On the Hardness of Robust Classification

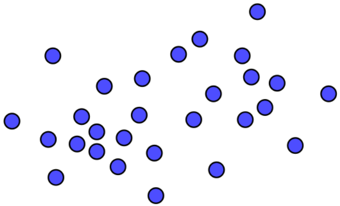
P. Gourdeau, V. Kanade, M. Kwiatkowska and J. Worrell

University of Oxford

NeurIPS 2019

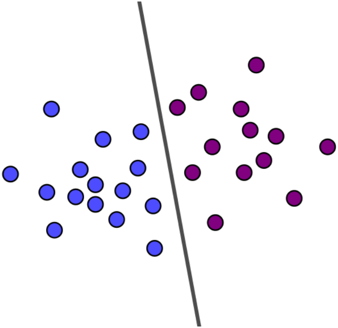
Problem Setting

Problem Setting



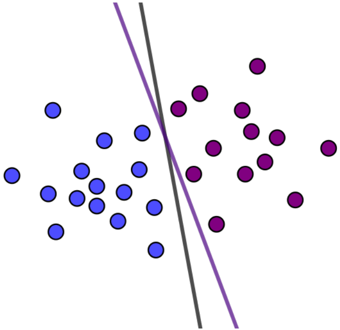
Our setting: Data i.i.d. from unknown distribution,

Problem Setting



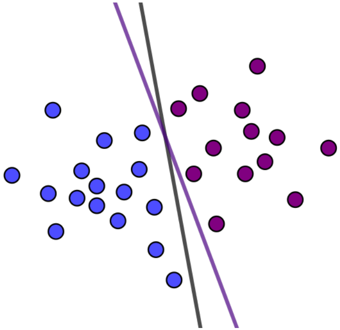
Our setting: Data i.i.d. from unknown distribution, *realizable setting*.

Problem Setting



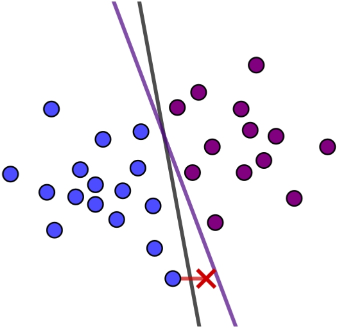
Our setting: Data i.i.d. from unknown distribution, *realizable setting*.
Goal: learn a function that will be robust (with high probability) against an adversary who can perturb the test data.

Problem Setting



Our setting: Data i.i.d. from unknown distribution, *realizable setting*.
Goal: learn a function that will be robust (with high probability) against an adversary who can perturb the test data.

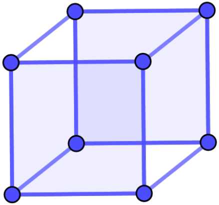
Problem Setting



Our setting: Data i.i.d. from unknown distribution, *realizable setting*.
Goal: learn a function that will be robust (with high probability) against an adversary who can perturb the test data.

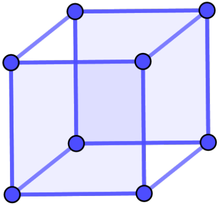
Sample Complexity

Input space: boolean cube $\mathcal{X} = \{0, 1\}^n$.



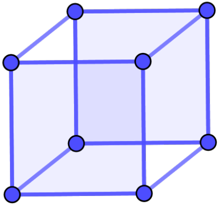
Sample Complexity

Input space: boolean cube $\mathcal{X} = \{0,1\}^n$.
Requirement: *polynomial* sample complexity (*efficient robust learning*).



Sample Complexity

Input space: boolean cube $\mathcal{X} = \{0,1\}^n$.
Requirement: *polynomial* sample complexity (*efficient robust learning*).

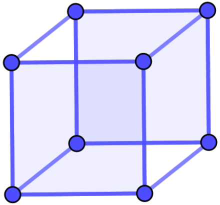


Theorem

$\mathcal{C}$ is efficiently distribution-free robustly learnable iff it is trivial.

