Lecture 8

Oct 11, 2023

Goal:

- PAC learnability
- Uniform convergence
- VC dim.

Recall:

Probably Approximately Correct (PAC)

X instance space — set of all instances
(input space / domain)

$c: X \to \{+1, -1\}$ concept — a function to label elements

C concept class — a collection of labeling functions

c^* target concept — $c^* \in C$ and label all instances correctly

D target distribution — distribution over instances

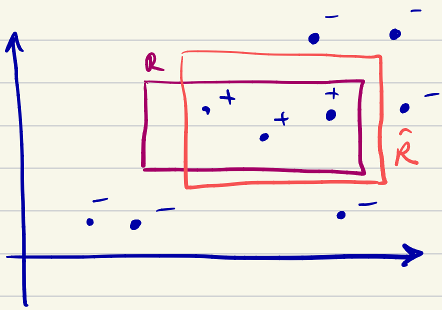
sample / training data set
$$\left\{\begin{array}{l} \langle x_1, c^*(x_1)\rangle \\ \langle x_2, c^*(x_2)\rangle \\ \vdots \\ \langle x_n, c^*(x_n)\rangle \end{array}\right.$$

Learning an axis-aligned rectangle R in $\mathbb{R}^2$

Samples : points $p_1, \ldots, p_n \sim D$ over $\mathbb{R}^2$

label $y_1, \ldots, y_n$

$$y_i = \begin{cases} +1 & \text{if } p_i \in R \\ -1 & \text{otherwise} \end{cases}$$



Goal : output $\hat{R}$ s.t. error of $\hat{R}$ is small (say ϵ) with high probability (say $1-\delta$)

Solution : Draw $m = \underline{\log 1/\delta}$ samples.

output a "consistent" rectangle.

What we did is called :

ERM : Empirical Risk Minimization

comes from samples $\nearrow$ (Empirical) error $\nearrow$ (Risk)

+ ERM could go very wrong if we overfit.

training set

$$\hat{R}(x) = \begin{cases} y_i & x = x_i \in T \\ 0 & x = x_i \in T \end{cases}$$

0 empirical error $\rightsquigarrow$ error 1 on any dist with a continuous domain

ERM has really bad error!

$\ast$ ERM works for a finite class C if we have enough samples.

- Problem setup:

samples $(x_1, y_1), \ldots, (x_m, y_m) \sim D$

$c \in C : \text{err}(c) := \Pr_{(x,y)\sim D} [c(x) \neq y]$

Realizable case

Assume $\exists\, c^* \in C$ s.t. $\text{err}(c^*) = 0$

- Goal

find $\hat{c} \in C$ s.t. with probability $1-\delta$, $\text{err}(\hat{c}) \leq \varepsilon$.

- Proof

Bad hypotheses $C_B := \{ c \in C \mid \text{err}(c) > \varepsilon \}$

training set

$$\hat{err}_T(c) := \frac{|\{(x,y) \in T \mid c(x) \neq y\}|}{|T|}$$

Misleading training samples

$$M := \{ T \mid \exists c \in C_B \text{ s.t. } \hat{err}_T(c) = 0 \}$$

upon observing T, we may pick c that is a bad choice, but it "looked" good from ERM perspective, since $\hat{err}_T(c) = 0$.

Our goal is to show observing a dataset $T \in M$ happens only with probability δ.

This is sufficient to prove $\star$.

fix $c \in C_B$.

