Lecture 15

Uniform convergence

overfitting

PAC learnability of finite classes.

No free lunch theorem

+ some proof witness

Last lecture :

Recall:

+ 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:

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

$$\boxed{\text{Uniform convergence implies agnostic PAC learnability via EMR.}}$$

+ 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 $\Bigg\}$ 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 samples.

- Problem setup:

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

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

Realizable case

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

- Goal

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

- Proof

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

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 ⋆.

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 $\rightarrow$

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

$err(C) > \varepsilon \rightarrow$

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

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.

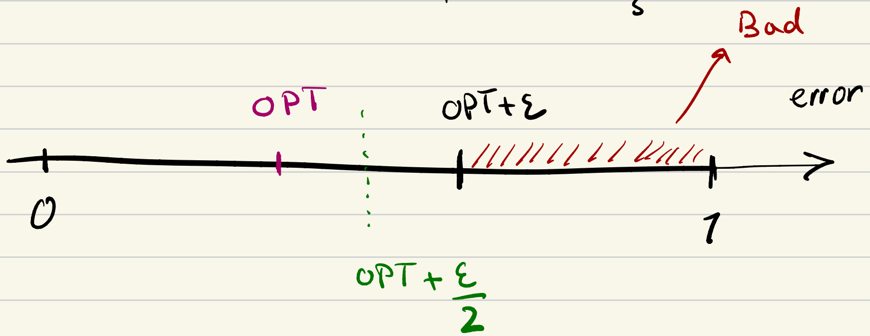
$$err(\hat{c}) < \underbrace{\min_{c \in C} err(c)}_{= OPT} + \varepsilon$$

the best possible option

$$\boxed{\text{Uniform convergence implies agnostic PAC learnability via EMR.}}$$

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

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



Exercise!

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:

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

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]

Suppose we have a set of $2m$ points

There are 2^{2m} possible labelings of these $2m$ points.

Suppose C is the class of 2^{2m} func. that assigns these labelings to these points.

Assume this is the true labeling.

Fix a labeling of the points

Now assume D is the uniform distribution on the 2m points with their label.

T ← Draw m samples from D

(WLOG assume they are unique)

How many function in C label T correctly? 2^m

$P := \{ c \in C \mid \widehat{err}_T(c) = 0 \}$

↳ promising hypothese. $|P| = 2^{m/2}$

How many of them has error < ε?

c is misleading if $\begin{cases} err(c) > \varepsilon \\ \text{and } \hat{err}_T(c) = 0 \end{cases}$

$$\mathcal{M} := \{ c \in C \mid err(c) > \varepsilon \ \& \ \hat{err}_T(c) = 0 \}$$

$$|\mathcal{M}| = \frac{|\mathcal{M}|}{|P|} \cdot |P|$$

$$= 2^m \cdot \Pr_{c \sim_U P} [c \in \mathcal{M}]$$

a random concept in P

c makes $\geq m \cdot \varepsilon$ mistakes in expectation

$$= 2^m \cdot \Pr \left[\frac{\#\text{mistake}}{m} < \varepsilon \right]$$

$$= 2^m \left(1 - \Pr \left[\frac{\#\text{mistakes}}{m} < \frac{1}{2} - \left(\frac{1}{2} - \varepsilon\right) \right] \right)$$

$$\geq 2^m \left(1 - e^{\left(-2m\left(\frac{1}{2}-\varepsilon\right)^2\right)} \right)$$

Hoeffding bound $\qquad \geq 2^{m/2} \cdot 0.99$

$\varepsilon \leq \frac{1}{4} \qquad m \geq 40$

$\Rightarrow$ 0.99 % of the promising concept are bad!