Sample Complexity

Input space: boolean cube $\mathcal{X} = \{0,1\}^n$.
Requirement: *polynomial* sample complexity (*efficient robust learning*).



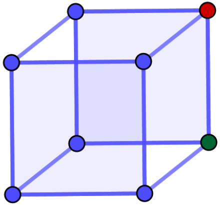
Theorem

$\mathcal{C}$ is efficiently distribution-free robustly learnable iff it is trivial.

$\exists c_1 \neq \neg c_2$

Sample Complexity

Input space: boolean cube $\mathcal{X} = \{0,1\}^n$.
Requirement: *polynomial* sample complexity (*efficient robust learning*).



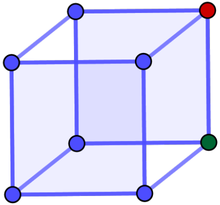
Theorem

$\mathcal{C}$ is efficiently distribution-free robustly learnable iff it is trivial.

$\exists c_1 \neq \neg c_2 \;\; \exists x, x$ s.t. $d_H(x,x) = 1$,

Sample Complexity

Input space: boolean cube $\mathcal{X} = \{0, 1\}^n$.
Requirement: *polynomial* sample complexity (*efficient robust learning*).



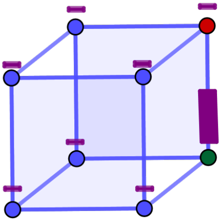
Theorem

$\mathcal{C}$ is efficiently distribution-free robustly learnable iff it is trivial.

$\exists c_1 \neq \neg c_2 \;\; \exists x, x$ s.t. $d_H(x,x) = 1, \;\; c_1(x) = c_2(x)$ and $c_1(x) \neq c_2(x)$

Sample Complexity

Input space: boolean cube $\mathcal{X} = \{0,1\}^n$.
Requirement: *polynomial* sample complexity (*efficient robust learning*).



Theorem

$\mathcal{C}$ is efficiently distribution-free robustly learnable iff it is trivial.

$\exists c_1 \neq \neg c_2 \;\; \exists x, x$ s.t. $d_H(x,x) = 1, \;\; c_1(x) = c_2(x)$ and $c_1(x) \neq c_2(x)$

"Nice" Distributions

Idea: We need distributional assumptions to have efficient robust learning.

"Nice" Distributions

Idea: We need distributional assumptions to have efficient robust learning.

Log-Lipschitz distributions: D is α-log-Lipschitz if the logarithm of the density function is $\log(\alpha)$-Lipschitz w.r.t. the Hamming distance.

“Nice” Distributions

Idea: We need distributional assumptions to have efficient robust learning.

Log-Lipschitz distributions: D is α-log-Lipschitz if the logarithm of the density function is $\log(\alpha)$-Lipschitz w.r.t. the Hamming distance.

$$\begin{array}{l} x_1 = (0, \dots, 1, 1, 1, \dots, 0) \\ x_2 = (0, \dots, 1, 0, 1, \dots, 0) \end{array} \implies \frac{D(x_1)}{D(x_2)} \leq \alpha \ .$$

"Nice" Distributions

Idea: We need distributional assumptions to have efficient robust learning.

Log-Lipschitz distributions: D is α-log-Lipschitz if the logarithm of the density function is $\log(\alpha)$-Lipschitz w.r.t. the Hamming distance.

$$\begin{matrix} x_1 = (0,\dots,1,1,1,\dots,0) \\ x_2 = (0,\dots,1,0,1,\dots,0) \end{matrix} \implies \frac{D(x_1)}{D(x_2)} \leq \alpha \ .$$

Intuition: input points that are close to each other cannot have vastly different probability masses.

"Nice" Distributions

Idea: We need distributional assumptions to have efficient robust learning.

