Sample Complexity Bounds for Robustly Learning Decision Lists against Evasion Attacks

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF OXFORD

Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska and James Worrell

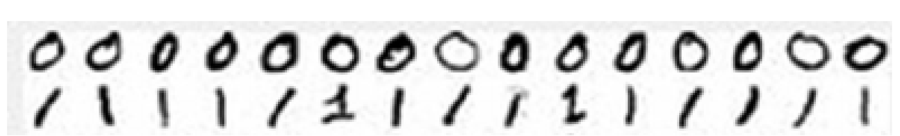
Question

How many data are needed for robust learning against evasion attacks under smooth distributions?

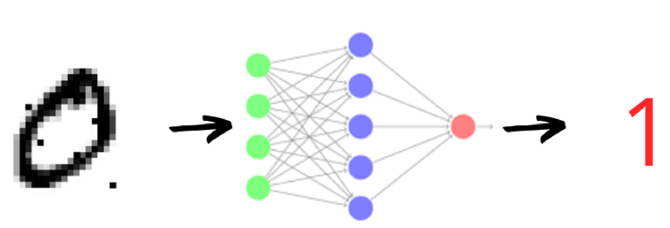
Problem Setting

- Binary classification
- Feature vectors (input space: $\mathcal{X} = \{0,1\}^n$)
- An adversary can modify input bits after training (*evasion attacks*)

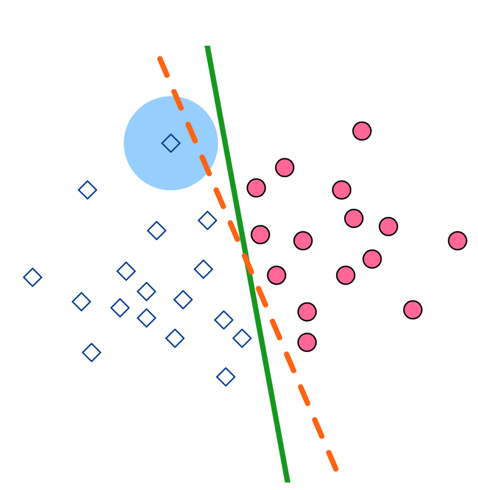
For example, we wish to be able to differentiate between 0's and 1's:



The image of a 0 should not be classified as a 1 if it is slightly perturbed by an adversary:



Efficient Robust Learning:
We want to prove or disprove the existence of an algorithm with *polynomial sample complexity* (in the learning parameters and input dimension n) that will output a hypothesis such that the probability of drawing a new point that can be perturbed by an adversary and resulting in a misclassification to be small:



Exact-in-the-ball Robust Risk:

$$\mathsf{R}^E_\rho(h,c) = \underset{x\sim D}{\mathbb{P}}\left(\exists z \in B_\rho(x) : h(z) \neq c(z)\right)$$

Take Away

- Adversary's budget is a fundamental quantity determining the sample complexity of robust learning
- We can efficiently use standard PAC algorithms as black boxes for some robust learning problems
- Our paper: decision lists under smooth distributions

Open Problem

Is a sample-efficient PAC-learning algorithm for concept class $\mathcal{C}$ also a sample-efficient $\log(n)$-robust learning algorithm for $\mathcal{C}$ under the uniform distribution?

- From previous work [1]
- Focus on the boolean hypercube $\{0,1\}^n$
- **Our sample complexity upper bound for decision lists** adds to the body of positive evidence for this problem.

Decision Lists

A decision list $f \in k\text{-}\mathsf{DL}$ is a list of pairs

$$(K_1, v_1), \dots, (K_r, v_r),$$

K_j: conjunction of size at most k with literals drawn from $\{x_1, \bar{x}_1, \dots, x_n, \bar{x_n}\}$, $v_j \in \{0,1\}$, $K_r = \texttt{true}$. The output $f(x)$ on $x \in \{0,1\}^n$ is v_j, where j is the least index s.t. K_j evaluates to $\texttt{true}$. Example:

$$\begin{array}{ccccccc} x_3 & \longrightarrow & x_1 \wedge \bar{x}_5 & \longrightarrow & x_2 \wedge \bar{x}_3 & \longrightarrow & \mathsf{true} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ 1 & & 0 & & 1 & & 0 \end{array}$$

Smooth Distributions

α-Log-Lipschitz Distributions:

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

For e.g.: uniform distribution, product distribution where the mean of each variable is bounded, etc.

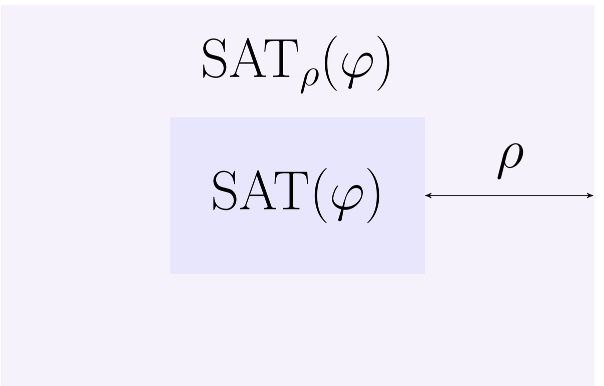
Intuition: input points that are close to each other cannot have vastly different probability masses.

Decision List Sample Complexity

Theorem: *Decision lists are efficiently $\log(n)$-robustly learnable under smooth distributions.*

- A *polynomial* number of examples is enough to return a hypothesis with small robust risk (with high probability).

A Unifying Result.



- $\varphi \in k$-CNF: $\varphi(x) = \wedge_{i\in I} \vee_{1\leq j\leq k} l_{ij}$
- $\rho(n) = \log n$
- $\mathrm{SAT}(\varphi) = \{x \in \mathcal{X} \mid \varphi(x) = 1\}$
- $|\mathrm{SAT}(\varphi)| \leq \mathrm{poly}(\varepsilon, 1/n) \implies |\mathrm{SAT}_{\log n}(\varphi)| \leq \varepsilon$

Proof Idea

1. Event of an exit at depths d_1, d_2 for two k-DL = a **k-CNF formula** $\varphi = \wedge_i \vee_{j=1}^k z_{ij}$
   - Error between the hypothesis and ground truth
2. **Induction on k:** the $\log(n)$-expansion of satisfying assignments of φ (i.e., the robust risk) isn't too large
   - Unifying result above
3. Controlling the standard risk $\implies$ controlling the **robust** risk
   - The standard learning algorithm for k-DL is a robust learner!

Monotone Conjunctions

"AND" of boolean variables:

thesis $\wedge$ sleep deprivation $\wedge$ caffeine

Concept classes that subsume monotone conjunctions:

- Decision lists
- Decision trees
- Linear classifiers

Sample complexity lower bound for monotone conjunctions holds for these classes as well.

Lower Bound

Theorem: *Any $\rho(n)$-robust learning algorithm for monotone conjunctions has a sample complexity lower bound of $\Omega(2^{\rho(n)})$ under the uniform distribution.*

- Any concept class that subsumes monotone conjunctions require a sample size for robust learning where there is an *exponential* dependence on the adversary's budget.

Proof Idea

- Two disjoint monotone conjunctions c_1, c_2 of length 2ρ have robust risk $R_\rho(c_1, c_2)$ bounded below by a constant
- A random sample of size $m = \Omega(2^\rho)$ won't be able to distinguish c_1 from c_2 w.p. $> 1/2$

References

[1] P. Gourdeau, V. Kanade, M. Kwiatkowska, and J. Worrell. On the hardness of robust classification. *Journal of Machine Learning Research*, 22, 2021.