AML

Assignment - 04

July 21, 2025

Group - 14

Submitted By :

Rohan Baghel
202116011

Raja Babu Meena
202116010

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 $r \sim (0, I)$ to x_0.
Show that the probability of a successful attack

$\mathbb{P}\left[\frac{1}{d}\sum_{1-j}^{d} w_j r_j \geq \varepsilon\right]$ at a tolerance level E is upper bounded by,

$$\mathbb{P}\left[\frac{1}{d}\sum_{1-j}^{d} w_j r_j \geq \varepsilon\right] \leq \frac{||w||}{\varepsilon d\sqrt{2\pi}} e^{-d^2 \frac{\varepsilon^2}{2||w||^2}}. \tag{1}$$

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

Sol

In Linear classifier we have seen that
$x = x_0 + \lambda w$
i.e. we move x_o along w by amount λ to misclassify the x.

if we move along w but along a random vector r, such that

$$x = x_0 + \sigma_r r, \qquad where, r \sim N(0, I) \tag{2}$$

- if $w^T r > 0$, then r and w will form an acute angle and so for sufficient step size we will be able to move x_0 to another class.
- if $w^T r < 0$, then w and r are form an obtuse angle and so r will move x_0 to an opposite direction.

Attacking the linear classifier with i.i.d. noise is equivalent to putting an uncertainty circle around x_0 with radius $_r$. The possible attack directions are those that form acute angle with the normal vector w. Therefore, among all the possible r's, we are only interested in those such that $w^T r > 0$, which occupy half of the space.

Now let us illustrate probability of $w^T r \geq \epsilon$ for $\epsilon > 0$
For this let us consider

$$\mathbb{P}\left[\frac{1}{d} w^T r \geq \varepsilon\right] = \mathbb{P}\left[\frac{1}{d}\sum_{1-j}^{d} w_j r_j \geq \varepsilon\right] \tag{3}$$

Here

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

Now We know :
According to Central Limit Theorem,
if we look at the inner product

$$w^T r = \sum_{j-1}^{d} w_j r_j \tag{4}$$

if $\mathbb{E}[r_i] = 0$ for all i
We can see that if we increase d then the random estimate

$$\sum_{j-1}^{d} w_j r_j \tag{5}$$

approaches its expectation.

Other than his we can see that for $r_i's$ there are +ve and -ve values are present in sum.most likes terms will cancel out each other.

If we look after the high-dimensional space, the concentration inequality says that, instead of having a half sphere event id actually concentrated at the high dimension sphere. event is

$$\frac{1}{d} \sum_{j-1}^{d} w_j r_j \tag{6}$$

We can say that if we increase the value of d then, there is very less probability or chance that we can find a point not on the equator of the sphere.

We can determine the i.i.d noise attack magnitude by determining the cosine angle,

$$cos\theta = \frac{w^T}{||w||_2 ||r||_2} \tag{7}$$

which is equivalent to

$$\frac{\lambda *}{\sigma_e} \tag{8}$$

If we take angle between w and r then $cos\theta$ is

$\frac{w^T r}{||w||_2||r||_2} = cos\theta$

since, The shortest distance between x_0 and decision boundary $w^T x + w_0 = 0 is \lambda *$

The distance from x_0 to the decision boundary along the direction r is

$$cos\theta = \frac{\lambda *}{\sigma_e} \tag{9}$$

Example :

Let's consider $w = 1_{dx1}$

i.e. a d-dimensional all-one vector, and $r \sim \mathbb{N}(0, I)$

in this case we define the Avg as.

$$\mathbb{Y} \stackrel{def}{=} \frac{1}{d} \sum_{j=1}^{d} r_j. \tag{10}$$

We know, linear combination of Gaussian remains a Gaussian.

So, $\mathbb{Y}$ is a Gaussian random variable

- Mean $\mathbb{E}[Y] = 0$
- Variance $Var[Y] = \frac{1}{d}$

$\therefore$ the probability of event $Y > \epsilon$ is

$$\mathbb{P}[Y > \epsilon] = \int_{\epsilon}^{\infty} \frac{1}{\sqrt{2\pi/d}} exp\left\{-\frac{t^2}{2/d}\right\} dt \tag{11}$$

$$= \frac{1}{2} erfc\left(\epsilon\sqrt{d/2}\right) \tag{12}$$

here, erfc is complementary arror function.

Now, if Y is a Gaussian random variable with $Y \sim N(\mu, \sigma)$, then

$$\mathbb{P}\left[Y \geq \mu + \sigma\epsilon\right] \leq \frac{1}{\epsilon} \frac{e^{-\Sigma^2} \Sigma \sqrt{2\pi}}{}$$

equation

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

$$\mu = \mathbb{E}[Y] = 0 and \sigma^2 = Var[Y] = \frac{1}{d^2}\sum^{d} j = 1 w_j^2 = \frac{||w||^2}{d^2} \tag{14}$$

$\therefore bysubstituting \epsilon = \sigma\epsilon$

for d $\rightarrow \infty$, it holds that $\mathbb{P}[\frac{1}{d}\sum^{d} j = 1 w_j r_j \geq \epsilon] \rightarrow 0$
That means, the probability of getting a "good attack direction" is diminishing to zero exponentially.