lesson from quiz 1:

$$\forall h \forall n \forall P \underset{S \sim P^n}{\mathbb{E}} [\mathcal{L}_S(h)] \;=\; \mathcal{L}_P(h)$$

The expectation of the empirical loss is the true loss of h!

Lesson from quiz 2:

$$\exists \mathcal{A} \exists P \forall n : \; | \underset{S \sim P^n}{\mathbb{E}} [\mathcal{L}_S(\mathcal{A}(S))] - \mathcal{L}_P(\mathcal{A}(S)| = 1$$

For some learning algorithms, the empirical loss is not a good indicator of the true loss at all.

What's the difference?

Answer:
In the second claim, the predictor $h = \mathcal{A}(S)$ depends on the data!

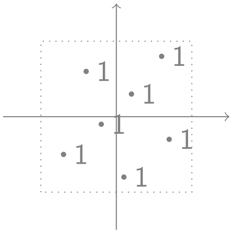
Lesson learned

In Machine Learning, we want the output predictor h to be learned from the data.

For $h = \mathcal{A}(S)$, we can not assume that the empirical loss is close to the true loss.

We need to prevent overfitting to control generalization.

The stubborn learner just memorizes the data...



- $\mathcal{X} = [-1, 1]^2$
- P uniform over $\mathcal{X} \times \{1\}$
- $\Rightarrow$ all samples of the form:
  $S = ((x_1, 1), (x_2, 1) \ldots (x_n, 1))$
- $[\mathcal{A}(S)](x) = \begin{cases} 1 \text{ if } (x, 1) \in S \\ 0 \text{ otherwise} \end{cases}$

$\Rightarrow$ This learner had too much flexibility to adapt its output to the data. One way to prevent overfitting is to restrict the output of the learner to a hypothesis class of bounded capacity.

Success criterion for learnability

Fix a hypothesis class $H \subseteq \{0,1\}^{\mathcal{X}}$.
Best loss in class: $\mathrm{opt}_P(H) = \inf_{h \in H} \mathcal{L}_P(h)$.

Algorithm $\mathcal{A}$ learns H

$\forall\ \epsilon, \delta > 0 \quad \exists\ m(\epsilon, \delta) \in \mathbb{N}$ such that $\forall\ P$

$$\Pr_{S \sim P^{m(\epsilon,\delta)}} \left[\, \mathcal{L}_P(\mathcal{A}(S)) \;\leq\; \mathrm{opt}_P(H) + \epsilon \,\right] \;\geq\; 1 - \delta$$

$\Rightarrow$ Distribution free, finite sample bound
$\Rightarrow$ If such a learner $\mathcal{A}$ exists, we call the class H (PAC-)learnable.

PAC-learnability

Question: Which hypothesis classes are (PAC-)learnable?

A (binary) class H is learnable if and only if it has finite VC-dimension.

In particular, every finite hypothesis class is learnable.

For binary classification, learnability is equivalent to uniform convergence, and Empirical Risk Minimization (ERM) is a successful PAC learner.

Empirical Risk Minimization (ERM)

Let H be some hypothesis class. A learner $\mathcal{A}$ is an Empirical Risk Minimizer (ERM), if for any data-sample S, we have

$$\hat{h} \ = \ \mathcal{A}(S) \in \mathrm{argmin}_{h \in H} \mathcal{L}_S(h)$$

That is, the algorithm outputs a predictor $\hat{h}$ from the class H, that makes the fewest mistakes on the dataset.

Finite classes are (PAC-)learnable

Theorem:

- Let $H = \{h_1, h_2, \ldots, h_N\}$ be a finite hypothesis class.
- Let P be any data-generating distribution over $\mathcal{X} \times \{0, 1\}$.
- Let $\delta > 0$.

With probability at least $1 - \delta$ (over the draw of an iid $S \sim P^n$), we have

$$\mathcal{L}_P(\hat{h}) \; \leq \; \mathcal{L}_P(h^*) + \sqrt{\frac{2(\log(2N) + \log(1/\delta))}{n}}$$

where $h^* \in \mathrm{argmin}_{h \in H} \mathcal{L}_P(h)$.

Finite classes are learnable

Hoeffding's Inequality
Let $Z_1, Z_2, \ldots, Z_n$ be iid random variables, taking values in $[0, 1]$.
Then, for all $\epsilon > 0$, we have

$$\Pr\left[\; \left| \frac{1}{n} \sum_{i=1}^{n} Z_i \; - \; \mathbb{E}[Z_i] \right| > \epsilon \; \right] \leq 2\mathrm{e}^{-2n\epsilon^2}$$

Finite classes are learnable - Proof

We want to show:

$$\mathcal{L}_P(\hat{h}) \;\leq\; \mathcal{L}_P(h^*) + \sqrt{\frac{2(\log(2N) + \log(1/\delta))}{n}}$$

We have:

$$\begin{aligned}
\mathcal{L}_P(\hat{h}) - \mathcal{L}_P(h^*) &\leq \mathcal{L}_P(\hat{h}) - \mathcal{L}_S(\hat{h}) + \mathcal{L}_S(h^*) - \mathcal{L}_P(h^*) \\
&\leq |\mathcal{L}_P(\hat{h}) - \mathcal{L}_S(\hat{h})| + |\mathcal{L}_S(h^*) - \mathcal{L}_P(h^*)| \\
&\leq 2 \cdot \sup_{h \in H} |\mathcal{L}_P(h) - \mathcal{L}_S(h)|
\end{aligned}$$

Setting $Z_i = \mathbf{1}[h(x_i) \neq y_i]$ in Hoeffding's in equality, we get for some fixed $h \in H$:

$$\begin{aligned}
\Pr\left[\; \left|\frac{1}{n}\sum_{i=1}^{n} Z_i \;-\; \mathbb{E}[Z_i]\right| > \epsilon \;\right] &= \Pr\left[\; |\mathcal{L}_S(h) \;-\; \mathcal{L}_P(h)| > \epsilon \;\right] \\
&\leq 2\mathrm{e}^{-2n\epsilon^2}
\end{aligned}$$

Finite classes are learnable - Proof

$$\begin{aligned}
& \Pr\left[\mathcal{L}_P(\hat{h}) \ - \ \mathcal{L}_P(h^*) > \epsilon\right] \\
\leq \ & \Pr\left[\max_{h \in H} |\mathcal{L}_P(h) \ - \ \mathcal{L}_S(h)| > \epsilon/2\right] \\
\leq \ & \Pr\left[\bigvee_{h \in H} |\mathcal{L}_P(h) \ - \ \mathcal{L}_S(h)| > \epsilon/2\right] \\
\leq \ & |H| \cdot 2\mathrm{e}^{-2n\epsilon^2} \ = \ N \cdot 2\mathrm{e}^{-2n\epsilon^2} := \delta
\end{aligned}$$

Solving for ϵ yields:

$$\epsilon = \sqrt{\frac{2(\log(2N) + \log(1/\delta))}{n}}$$

Thus

$$\Pr\left[\mathcal{L}_P(\hat{h}) \ - \ \mathcal{L}_P(h^*) > \sqrt{\frac{2(\log(2N) + \log(1/\delta))}{n}}\right] \ \leq \ \delta$$

Thank You!