Dongheng Lin
2023/11/7

CS 442: Trustworthy Machine Learning
Homework 3

Solutions.

Q1

1.1

Since we know that the MSE loss is a convex function, by Jensen's Inequality, we know that for a convex function $\phi(z)$ and a random variable Z:

$$\phi(\mathbb{E}[Z]) \leq \mathbb{E}[\phi(Z)]$$

Applying this to the mean squared error (MSE), where $\phi(z) = z^2$ and $Z = f(x) - y$, we have:

$$(\mathbb{E}[f(x) - Y|X = x])^2 \leq \mathbb{E}\left[(f(x) - Y)^2|X = x\right]$$

Since $f(x)$ is a deterministic function and not a random variable, $\mathbb{E}[f(x)|X = x] = f(x)$. Thus, the inequality simplifies to:

$$(f(x) - \mathbb{E}[Y|X = x])^2 \leq \mathbb{E}\left[(f(x) - Y)^2|X = x\right]$$

This expression is minimized when $f(x) = \mathbb{E}[Y|X = x]$, which makes the left-hand side of the inequality zero. Therefore, the function $f^*(x)$ that minimizes the expected squared loss is the conditional expectation of Y given $X = x$:

$$f^*(x) = \mathbb{E}[Y|X = x]$$

This is the Bayes optimal predictor under the mean squared error criterion.

Q2

2.1

Since x_1 is directly correlated with y, we want to give it a high weight. Let's denote this weight as w_1.

If we set $w_1 = \alpha$, and $w_i = \beta, i = 2 \ldots d$. The classifier is then:

$$f_w(x) = \text{sgn}(w_1 x_1 + \sum_{i=2}^{d} \beta x_i)$$

Since x_i follows $\mathcal{N}(2y/\sqrt{d}, 1)$, then:

$$\sum_{i=2}^{d} \beta x_i \sim \mathcal{N}(2(d-1)\beta y/\sqrt{d}, (d-1)\beta^2)$$

We simply set $\beta = \frac{\sqrt{d}}{2\sqrt{d-1}}$, therefore we have the classifier:

$$\begin{aligned} f_w(x) &= \text{sgn}(\alpha x_1/\sqrt{d} + Z) \\ Z &\sim \mathcal{N}(y, \frac{d}{4}) \end{aligned}$$

Since $Z \sim \mathcal{N}(y, \frac{d}{4})$, which means $\Pr(Z-y \leq -t\frac{d}{4}) \leq \exp\left(-t^2/2\right), \forall t \geq 0$ and Gaussian distribution is symmetric along the vertical line of mean value, we know that when $x_1 y > 0$, the probability for a wrong prediction can be calculated:

$$\begin{aligned} & \Pr_{y=1,x=1}(Z+\alpha \leq 0) + \Pr_{y=-1,x=-1}(Z-\alpha > 0) \\ = & \; 2\Pr_{y=1,x=1}(Z+\alpha \leq 0) \\ = & \; 2\Pr_{y=1,x=1}(Z - y \leq -\alpha - 1) \\ = & \; 2e^{-(\frac{4\alpha-1}{d})^2/2} \end{aligned}$$

Similarly, the probability of making a correct prediction when $x_1 y < 0$ is then:

$$\begin{aligned} & 2\Pr_{y=-1,x=1}(Z+\alpha \leq 0) \\ = & \; 2\Pr_{y=-1,x=1}(Z - y \leq -\alpha + 1) \\ = & \; 2e^{-(\frac{4\alpha+1}{d})^2/2} \end{aligned}$$

Therefore, we can express the general accuracy as:

$$p * (1 - 2e^{-(\frac{4\alpha-1}{d})^2/2}) + (1-p) * (2e^{-(\frac{4\alpha+1}{d})^2/2}) \geq 0.85$$

Since $0.5 < p \leq 0.8, d \geq 25$, it's easy to find a solution for inequality above such that the accuracy for f_{w_n} is at least 0.85.

e.g. when $p = 0.5$, take $\frac{(\frac{4\alpha}{d})^2}{2} \approx 1.38$ such that $e^{\frac{(\frac{4\alpha}{d})^2}{2}} \approx 0.24$:

$$accuracy \approx 0.5(1 - 0.24 * 2) + 0.5 * (0.24 * 2) \approx 1$$

2.2.1

Since we know that the adversary budget $\epsilon = \frac{4}{\sqrt{d}} \leq |y|$, It is manifest that any perturbation to x_1 can not have impact on the sign of the output if we set $w'_1 = 1, w'_i = 0, \forall i \geq 2$, because in such w', the classifier f' is:

$$f_{w'}(x + \triangle x) \quad = \quad \text{sgn}(x_1 + \triangle x_1)$$

And for $w \in \mathbb{R}^d, \exists i \geq 2, w_i \neq 0$ such that we have:

$$f_{w'}(x + \triangle x) \quad = \quad \text{sgn}(x_1 + \triangle x_1 + \sum_{i=2}^{d} w_i(x_i + \triangle x_i))$$

To compare the $\ell_r(w'), \ell_r(w)$ consider the impact of the ℓ^∞ perturbation. For w, the adversary can induce additional misclassification by perturbing $x_2, \ldots, x_d$. However, for w', the adversary's impact is limited as x_1 can only take values $\pm y$, and small perturbation by at most ϵ do not change its sign. Thus, there exists a $\ell_r(w')$ is less than $\ell_r(w)$, proving that a better classifier in terms of robust error exists.

