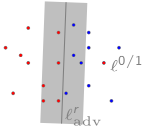
On the (Un-)Avoidability of Adversarial Examples

Sadia Chowdhury, Ruth Urner. ICML2021 Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI

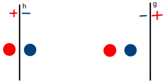
Phenomenon:



- Adversarial examples pose safety concerns.
- The usual formulation of the adversarial loss

$$\ell^r_{\mathrm{adv}}(h, x, y) = 1\left[\exists x' \in \mathcal{B}_r(x) \;:\; h(x') \neq y\right]$$

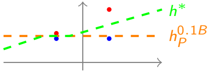
leads to inconsistencies with accuracy.



$\mathrm{err}_h = 0,\ \mathrm{mar}^r_h = 1$
$\Rightarrow \mathcal{L}^r_P(h) = 1$

$\mathrm{err}_g = 0.5,\ \mathrm{mar}^r_g = 0$
$\Rightarrow \mathcal{L}^r_P(g) = 0.5$

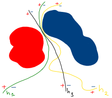
Question: Is it capturing what was intended?



$\Rightarrow$ We really want locally maximal robustness!

Our contributions:

- We define the notion of a margin canonical Bayes classifier h^B_P.

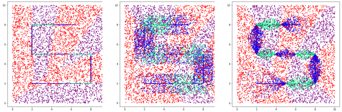


- We propose to re-define the robust loss as a locally adaptive requirement with respect to to a margin canonical Bayes classifier:

$$\ell^{ar}(h, x, y) = 1\left[h(x) \neq y \ \vee\ \mathcal{B}^{h^B_P}(x) \not\subseteq \mathcal{B}^h(x)\right],$$

where $\mathcal{B}^h(x)$ is a largest ball around x that a classifier h labels homogeneously. We also define an empirical version of the adaptive robust loss.

- We introduce an adaptive data augmentation scheme, evaluate it empirically and prove that it maintains consistency of a nearest neighbor classifier.



Conclusions:

- Robustness is not at odds with accuracy.
- It's important to analyze (and highlight) to what degree mathematical phenomena are artifacts of a framework for analysis.