ECE 595: Homework 5

Yi Qiao, Class ID 187
(Spring 2019)

Exercise 1: Adversarial Attacks on Gaussian Classifier

(a) minimum-norm attacks

(i) minimum l_2 and l_∞ attack

Since we only have 2 classes, the question becomes

$$\begin{aligned} &\underset{\boldsymbol{x}}{minimize}\ ||\boldsymbol{x}-\boldsymbol{x}_0|| \\ &subject\ to\ \boldsymbol{w}^T\boldsymbol{x}+w_0=0 \end{aligned} \tag{1}$$

using l_2 norm
the problem is the same as

$$\begin{aligned} &\underset{\boldsymbol{x}}{minimize}\ \frac{1}{2}||\boldsymbol{x}-\boldsymbol{x}_0||_2^2 \\ &subject\ to\ \boldsymbol{w}^T\boldsymbol{x}+w_0=0 \end{aligned} \tag{2}$$

The lagrangian is

$$\mathcal{L}(\boldsymbol{x},\lambda)=\frac{1}{2}||\boldsymbol{x}-\boldsymbol{x}_0||_2^2+\lambda(\boldsymbol{w}^T\boldsymbol{x}+w_0) \tag{3}$$

Taking the derivative

$$\begin{aligned} \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\lambda)&=\boldsymbol{x}-\boldsymbol{x}_0+\lambda\boldsymbol{w}=0 \\ \frac{\partial}{\partial\lambda}\mathcal{L}(\boldsymbol{x},\lambda)&=\boldsymbol{w}^T\boldsymbol{x}+w_0=0 \end{aligned} \tag{4}$$

$$\begin{aligned} \lambda\boldsymbol{w}&=\boldsymbol{x}_0-\boldsymbol{x} \\ \lambda\boldsymbol{w}^T\boldsymbol{w}&=\boldsymbol{w}^T\boldsymbol{x}_0-\boldsymbol{w}^T\boldsymbol{x} \\ \lambda&=\left(\boldsymbol{w}^T\boldsymbol{w}\right)^{-1}\left(\boldsymbol{w}^T\boldsymbol{x}_0+w_0\right) \end{aligned} \tag{5}$$

$$\begin{aligned} \boldsymbol{x}&=\boldsymbol{x}_0-\lambda\boldsymbol{w} \\ &=\boldsymbol{x}_0-\frac{\boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{x}_0+w_0)}{||\boldsymbol{w}||_2^2} \end{aligned} \tag{6}$$

using l_∞ norm

$$\begin{aligned} &\underset{\boldsymbol{x}}{minimize}\ ||\boldsymbol{x}-\boldsymbol{x}_0||_\infty \\ &subject\ to\ \boldsymbol{w}^T\boldsymbol{x}+w_0=0 \end{aligned} \tag{7}$$

Let $\boldsymbol{r}=\boldsymbol{x}-\boldsymbol{x}_0$, $b_0=-(\boldsymbol{w}^T\boldsymbol{x}_0+w_0)$, the problem becomes:

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}||\boldsymbol{x}-\boldsymbol{x}_0||_\infty \\ &subject\ to\ \boldsymbol{w}^T\boldsymbol{r}=b_0 \end{aligned} \tag{8}$$

The lagrangian is

$$\mathcal{L}(\boldsymbol{r}, \lambda) = ||\boldsymbol{r}||_\infty + \lambda(b_0 - \boldsymbol{w}^T\boldsymbol{r}) \tag{9}$$

Taking derivative,

$$\frac{\partial}{\partial\lambda}\mathcal{L}(\boldsymbol{r}, \lambda) = b_0 - \boldsymbol{w}^T\boldsymbol{r} = 0 \tag{10}$$

By Holder's Inequality:

$$\begin{aligned} |b_0| = |\boldsymbol{w}^T\boldsymbol{r}| &\leq ||\boldsymbol{w}||_1||\boldsymbol{r}||_\infty \\ ||\boldsymbol{r}||_\infty &\geq \frac{|b_0|}{||\boldsymbol{w}||_1} \end{aligned} \tag{11}$$

Consider $\boldsymbol{r} = \eta \cdot sign(\boldsymbol{w})$, for some constant η tbd. We can show that

$$||\boldsymbol{r}||_\infty = \underset{i}{argmax}\ |\eta \cdot sign(w_i)| = |\eta| \tag{12}$$

let $\eta = \frac{b_0}{||\boldsymbol{w}||_1} \cdot sign(\boldsymbol{w})$, then we have,

$$||\boldsymbol{r}||_\infty = |\eta| = \frac{b_0}{||\boldsymbol{w}||_1} \tag{13}$$

Lower bound is achieved, thus the solution is,

$$\boldsymbol{r} = \frac{|b_0|}{||\boldsymbol{w}_1||} \cdot sign(\boldsymbol{w}) \tag{14}$$

(ii) DeepFool attack

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}\ ||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 \\ &subject\ to\ g(\boldsymbol{x}) = 0 \end{aligned} \tag{15}$$

First order approximation

$$g(\boldsymbol{x}) \approx g(\boldsymbol{x}^{(k)}) + \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})^T(\boldsymbol{x} - \boldsymbol{x}^{(k)}) \tag{16}$$

Then the problem can be approximate by

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}\ ||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 \\ &subject\ to\ g(\boldsymbol{x}^{(k)}) + \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})^T(\boldsymbol{x} - \boldsymbol{x}^{(k)}) = 0 \end{aligned} \tag{17}$$

Let $\boldsymbol{w}^{(k)} = \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})$ and $w_0^{(k)} = g(\boldsymbol{x}^{(k)}) - \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})^T\boldsymbol{x}^{(k)}$
Then the problem is equivalent to

$$\begin{aligned} &\underset{\boldsymbol{x}}{argmin}\ ||\boldsymbol{x} - \boldsymbol{x}_0||_2^2 \\ &subject\ to\ (\boldsymbol{w}^{(k)})^T\boldsymbol{x} + w_0^{(k)} = 0 \end{aligned} \tag{18}$$

This is the same problem as minimum l_2 norm attack, Thus the solution will be,

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{((\boldsymbol{w}^{(k)})^T x^{(k)} + w_0^{(k)})\boldsymbol{w}^{(k)}}{||\boldsymbol{w}^{(k)}||_2^2} \tag{19}$$

substitute $\boldsymbol{w}$ and w_0 back, we get

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(\frac{g(\boldsymbol{x}^{(k)})}{||\nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)})||^2}\right) \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{(k)}) \tag{20}$$

(iii) An example DeepFool never converge