Log-Lipschitz distributions: D is α-log-Lipschitz if the logarithm of the density function is $\log(\alpha)$-Lipschitz w.r.t. the Hamming distance.

$$\begin{matrix} x_1 = (0, \dots, 1, 1, 1, \dots, 0) \\ x_2 = (0, \dots, 1, 0, 1, \dots, 0) \end{matrix} \implies \frac{D(x_1)}{D(x_2)} \leq \alpha \ .$$

Intuition: input points that are close to each other cannot have vastly different probability masses.

Examples: uniform distribution, product distribution where the mean of each variable is bounded, etc.

A Robustness Threshold

Question: Given a concept class $\mathcal{C}$, what is the threshold ρ such that we can *efficiently* robustly learn against adversary with budget ρ?

A Robustness Threshold

Question: Given a concept class $\mathcal{C}$, what is the threshold ρ such that we can *efficiently* robustly learn against adversary with budget ρ?

Our paper: We study the class MON-CONJ of monotone conjunctions, e.g. $x_1 \wedge x_2 \wedge x_5$

A Robustness Threshold

Question: Given a concept class $\mathcal{C}$, what is the threshold ρ such that we can *efficiently* robustly learn against adversary with budget ρ?

Our paper: We study the class MON-CONJ of monotone conjunctions, e.g. $x_1 \wedge x_2 \wedge x_5$

Theorem

The threshold to efficiently robustly learn MON-CONJ under log-Lipschitz distributions is $\rho(n) = O(\log n)$.

A Robustness Threshold

Question: Given a concept class $\mathcal{C}$, what is the threshold ρ such that we can *efficiently* robustly learn against adversary with budget ρ?

Our paper: We study the class MON-CONJ of monotone conjunctions, e.g. $x_1 \wedge x_2 \wedge x_5$

Theorem

The threshold to efficiently robustly learn MON-CONJ under log-Lipschitz distributions is $\rho(n) = O(\log n)$.

$\boldsymbol{\rho(n) = O(\log n)}$: PAC learning algorithm with larger (but still polynomial) sample complexity is a robust learner.

A Robustness Threshold

Question: Given a concept class $\mathcal{C}$, what is the threshold ρ such that we can *efficiently* robustly learn against adversary with budget ρ?

Our paper: We study the class MON-CONJ of monotone conjunctions, e.g. $x_1 \wedge x_2 \wedge x_5$

Theorem

The threshold to efficiently robustly learn MON-CONJ under log-Lipschitz distributions is $\rho(n) = O(\log n)$.

$\boldsymbol{\rho(n) = O(\log n)}$: PAC learning algorithm with larger (but still polynomial) sample complexity is a robust learner.

$\boldsymbol{\rho(n) = \omega(\log(n))}$: no sample-efficient learning algorithm exists to robustly learn MON-CONJ under the uniform distribution.

Take Away

- The definitions and models come from previous work in adversarial machine learning theory.

Take Away

- The definitions and models come from previous work in adversarial machine learning theory.
- At first glance, they seem in many ways *natural* and *reasonable*.

Take Away

- The definitions and models come from previous work in adversarial machine learning theory.
- At first glance, they seem in many ways *natural* and *reasonable*.
  - Their *inadequacies* surface when viewed under the lens of computational learning theory.

Take Away

- The definitions and models come from previous work in adversarial machine learning theory.
- At first glance, they seem in many ways *natural* and *reasonable*.
  - Their *inadequacies* surface when viewed under the lens of computational learning theory.
- It may be possible to only solve "easy" robust learning problems with strong *distributional assumptions*.

Take Away

- The definitions and models come from previous work in adversarial machine learning theory.
- At first glance, they seem in many ways *natural* and *reasonable*.
  - Their *inadequacies* surface when viewed under the lens of computational learning theory.
- It may be possible to only solve "easy" robust learning problems with strong *distributional assumptions*.
- Other learning models, e.g. when one has access to *membership queries*.

Thank you!

Poster Information...