A trichotomy of rates in supervised learning

Amir Yehudayoff (Technion)

Olivier Bousquet (Google)
Steve Hanneke (TTIC)
Shay Moran (Technion & Google)
Ramon van Handel (Princeton)

background

learning theory

PAC learning is standard definition

sometimes fails to provide valuable information
– specific algorithms (nearest neighbor, neural nets, ...)
– specific problems

learning rates

framework

input: sample of size n

$$S = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (\mathcal{X} \times \{0,1\})^n$$

output: an hypothesis

$$S \underset{A}{\mapsto} h \in \{0,1\}^{\mathcal{X}}$$

learning algorithm A

generalization

goal: PAC learning

if $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ is i.i.d. from unknown μ

then $h = A(S)$ is typically close to μ

closeness is measured by

$$err(h) = \Pr_{(x,y)\sim\mu}[h(x) \neq y]$$

context

without "context" learning is "impossible"
what is next element of $1, 2, 3, 4, 5, \ldots$?

few possible definitions

for a class $\mathcal{H}$, the distribution μ is **realizable** if

$$\inf\{err(h) : h \in \mathcal{H}\} = 0$$

where $err(h) = \Pr_{(x,y)\sim\mu}[h(x) \neq y]$

PAC learning

error of algorithm for sample size n

$$ERR_n(A, \mathcal{H}) = \sup \left\{ \mathop{\mathbb{E}}_{S \sim \mu^n} err(A(S)) : \mu \text{ is } \mathcal{H}\text{-realizable} \right\}$$

the class $\mathcal{H}$ is **PAC learnable** if there is A so that

$$\lim_{n \to \infty} ERR_n(A, \mathcal{H}) = 0$$

VC theory

theorem [Vapnik-Chervonenkis, Blumer-Ehrenfeucht-Haussler-Warmuth, ...]

$\mathcal{H}$ is PAC learnable $\Leftrightarrow$ VC dimension of $\mathcal{H}$ is finite

learning curve [Schuurmans]

error "should" decrease as more examples are seen

this improvement is important (predict, estimate, ...)

rates

usually: μ is unknown but fixed

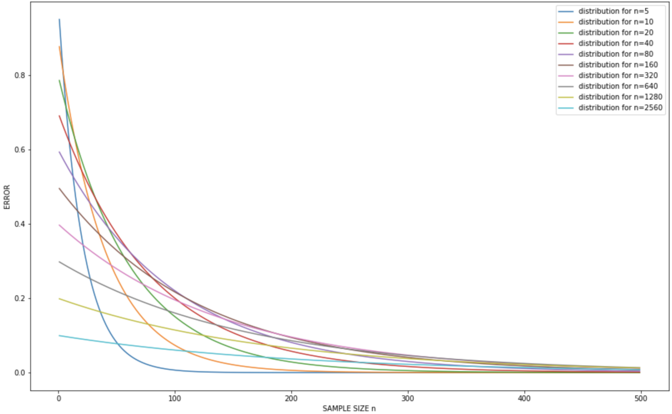
want definition to capture this

the **rate** of algorithm A with respect to μ is

$$\mathrm{rate}(n) = \mathrm{rate}_{A,\mu}(n) = \mathop{\mathbb{E}}_{S} err(A(S))$$

where $err(h) = \Pr_{(x,y)\sim\mu}[h(x) \neq y]$ and $|S| = n$

VC classes



thm: upper envelope $\approx \frac{VC}{n}$ [Vapnik-Chervonenkis, Blumer-Ehrenfeucht-Haussler-Warmuth, ...]

experiments: $rate(n) \lesssim \exp(-n)$ for fixed μ [Cohn-Tesauro]

rate of class

$R : \mathbb{N} \to [0, 1]$ is a rate function

the class $\mathcal{H}$ has **rate** $\leq R$ if

$$\exists A \; \forall \mu \; \exists C \; \forall n \qquad \mathbb{E}\, err(A(S)) < CR\left(\frac{n}{C}\right)$$

the class $\mathcal{H}$ has **rate** $\geq R$ if

$$\exists C \; \forall A \; \exists \mu \text{ for } \infty \text{ many } n \qquad \mathbb{E}\, err(A(S)) > \frac{R(Cn)}{C}$$

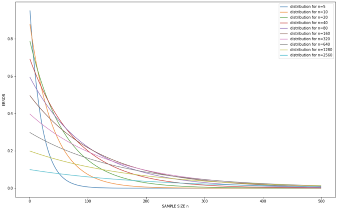
the class $\mathcal{H}$ has **rate** R if both

rates: comments

$$\text{rate} \leq R \text{ if } \exists A\ \forall \mu\ \exists C\ \forall n\ \ \mathbb{E}\, err(A(S)) < CR(n/C)$$

algorithm A does not know distribution μ

the "complexity" of μ is captured by delay factor $C = C(\mu)$



trichotomy theorem*

the rate of $\mathcal{H}$ can be

– exponential (e^{-n})

