IT581 Adversarial Machine Learning

Lab Assignment - 04

Group 05

2021 17013: Shaikh Faizan Ahmed
2021 17014: Sonam Bharti

Question 1.

$w^t x + w_0 = 0$ is the decision boundary of a linear classifier, and let $x_0 \in R^d$ be an input data point. Suppose we attack the classifier by adding i.i.d. Gaussian noise $r \sim N(0, I)$ to x_0.
Show that the probability of a successful attack $P[\frac{1}{d}\sum_{j=1}^{d} w_j r_j \geq \epsilon]$ at a tolerance level E is upper bounded by,

$$P[\frac{1}{d}\sum_{j=1}^{d} w_j r_j \geq \epsilon] \leq \frac{||w||}{\epsilon\, d\sqrt{2\pi}} e^{-d^2 \frac{\epsilon^2}{2||w||^2}}$$

And with practical example show that, as $d \to \infty$ it becomes gradually more difficult for i.i.d. Gaussian noise to succeed in attacking.
Comment your observation.

Answer

When we apply attack on linear classifier the perturbation will always be $x = x_0 + \lambda w$. It means we want to change x_0 along w by some λ value so that it can miss-classify the model. Now instead of moving along w, we move along some random vector r such that

$$x = x_0 + \sigma_r\, r,$$

where $r \sim \mathcal{N}(0, I)$, then we will see if can we still misclassify the data point x or not. It is clear that it requires to check whether $w^T r > 0$. If $w^T r > 0$, then there will be an acute angle form from r and w which can be seen which are sufficient step size to move x_0 to another class. Otherwise, ($w^T r < 0$), w and r will form an obtuse angle and r will move x_0 to the opposite direction.

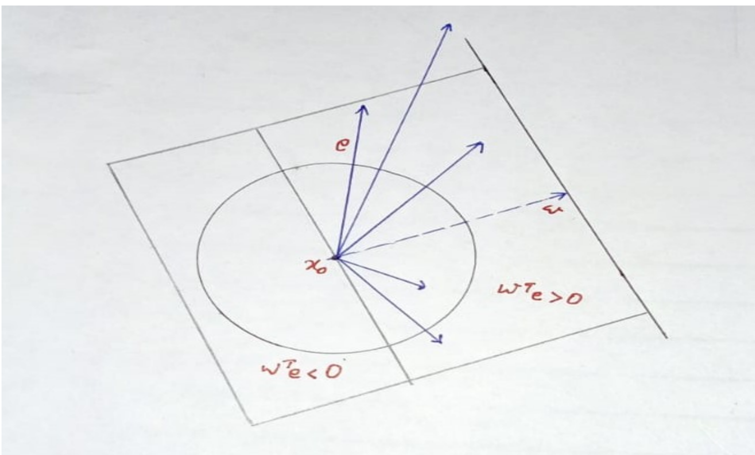

Figure 1: Attacking the linear classifer with i.i.d. noise is equivalent to putting an uncertainty circle around x_0 with radius σ_r.

From the above given figure, our intuition towards this is that i.i.d noise can reach the attack rate of 50% due to the fact that $w^T r > 0$ occupies half of the space. The problem of such argument is that in high dimension, the probability of $w^T r > 0$ is diminishing very quickly as the dimensionality of r grows. This is a well-known phenomenon called curse of dimensionality. To illustrate the idea, let us evaluate the probability of $w^T r \geq \epsilon$ for some $\epsilon > 0$. To this end, let us consider

$$P[\frac{1}{d} w^T r \geq \epsilon] \; = \; P[\frac{1}{d} \sum_{j=1}^{d} w_j r_j \geq \epsilon]$$

where d is the dimensionality of w, i.e., $w \in R^d$. The tolerance level ϵ is a small positive constant that stays away from 0.

