6.7900 Fall 2024: Lecture Notes 14

1 Robustness

We discuss distributionally robust optimization. We have training data in the form of (x, y) pairs drawn from a distrubtion P but the test data comes from a distribution Q different from P. However we know Q is not too far, within a KL-divergence of ϵ. We write down the distributionally robust optimization criterion as

$$\min_{\theta} \max_{D(P||Q)} \mathbb{E}_{z\sim Q}[L(z, \theta)]$$

If this optimization were tractable, and the true test distribution Q^* satisfies $D(Q^*||P) \leq \epsilon$. Then the test error is upper bounded by the DRO objective

$$\mathbb{E}_{z\sim Q^*}[L(z, \theta)] \leq \max_{D(P||Q)} \mathbb{E}_{z\sim Q}[L(z, \theta)]$$

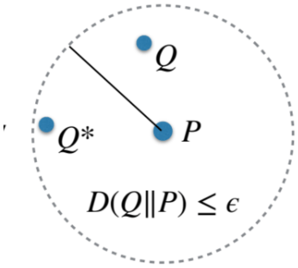

Figure 1: Distributionally robust optimization posits a min max formulation of learning objective to handle any distribution Q^* in KL ball of Q.

1.1 Adversarial Robustness

For adversarial robustness, we want to make sure our learned classifier is robust against any perturbation in a ball.

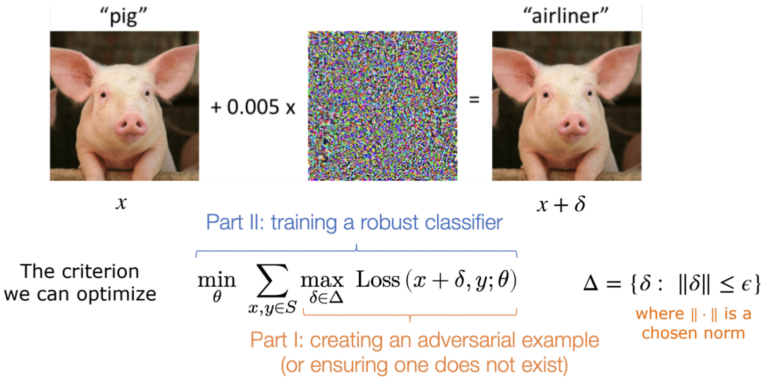
"pig"
+ 0.005 x
=
"airliner"
x
$x+\delta$
Part II: training a robust classifier
The criterion we can optimize
$\min_{\theta} \sum_{x,y \in S} \max_{\delta \in \Delta} \text{Loss}\,(x+\delta, y; \theta)$
$\Delta = \{\delta : \|\delta\| \leq \epsilon\}$
where $\|\cdot\|$ is a chosen norm
Part I: creating an adversarial example (or ensuring one does not exist)

Figure 2: For adversarial robustness we're minimizing a min max criterion where the minimization is taken over learned parameters and the max is taken over a perturbation ball.

$$\min_{\theta} \sum_{x,y \in S} \max_{\delta \in \Delta} \text{Loss}(x+\delta, y; \theta)$$

For finding an adversarial example we can solve the inner maximization

$$\max_{\delta \in \Delta} \text{Loss}(x+\delta, y; \theta)$$

We can solve this via projected gradient descent onto a feasible set Δ in this case the euclidean ball of radius ϵ.

$$\delta = \mathcal{P}_{\Delta}(\delta + \alpha \nabla_{\delta} \text{Loss}(x+\delta, y; \theta))$$

Now we move onto learning the classifier. To do this we use Danskin's theorem.

$$\nabla_{\theta} \max_{\delta \in \Delta} \text{Loss}(x+\delta, y; \theta) = \nabla_{\theta} \text{Loss}(x+\delta^*, y; \theta)$$

where $\delta^* = \max_{\delta \in \Delta} \text{Loss}(x+\delta, y; \theta)$. This implies we can optimize through the max by just finding the maximum value. This leads us to a simple adversarial training algorithm

1. Select minibatch B
2. For each $(x, y) \in B$ compute adversarial example $\delta^*(x)$
3. Update parameters

$$\theta = \theta - \frac{\alpha}{|B|} \sum_{x,y \in B} \nabla_{\theta} \text{Loss}(x + \delta^*(x), y; \theta)$$

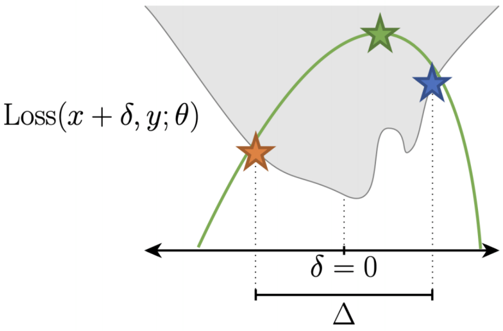

Figure 3: The goal is to find a perturbation δ where the classificaiton moves across the decision boundary.

At this point we list some additional considerations. Yes, we need extra expressive power to find a well-fitting mdoel that is robust on the training set. The sample complexity of robust generalization is greater than for regular learning. As it turns out, it's possible to "certify" robustness and show that no adversarial examples exist nearby a specific new example.

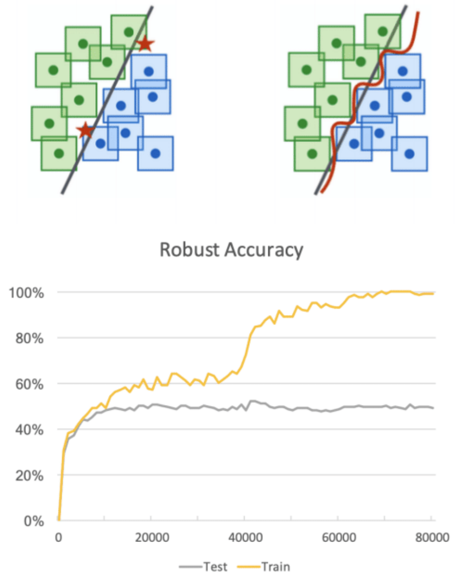

Figure 4: Robust classifiers are harder to learn than brittle classifiers. They require more expressive power at the decision boundary. They also require more data to learn the small fluctuations at the decision boundaries.

1.2 Fairness

Suppose our data distributions consists of K unknown subgroups. We would like to ensure that the error rates of the estimated predictor are the same across these groups to the extent possible. For $z \sim P_k$ being the data drawn from the k'th group (to be clear the group identification is unknown). For $\mathbb{E}_{z\sim P_k}[L(z,\theta)]$ being the expected loss within group k (we cannot calculate this). Let α_k be the fraction of individual that belong to group k. To balance subgroup error we minimize

$$\min_\theta \max_k \mathbb{E}_{z\sim P_k}[L(z,\theta)]$$

We can solve this optimizaiton through a robust DRO objective without any knowledge of the underlying demographics except for the size of the smallest group. Here we observe

$$\max_k \mathbb{E}_{z\sim P_k}[L(z,\theta)] \leq \max_{Q:D_{\chi^2}(Q||P)\leq r}[L(z,\theta)]$$

for $r = \max_k(1/\alpha_k - 1)^2 = (1/\min_k \alpha_k - 1)^2$ where $D_{\chi^2}(Q||P) = \int(\frac{Q(z)}{P(z)} - 1)^2 P(z)dz = \int \frac{Q(z)^2}{P(z)}dz - 1$.