

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)

$c: X \rightarrow \{+1, -1\}$ 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

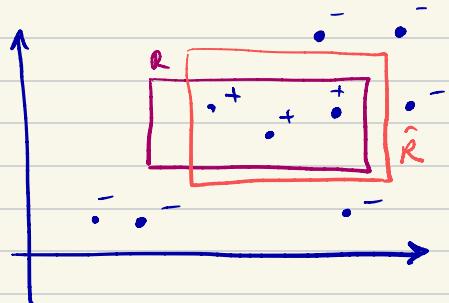
sample / training data set.

$\langle x_1, c^*(x_1) \rangle$
 $\langle x_2, c^*(x_2) \rangle$
⋮
 $\langle x_n, c^*(x_n) \rangle$

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

Samples : points $p_1, \dots, p_n \sim D$ over \mathbb{R}^2
label y_1, \dots, 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 $m = \log \frac{1}{\delta}$ samples.
output a "consistent" rectangle.

What we did is called:

ERM : Empirical Risk Minimization

↑ ↑
comes from samples error

- + ERM could go very wrong if we overfit.
training set

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

0 empirical error } error 1 on any dist
with a continuous domain

ERM has really bad error! ↴

*

ERM works for a finite class C if we have enough sample.

- Problem setup:

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

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

Realizable case

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

- Goal

find $\hat{c} \in C$ s.t. with probability

$$1 - \delta, \quad \text{err}(\hat{c}) \leq \varepsilon.$$

- Proof

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

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

↑
training set

Misleading training samples

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

Upon observing T , we may pick c that is a bad choice, but it "looked" good from ERM perspective, since

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

Our goal is to show observing a dataset $T \in M$ happens only with probability 0.

This is sufficient to prove \star .

fix $c \in C_B$

what is the probability of

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

$$\Pr_{T \sim D^n} [\hat{\text{err}}_T(c) = 0]$$

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

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

$\text{err}(c) > \epsilon \rightarrow < (1 - \epsilon)^m \leq e^{-\epsilon m}$

Now, we are ready to bound

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

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

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

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

$$\text{set } m = \frac{\log(|C|/\delta)}{\epsilon}$$

$$\Rightarrow \Pr [\text{outputting a misleading } c]$$

$$\leq \delta$$

□

The agnostic case:

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

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

Goal

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

$$\text{err}(\hat{c}) < \min_{c \in C} \text{err}(c) + \varepsilon$$


= OPT

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:

$$|\hat{\text{err}}_S(c) - \text{err}(c)| < \frac{\varepsilon}{2}$$

+ Uniform convergence. (UC)

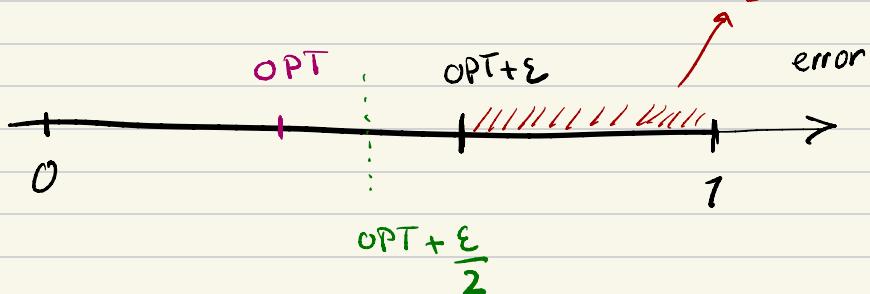
Class C has the uniform convergence property if $\forall \epsilon, \delta \in (0,1)$, $\text{dist } D$
 $\exists m$ (as a function of ϵ, δ, H , but not D since we don't know D). s.t. for
 a training set of size m :

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

Uniform convergence implies agnostic PAC learnability via EMR.

$$UC \Rightarrow \forall c \in C_B \quad \hat{\text{err}}_s(c) > \text{OPT} + \frac{\epsilon}{2}$$

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



There are two types of error
in the agnostic setting:

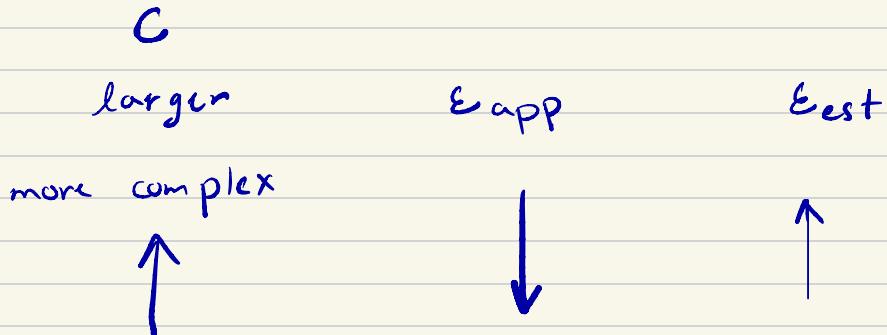
$$\text{err}(\hat{c}) < \min_{c \in C} \text{err}(c) + \underbrace{\epsilon}_{\text{E}_{\text{app}} = \text{approximation error}}$$

ϵ $E_{\text{est}} := \text{estimation error}$



depends only to the choice
of the class c

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



No free lunch theorem says if
there is no universal learner ;
for a complex C even when
 E_{app} is 0 , $E_{est} \gg$ constant
with some constant probability
[unless we have $\Omega(1/X)$ samples]

VC dimension

- infinite classes can still be PAC-learnable.

\Rightarrow size is not determinant of learnability.

So, what is then?

VC-dim of C characterizes its learnability!