"Robustness May Be at Odds with Accuracy" Tsipras et. al 2019

A review by Zen Tang
S&DS 659 Statistical Learning Theory

April 23, 2019

1 Background

Viewed as an empirical risk minimization problem, adversarial learning seeks to find a classifier C such that the adversarial loss

$$\mathbb{E}_{(x,y)\in\mathcal{D}} \max_{\delta\in\Delta} \mathcal{L}(x+\delta, y) \tag{1}$$

is minimized. A classifier which can minimize this is termed *robust*, since no matter what δ you pick, the maximum loss is not too great.

Intuitively, one might expect that minimizing this *adversarial loss* is similar to minimizing the actual expected loss, where one wants to achieve the best accuracy on out-of-sample data. However, in this paper, the authors show that this is not the case - in fact, even with infinite data, there is a distinction between the two scenarios.

1.1 Contributions

The main contribution of the authors is to illustrate with a simple statistical model a provable difference between standard and adversarial accuracy.

1.1.1 The binary classification task

The task is as follows: we have a bunch of (x, y) pairs, where $y \in \{-1, +1\}$, and each x_i is sampled from a normal distribution with mean ηy and a variance of 1, *except for the first one*, where x_0 is $+y$ or $-y$ with probability p and $1-p$, respectively. If you want to build a regressor, you can minimize the risk by simply ignoring the first one - the rest of the features, even when η is small, are in the aggregate informative.

However, in the adversarial case, a perturbation $\epsilon > 2\eta$ can shift $x_i, ..., x_n$ to seem like they were sampled from $\mathcal{N}(-\eta y, 1)$ instead.

This serves as a motivating example for why adversarial accuracy could be different from standard accuracy, rooted in the distinction between *robust* features, which are reliable in the face of the adversary (but which might not be incredibly predictive), and *non-robust* features, which boost accuracy in the standard case but are not reliable in the adversarial case. Thus, the reasoning is that a robust classifier has to make do with only the robust features, and miss out on any standard accuracy that looking at the non-robust features would provide.

2 Theory

What are some of the assumptions that this example makes?

If we had infinite data, could we have a standard classifier, like the Bayes classifier, which is also robust?

3 Concluding remarks

Goodfellow 2014 observed that adversarial training has the empirical result of relying on fewer input features - we are essentially pruning the features that are not robust, and hence are more invariant to perturbations. But is there a way that we can capture the nature of this perturbation? (And becovariant?) This would essentially be tantamount to having a manipulable model which can capture not only the distribution of data, but also the nature and distribution of perturbations (!).