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 $n$

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

**output:** an hypothesis

\[ S \overset{A}{\rightarrow} h \in \{0, 1\}^\mathcal{X} \]

learning algorithm $A$
generalization

goal: PAC learning

if \( S = ((x_1, y_1), \ldots, (x_n, y_n)) \) is i.i.d. from unknown \( \mu \)
then \( h = A(S) \) is typically close to \( \mu \)

closeness is measured by

\[
\text{err}(h) = \Pr_{(x,y) \sim \mu} [h(x) \neq y]
\]
without “context” learning is “impossible”
what is next element of 1, 2, 3, 4, 5, …?

few possible definitions

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

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

where $\text{err}(h) = \Pr_{(x,y) \sim \mu}[h(x) \neq y]$
error of algorithm for sample size $n$

$$ERR_n(A, \mathcal{H}) = \sup \{ \mathbb{E} \text{ err}(A(S)) : \mu \text{ is } \mathcal{H}\text{-realizable} \}$$

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, ...]

\[ \mathcal{H} \text{ is PAC learnable } \iff \text{ VC dimension of } \mathcal{H} \text{ is finite} \]
error “should” decrease as more examples are seen

this improvement is important (predict, estimate, ...)
usually: $\mu$ is unknown but fixed
want definition to capture this

the rate of algorithm $A$ with respect to $\mu$ is

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

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: $rate(n) \preceq \exp(-n)$ for fixed $\mu$  [Cohn-Tesauro]
rate of class

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

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

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

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

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

the class $\mathcal{H}$ has rate $R$ if both
rates: comments

rate ≤ R if ∃A ∀µ ∃C ∀n 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(µ)
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
rate $2^{-\sqrt{n}}$ e.g. is not an option

Schuurmans proved a special case \textit{(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 $\leq e^{-n}$

$$\exists A \forall \mu \exists C \forall n \quad \mathbb{E}\{\text{err}(A(S))\} < Ce^{-n/C}$$
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 $\leq e^{-n}$

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

**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 \implies\ algorithm

simplest example:

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

duality for Gale-Stewart games:
one of players have a winning strategy
duality (LP, games,...)

no tree ⇒ algorithm

simplest example:

no point in intersection of two convex bodies
⇒ 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

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

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

**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 characterized 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