what is the probability of

$\hat{err}_T(c) = 0$

$$\Pr_{T \sim D^m}\left[\hat{err}_T(c) = 0\right]$$

$$= \Pr_{T \sim D^m}\left[\forall (x,y) \in T \,.\, c(x) = y\right]$$

iid samples

$$= \left(\Pr_{(x,y) \sim D}\left[c(x) = y\right]\right)^m$$

err(C) > ε

$$< (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

Now, we are ready to bound

$$\Pr_{T \sim D^m} [T \in \mathcal{U}]$$

$$= \Pr_{T \sim D^m} [\exists c \in C_B \text{ s.t. } \hat{err}_T(c) = 0]$$

$$= \sum_{c \in C_B} \Pr_{T \sim D^m} [\hat{err}_T(c) = 0]$$

$$\leq |C_B| \cdot e^{-\varepsilon m} \leq |C| \cdot e^{-\varepsilon m}$$

set $m = \dfrac{\log(|C|/\delta)}{\varepsilon}$

$\Rightarrow$ Pr[outputting a misleading c]

$\leq \delta$ $\square$

The agnostic case :

What if there is no perfect $c \in C$?

$$\forall\ c \in C \quad err(c) > 0$$

Goal

Find $\hat{c} \in C$ s.t.

$$err(\hat{c}) < \underbrace{\min_{c \in C} err(c)}_{= OPT} + \varepsilon$$

the best possible option

Exercise 1.

Suppose we have a finite class C, and $m = O\left(\frac{\log |C|/\delta}{\varepsilon^2}\right)$. then w.p. at least $1-\delta$, for all $c \in C$, we have:

$$\left| \hat{err}_S(c) - err(c) \right| < \varepsilon/2$$

+ Uniform convergence. (UC)

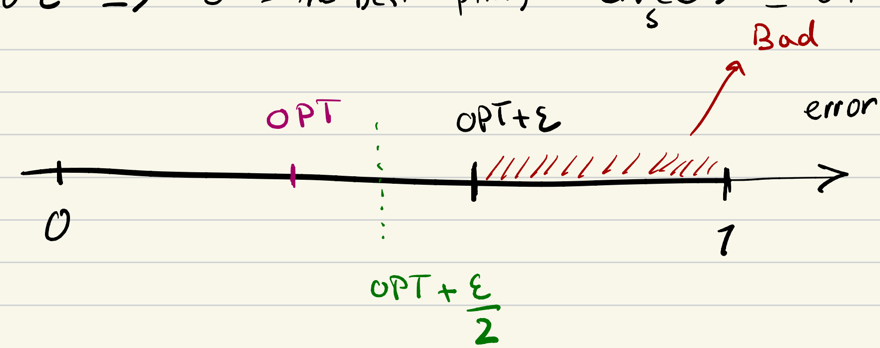
Class C has the uniform convergence property if $\forall\ \varepsilon, \delta \in (0,1)$, dist D $\exists\ m$ (as a function of $\varepsilon, \delta, \mathcal{H}$, but not D since we don't know D). s.t. for a training set of size m:

$$\Pr_{T \sim D^m}\left[\forall c \in C : \left|\hat{err}_T(c) - err(c)\right| \leq \varepsilon\right] \geq 1-\delta$$

Uniform convergence implies agnostic PAC learnability via EMR.

UC $\Rightarrow$ $\forall c \in C_B$ $\quad \hat{err}_S(c) > OPT + \varepsilon/2$

UC $\Rightarrow$ c^* = the best option) $\quad \hat{err}_S(c^*) \leq OPT + \varepsilon$



There are two types of error in the agnostic setting:

$$\mathrm{err}(\hat{c}) < \underbrace{\min_{c \in C} \mathrm{err}(c)}_{\varepsilon_{app} := \text{approximation error}} + \underbrace{\varepsilon}_{\varepsilon_{est} := \text{estimation error}}$$

↓

depends only to the choice of the class C

- Is C rich enough to capture how data is labeled?

C larger more complex	ε_{app}	ε_{est}
↑	↓	↑

No free lunch theorem says if there is no universal learner :

for a complex C even when ε_{app} is 0, $\varepsilon_{est} \gg$ constant with some constant probability

[unless we have $\Omega(|X|)$ samples]

VC dimension

- infinite classes can still be PAC-learnable.

$\Rightarrow$ size is <u>not</u> determinant of learnability.

So, what is then?

VC-dim of C characterizes its learnability!