Marginal-Nonuniform PAC Learnability

NeurIPS 2025

Steve Hanneke, Purdue University
Shay Moran, Technion and Google Research
Maximilian Thiessen, TU Wien

Joint work with. . .

Steve Hanneke
Purdue University

Shay Moran
Technion and Google Research

PAC learning

- Binary classification problem $|Y| = 2$

PAC learning

- Binary classification problem $|Y| = 2$
- Learn a classifier $h : X \to Y$ mapping data points $x \in X$ to labels in $y \in Y$

PAC learning

- Binary classification problem $|Y| = 2$
- Learn a classifier $h : X \to Y$ mapping data points $x \in X$ to labels in $y \in Y$
- Hypothesis class $\mathcal{H}$ of classifiers $f : X \to Y$

PAC learning

- Binary classification problem $|Y| = 2$
- Learn a classifier $h : X \to Y$ mapping data points $x \in X$ to labels in $y \in Y$
- Hypothesis class $\mathcal{H}$ of classifiers $f : X \to Y$
- Training examples are pairs (x, y) where:
  - x is drawn from a fixed but unknown distribution $\mathcal{D} \in \Delta(X)$ over X
  - $y = f^*(x)$ for some unknown target $f^* \in \mathcal{H}$

PAC learning

- Binary classification problem $|Y| = 2$
- Learn a classifier $h : X \to Y$ mapping data points $x \in X$ to labels in $y \in Y$
- Hypothesis class $\mathcal{H}$ of classifiers $f : X \to Y$
- Training examples are pairs (x, y) where:
    - x is drawn from a fixed but unknown distribution $\mathcal{D} \in \Delta(X)$ over X
    - $y = f^*(x)$ for some unknown target $f^* \in \mathcal{H}$
- Training sample $S_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ i.i.d. from $\mathcal{D}$

PAC learning

- Binary classification problem $|Y| = 2$
- Learn a classifier $h : X \to Y$ mapping data points $x \in X$ to labels in $y \in Y$
- Hypothesis class $\mathcal{H}$ of classifiers $f : X \to Y$
- Training examples are pairs (x, y) where:
  - x is drawn from a fixed but unknown distribution $\mathcal{D} \in \Delta(X)$ over X
  - $y = f^*(x)$ for some unknown target $f^* \in \mathcal{H}$
- Training sample $S_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ i.i.d. from $\mathcal{D}$
- Learning algorithm $A : (S_n) \mapsto \hat{h}_n$
- Error of classifier $h : X \to Y$

$$\mathrm{er}_{\mathcal{D},f^*}(h) = \mathbb{P}_{x\sim\mathcal{D}}\big(h(x) \neq f^*(x)\big)$$

PAC learning

A PAC learner for $\mathcal{H}$ is a learning algorithm that for any $f^* \in \mathcal{H}$, $\varepsilon > 0$, and data distribution $\mathcal{D}$, uses a sample size $R(\frac{1}{\varepsilon})$ and outputs a $h : X \to Y$ such that $\mathrm{er}_{\mathcal{D},f^*}(h) \leq \varepsilon$ with high probability

When is PAC learning possible?

Classical answer [Vapnik & Chervonenkis, Blumer et al., Haussler et al., . . .]

- $\mathcal{H}$ is PAC learnable $\iff$ $\mathcal{H}$ has finite VC dimension.
- VC $=$ largest shatterable set: points $\{{\color{gray}\bullet}, {\color{gray}\bullet}\}$ with all possible labels

$$\{{\color{red}\bullet}, {\color{red}\bullet}\} \qquad \{{\color{blue}\bullet}, {\color{red}\bullet}\}$$
$$\{{\color{red}\bullet}, {\color{blue}\bullet}\} \qquad \{{\color{blue}\bullet}, {\color{blue}\bullet}\}$$

When is PAC learning possible?

Classical answer [Vapnik & Chervonenkis, Blumer et al., Haussler et al., . . .]

- $\mathcal{H}$ is PAC learnable $\iff$ $\mathcal{H}$ has finite VC dimension.
- VC $=$ largest shatterable set: points {●, ●} with all possible labels

{●, ●} {●, ●}

{●, ●} {●, ●}

Error rate

$$\boxed{\mathbb{E}_{S_n}[\mathrm{er}_{\mathcal{D},f^*}(\hat{h}_n)] = \Theta\left(\frac{\mathrm{VC}}{n}\right)}$$

PAC learning is uniform

PAC learning is uniform over all $f^* \in \mathcal{H}$ and distributions $\mathcal{D}$:

