Upper Bound on Malicious Noise Rate for PAC Learning

Dimitris Diochnos

Fall 2021

Abstract

Kearns and Li proved in [1, Theorem 1] that the rate tolerated by PAC learning algorithms in the malicious noise model is upper bounded by ε. They did so by using two oracles; one for positive and one for negative examples. This model of PAC learning (with two oracles) was popular at the time. Below we see a proof using just one oracle that returns both positive and negative examples.

Let $\varepsilon \leqslant 1/2$. Let the concept class $\mathcal{C}$ have at least two concepts c_1 and c_2 such that on two points $u, v \in X$ it holds $c_1(u) = c_2(u)$ and $c_1(v) \neq c_2(v)$; i.e., the concepts c_1 and c_2 agree on one of the instances and disagree on the other one. Now consider the following distribution $\mathcal{D}$ shown below.

	c_1	c_2	$\mathcal{D}$
u	1	1	$1-\varepsilon$
v	1	0	ε

- Any hypothesis that disagrees even in one of these two points, implies that it has error at least ε and thus it is not accepted as a solution satisfying the PAC criterion (with strict inequalities).
- Requiring two such points is meaningful. For example, let $u = (1, \ldots, 1)$ and $v = (1, \ldots, 1, 0)$ and the two concepts be $c_1 = x_1 \wedge x_2 \wedge \ldots \wedge x_{n-1}$ and $c_2 = c_1 \wedge x_n = x_1 \wedge x_2 \wedge \ldots \wedge x_{n-1} \wedge x_n$.

During the course of learning, the adversary presents a point drawn from $\mathcal{D}$ with probability $1-\eta$. Furthermore, with probability η it returns v with the opposite label. The induced distribution $\mathcal{D}'$ is shown below.

	c_1	c_2	$\mathcal{D}'$			
u	1	1	$(1-\eta)\cdot(1-\varepsilon)$	$+$	0	$\leftarrow$ always return label 1
v	1	0	$(1-\eta)\cdot\varepsilon$	$+$	η	$\leftarrow$ return both labels
			honest label		opposite label	

If we require now $(1-\eta)\varepsilon = \eta \Leftrightarrow \eta = \frac{\varepsilon}{1+\varepsilon}$, then it follows that the instance v is returned to the learner with both labels at the same rate $\varepsilon/(1+\varepsilon)$. However, the same distribution $\mathcal{D}'$ can be obtained when we use c_2 as the target concept and again return the instance v with rate η with opposite label. Therefore, any algorithm that produces an ε-good hypothesis with probability at least $1-\delta$ when the target is c_1, then with the same probability must produce an ε-bad hypothesis when the target is c_2. Hence, the malicious noise rate that can potentially be tolerated is strictly less than $\varepsilon/(1+\varepsilon)$.

Remark 1. If the adversary has more power of tampering than $\varepsilon/(1+\varepsilon)$, then they can always return an honest example for the excess part of the probability above this threshold and now repeat the above argument.

References

[1] Michael J. Kearns and Ming Li. Learning in the Presence of Malicious Errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.