Now, let $Y \; = \; \frac{1}{d} \sum_{j=1}^{d} w_j r_j$. Since a linear combination of Gaussian remains a Gaussian, it holds that Y is Gaussian and

$$\mu \; = \; E[Y] \; = \; 0, \qquad \textit{and} \qquad \sigma^2 \; = Var[Y] = \; \frac{1}{d^2} \sum_{j=1}^{d} w_j^2 \; = \; \frac{||w||^2}{d^2}$$

Therefore, by substituting $\epsilon = \sigma\epsilon$ we can show that

$$P[\frac{1}{d} \sum_{j=1}^{d} w_j r_j \geq \epsilon] = P[Y \geq \epsilon] \leq \frac{\sigma}{\epsilon} \frac{e^{-\frac{\epsilon^2}{2\sigma^2}}}{\sqrt{2\pi}} = \frac{||w||}{\epsilon d \sqrt{2\pi}} e^{-d^2 \frac{\epsilon^2}{2||w||^2}}$$

As $d \to \; \infty$, it holds that . That means, the probability of getting a attack direction" is diminishing to zero exponentially. Putting everything together we have the following theorem.

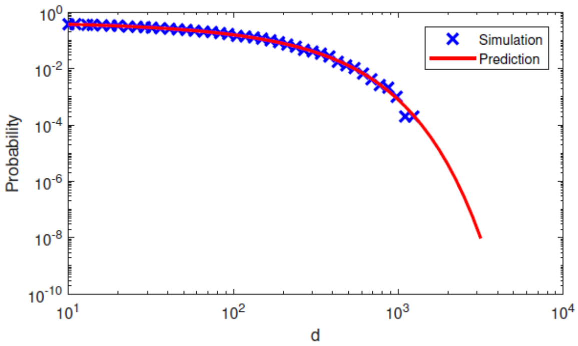

Figure 2: Empirical probability and the theoretically predicted value.

Hence,

Let $w^T x + w_0 \; = \; 0$ be the decision boundary of a linear classifier, and let $x_0 \; \in \; R^d$ be an input data point. Suppose we attack the classifier by adding i.i.d. Gaussian noise where $r \; \sim \mathcal{N}(0, I)$ to x_0. The probability of a successful attack at a tolerance level ϵ is $P[\frac{1}{d} \sum_{j=1}^{d} w_j r_j \geq \epsilon]$, and such probability is upper bounded by

$$P[\frac{1}{d} \sum_{j=1}^{d} w_j r_j \geq \epsilon] \leq \frac{||w||}{\epsilon d \sqrt{2\pi}} e^{-d^2 \frac{\epsilon^2}{2||w||^2}}$$

Therefore, as $d \to \; \infty$ it becomes increasingly more dicult for i.i.d. Gaussian noise to succeed in attacking.

Example

Consider a special case where $w = 1_{dx1}$, i.e., a d-dimensional all-one vector, and $r \sim \mathcal{N}(0, I)$ to x_0. In this case, we de
ne the average as

$$Y \stackrel{\text{def}}{=} \frac{1}{d} \sum_{j=1}^{d} r_j$$

It follows that Y is a Gaussian random variable because linear combination of Gaussian remains a Gaussian. The mean and variance are

$$E[Y] = 0,$$

and

$$Var[Y] = \frac{1}{d}$$

Therefore, the probability of the event $Y > \epsilon$ is

$$P[Y > \epsilon] = \int_{\epsilon}^{\infty} \frac{1}{\sqrt{2\pi/d}} e^{-\frac{t^2}{2/d}}\, dt$$

$$= \int_{\epsilon\sqrt{\frac{d}{2}}}^{\infty} \frac{1}{\sqrt{\pi}} e^{-t^2}\, dt$$

$$= \frac{1}{2} erfc(\epsilon\sqrt{d/2}),$$

where erfc is the complementary error function defined as $erfc(z) \stackrel{\text{def}}{=} \frac{2}{\sqrt{\pi}} \int_{z}^{\infty} e^{-t^2}\, dt$. As we can see in Figure 2, the probability drops rapidly as d increases.

Observation

As $d \to \infty$ it becomes increasingly more dicult for i.i.d. Gaussian noise to succeed in attacking.