– linear $(\frac{1}{n})$

– arbitrarily slow (for every $R \to 0$, at least R)

* realizable, $|\mathcal{H}| > 2$, standard measurability assumptions

trichotomy: comments

rate $2^{-\sqrt{n}}$ e.g. is not an option

Schuurmans proved a special case (dichotomy for chains)

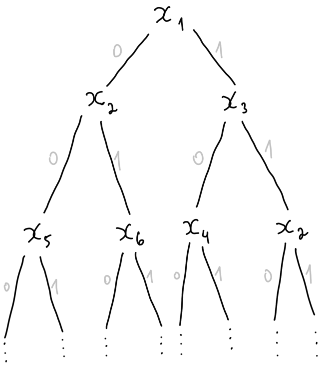
the higher the complexity of $\mathcal{H}$, the slower the rate

the complexity is characterized by "shattering capabilities"

exponential rate

proposition

the rate of $\mathcal{H}$ is exponential iff $\mathcal{H}$ does not shatter an infinite Littlestone tree



exponential rate

lower bound: if $|\mathcal{H}| > 2$ then rate is $\geq e^{-n}$

upper bound: if $\mathcal{H}$ does not shatter an infinite Littlestone tree then rate is $\leq e^{-n}$

$$\exists A \; \forall \mu \; \exists C \; \forall n \;\; \mathbb{E}\, err(A(S)) < Ce^{-n/C}$$

exponential rate

lower bound: if $|\mathcal{H}| > 2$ then rate is $\geq e^{-n}$

upper bound: if $\mathcal{H}$ does not shatter an infinite Littlestone tree then rate is $\leq e^{-n}$

$$\exists A \; \forall \mu \; \exists C \; \forall n \;\; \mathbb{E}\, err(A(S)) < Ce^{-n/C}$$

need: no tree $\Rightarrow$ algorithm

duality (LP, games,...)

no tree $\Rightarrow$ algorithm

simplest example:

no point in intersection of two convex bodies
$\Rightarrow$ a separating hyperplane

duality (LP, games,...)

no tree $\Rightarrow$ algorithm

simplest example:

no point in intersection of two convex bodies
$\Rightarrow$ a separating hyperplane

duality for Gale-Stewart games:
one of players have a winning strategy

duality (LP, games,...)

no tree $\Rightarrow$ algorithm

simplest example:

no point in intersection of two convex bodies
$\Rightarrow$ a separating hyperplane

duality for Gale-Stewart games:
one of players have a winning strategy

problem: how complex is this strategy?

measurability

value of position is an ordinal

measures “how many steps to victory”

n-steps to mate [Evans, Hamkins]

measurability

value of position is an ordinal

measures "how many steps to victory"

n-steps to mate [Evans, Hamkins]

the **Littlestone dimension** of $\mathcal{H}$ is the ordinal

$$LD(\mathcal{H}) = \begin{cases} 0 & |\mathcal{H}| = 1 \\ \infty & \mathcal{H} \text{ has } \infty \text{ tree} \\ \left(\sup_{x \in \mathcal{X}} \min_{y \in \{0,1\}} LD\left(\mathcal{H}\big|_{x \mapsto y}\right)\right) + 1 & \text{otherwise} \end{cases}$$

measurability

value of position is an ordinal

measures "how many steps to victory"

n-steps to mate [Evans, Hamkins]

the **Littlestone dimension** of $\mathcal{H}$ is the ordinal

$$LD(\mathcal{H}) = \begin{cases} 0 & |\mathcal{H}| = 1 \\ \infty & \mathcal{H} \text{ has } \infty \text{ tree} \\ \left(\sup_{x\in\mathcal{X}} \min_{y\in\{0,1\}} LD\left(\mathcal{H}\big|_{x\mapsto y}\right)\right) + 1 & \text{otherwise} \end{cases}$$

theorem (relies on [Kunen-Martin])

if $\mathcal{H}$ is measurable* then $LD(\mathcal{H})$ is countable

summary

learning rates capture distribution specific performance

there are 3 possible learning rates in realizable case

rate is characterizes by shattering capabilities
– shattering $\Rightarrow$ hard distribution via construction
– no shattering $\Rightarrow$ algorithm via duality

complexity of algorithm via ordinals etc.

to do

agnostic case

accurate bounds on rates

applications for shattering framework