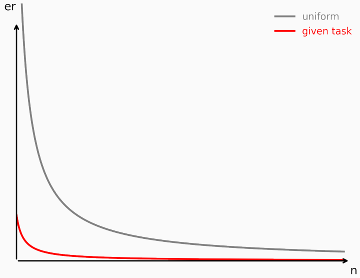
$$\exists \hat{h}_n \text{ s.t. } \exists C, c > 0 \text{ s.t. } \forall \mathcal{D} \in \Delta(X)\ \forall f^* \in \mathcal{H}\text{: } \mathbb{E}[\mathrm{er}_{\mathcal{D},f^*}(\hat{h}_n)] \leq CR(cn) \text{ for all } n.$$

PAC learning is uniform

PAC learning is uniform over all $f^* \in \mathcal{H}$ and distributions $\mathcal{D}$:

$$\exists \hat{h}_n \text{ s.t. } \exists C, c > 0 \text{ s.t. } \forall \mathcal{D} \in \Delta(X)\ \forall f^* \in \mathcal{H}\text{: } \mathbb{E}[\mathrm{er}_{\mathcal{D},f^*}(\hat{h}_n)] \leq CR(cn) \text{ for all } n.$$

In real-world applications we often have *simple* distributions and target concepts

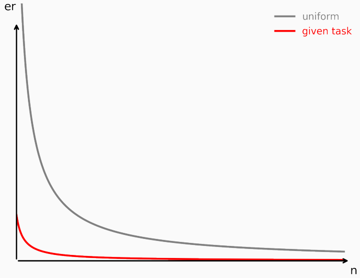


PAC learning is uniform

PAC learning is uniform over all $f^* \in \mathcal{H}$ and distributions $\mathcal{D}$:

$$\exists \hat{h}_n \text{ s.t. } \exists C, c > 0 \text{ s.t. } \forall \mathcal{D} \in \Delta(X)\ \forall f^* \in \mathcal{H}\text{: } \mathbb{E}[\mathrm{er}_{\mathcal{D},f^*}(\hat{h}_n)] \leq CR(cn) \text{ for all } n.$$

In real-world applications we often have *simple* distributions and target concepts



(Uniform) PAC learning is too worst-case!

Nonuniform PAC learning [Benedek & Itai, Ben-David et al., Lugosi & Zeger, ...]

Main idea:

- some target concepts are much easier to learn than others

Nonuniform PAC learning [Benedek & Itai, Ben-David et al., Lugosi & Zeger, . . .]

Main idea:

- some target concepts are much easier to learn than others
- e.g., $\mathcal{H}$ = all polynomials: degree 2 easier to learn than degree 10

Nonuniform PAC learning [Benedek & Itai, Ben-David et al., Lugosi & Zeger, . . .]

Main idea:

- some target concepts are much easier to learn than others
- e.g., $\mathcal{H}$ = all polynomials: degree 2 easier to learn than degree 10

$\rightarrow$ allow error rate to depend on the target concept $f^* \in \mathcal{H}$

Nonuniform PAC learning [Benedek & Itai, Ben-David et al., Lugosi & Zeger, ...]

Main idea:

- some target concepts are much easier to learn than others
- e.g., $\mathcal{H}$ = all polynomials: degree 2 easier to learn than degree 10

$\rightarrow$ allow error rate to depend on the target concept $f^* \in \mathcal{H}$

SRM [Vapnik], Occam's razor [Blumer et al.], Minimum description length [Rissanen], ...

~~Concept~~ Marginal-nonuniform learning [Ben-David et al.]

Main idea:

- some marginal distributions are much easier to learn than others

~~Concept~~ Marginal-nonuniform learning [Ben-David et al.]

Main idea:

- some marginal distributions are much easier to learn than others
- e.g., $\mathcal{H}$ = linear classifiers in $\mathbb{R}^d$: distribution supported on p-dim. subspace with $p \ll d$ easier to learn than $p = d$

~~Concept~~ Marginal-nonuniform learning [Ben-David et al.]

Main idea:

- some marginal distributions are much easier to learn than others
- e.g., $\mathcal{H} =$ linear classifiers in $\mathbb{R}^d$: distribution supported on p-dim. subspace with $p \ll d$ easier to learn than $p = d$

  $\rightarrow$ allow error rate to depend on the marginal distribution $\mathcal{D}$ over X

Characterization of rates

Theorem: Every class $\mathcal{H}$ satisfies (*m.nu. = marginal-nonuniform*)

1. $\mathcal{H}$ is m.nu. learnable with rate e^{-n} iff $|\mathcal{H}| < \infty$

Characterization of rates

Theorem: Every class $\mathcal{H}$ satisfies (*m.nu.* $=$ *marginal-nonuniform*)

1. $\mathcal{H}$ is m.nu. learnable with rate e^{-n} iff $|\mathcal{H}| < \infty$
2. $\mathcal{H}$ is m.nu. learnable with rate $1/n$ iff $|\mathcal{H}| = \infty$ and $\mathrm{vc}(\mathcal{H}) < \infty$

Characterization of rates

Theorem: Every class $\mathcal{H}$ satisfies (*m.nu.* = *marginal-nonuniform*)

1. $\mathcal{H}$ is m.nu. learnable with rate e^{-n} iff $|\mathcal{H}| < \infty$
2. $\mathcal{H}$ is m.nu. learnable with rate $1/n$ iff $|\mathcal{H}| = \infty$ and $\mathrm{vc}(\mathcal{H}) < \infty$
3. $\mathcal{H}$ requires arbitrarily slow rates to be m.nu. learnable iff $\mathrm{vc}(\mathcal{H}) = \infty$

Characterization of rates

Theorem: Every class $\mathcal{H}$ satisfies (*m.nu.* = *marginal-nonuniform*)

1. $\mathcal{H}$ is m.nu. learnable with rate e^{-n} iff $|\mathcal{H}| < \infty$
2. $\mathcal{H}$ is m.nu. learnable with rate $1/n$ iff $|\mathcal{H}| = \infty$ and $\mathrm{vc}(\mathcal{H}) < \infty$
3. $\mathcal{H}$ requires arbitrarily slow rates to be m.nu. learnable iff $\mathrm{vc}(\mathcal{H}) = \infty$

So marginal-nonuniform does not help when $|\mathcal{H}| = \infty$?

Main result: fine-grained rate

Combinatorial parameter *VC-eluder dimension* $d = \mathrm{VCE}(\mathcal{H})$ [Hanneke & Xu]

- for all classes $d \leq \mathrm{VC}(\mathcal{H})$ and often $d \ll \mathrm{VC}(\mathcal{H})$

Main result: fine-grained rate

Combinatorial parameter *VC-eluder dimension* $d = \mathrm{VCE}(\mathcal{H})$ [Hanneke & Xu]

- for all classes $d \leq \mathrm{VC}(\mathcal{H})$ and often $d \ll \mathrm{VC}(\mathcal{H})$

Main theorem: Each class $\mathcal{H}$ with $d < \infty$ is m.nu. learnable with rate d/n

Same rate as uniform learning but typically much better constants!

Overview

$$|\mathcal{H}| < \infty \iff$$
e^{-n}-marginal-nonuniform

Overview

$\mathrm{vc}(\mathcal{H}) < \infty \iff$

$|\mathcal{H}| < \infty \iff$

e^{-n}-marginal-nonuniform

${}^1\!/_n$-marginal-nonuniform $=$ ${}^1\!/_n$-uniform

Overview

$\mathrm{vc}(\mathcal{H}) < \infty \iff$

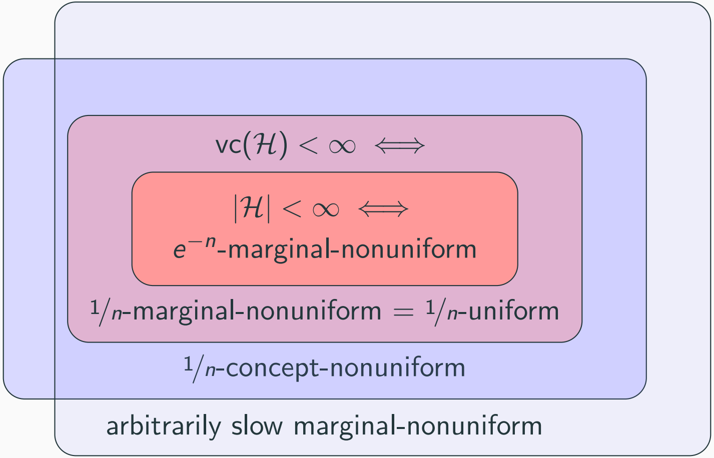
$|\mathcal{H}| < \infty \iff$

e^{-n}-marginal-nonuniform

$^1/_n$-marginal-nonuniform $=$ $^1/_n$-uniform

$^1/_n$-concept-nonuniform

Overview


$\mathrm{vc}(\mathcal{H}) < \infty \iff$
$|\mathcal{H}| < \infty \iff$
e^{-n}-marginal-nonuniform
$1/n$-marginal-nonuniform $=$ $1/n$-uniform
$1/n$-concept-nonuniform
arbitrarily slow marginal-nonuniform

More results in the paper:

- Tight $1/n$ rate for (concept-)nonuniform learning
- Relationship with *universal learning* [Bousquet et al.]

Open: when is marginal-nonuniform learning possible?

Thanks!

See you in San Diego and Copenhagen