2.2.2

In 2.2.1 we know that $w' = w_r$ remains robust under any perturbation on x_1, since $|x_1| > \epsilon \geq \triangle x_1$, the maximum 0-1 loss for such classifier is thus:

$$w_r \quad = \quad \{1, 0, 0, \ldots, 0\}$$

For w_r, the classifier decision is based solely on the sign of x_1, which matches the label y with probability p, since $x_1 = +y$ with probability p and $x_1 = -y$ with probability $1 - p$.

Therefore, the robust error $\ell_r(w_r)$ is the probability of misclassification under the worstcase perturbation. However, since w_r ignores $x_2, \ldots, x_d$, the only source of error is when x_1 does not match y, which happens with probability $1 - p$. Therefore, $\ell_r(w_r) = 1 - p$.

$$\ell_r(w_r) \quad = \quad \mathbb{E}\left[\max_{\|\Delta x\|_\infty \leq \frac{4}{\sqrt{d}}} \ell_{01}\left(f_{w_r}(x + \Delta x), y\right)\right] = Pr(x_1 = -y) = 1 - p$$

2.3

According to 2.2.2, f_{w_r} is a classifier that relies solely on the first feature x_1, with $w_r = (1, 0, 0, \ldots, 0)$. The standard error is computed as $E\left[\ell_{01}\left(f_{w_r}(X), Y\right)\right]$, which is the expected 0-1 loss.

$$\mathbb{E}\left[\max_{\|\Delta x\|_\infty \leq \frac{4}{\sqrt{d}}} \ell_{01}\left(f_{w_r}(x + \Delta x), y\right)\right] = Pr(x_1 = -y) = 1 - p$$

Compared with other w setting that we have shown in 2.1 which achieves higher accuracy than w_r, we have seen that there is a non-zero gap in terms of the standard accuracy between the robust classifier and the original classifier.

Q3

3.1

Since we have $c \geq 0$, to optimize $\min_{t,x} c^\top t$, we are essentially minimizing each component of t subject to the constraints:

$$t = \text{ReLU}(Ax)$$

But the constraints is not a linear function and does not span a convex feasible region for (Ax, t), making it unsolvable via Linear Programming.

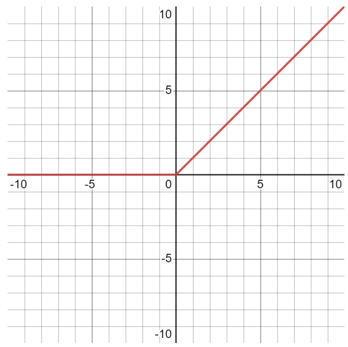

Figure 1: ReLU function

To linearize this, we need to express it in terms of linear inequalities:

- $t \geq 0$: This constraint comes directly from the definition of ReLU. Since ReLU never outputs negative values, t, which is the output of ReLU, must be non-negative.
- $t \geq Ax$: This constraint represents the case where $Ax > 0$. In this scenario, the ReLU function outputs Ax itself, so t, being the output of ReLU, must be at least Ax. This constraint does not contradict the case where $Ax \leq 0$ because when Ax is negative or zero, $t \geq Ax$ is still valid as $t \geq 0$ or more.

Therefore, after such relaxation, the convex feasible region is:

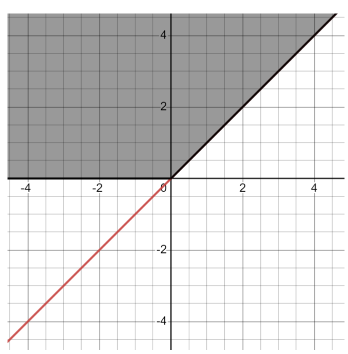

Figure 2: Linear Constraints

It is easy to find that the $\min_{t,x} c^\top t$ must have $t = \max(0, Ax)$, since they are lower bound of the constraints, which is always equals to the original constraint $t = \text{ReLU}(Ax)$

3.2.1

No, In 3.1, we have shown that the objective function is linear, specifically $c^T t$, and the ReLU function $t = \text{ReLU}(Ax)$ can be linearized under the condition that $c \geq 0$

But for (1)the objective function is $(e_y - e_t)^T (W_2 z_2)$, where the previous restrictions on each component of linear combinations no longer holds.

Therefore, minimizing the objective function is no longer equivalent to minimizing all the component of t, formally we have:

$$c' = (e_y - e_t) = \{0, 1, 0 \ldots, -1, 0\} \Longrightarrow \exists c'_i < 0$$

Therefore, under such scenario, minimizing the objective function involves maximizing corresponding i-th component of $W_2 z_2$ and therefore all the contributing element in $\text{ReLU}(W_1 z_1)$, but if we follow the same relaxation in 3.1, we will have no upper bound available to determine the max value of z_2 even with restriction $||z_1 - x||_\infty \leq \epsilon$. Hence we have proved that the method in 3.1 is not applicable to 3.2.

3.2.2

To show the 2 formulations are equivalent, we first consider a = 0, Then we have the constraints:

- $t - x \geq 0$
- $t \geq 0$
- $-t \geq 0$
- $x - l - t \geq 0$

Since we know that $l \leq x \leq u$, then $x - l \geq 0$, combining with the other constraints, they eventually simplifies to $y = 0$, then we consider the case $a = 1$, the constraints are then:

- $t - x \geq 0$
- $t \geq 0$
- $u - t \geq 0$
- $x - t \geq 0$

Similarly, we can simplify them to $t = x$.

Since $\text{ReLU}(x) = \max(0, x)$, if we set a = 0 when$x <= 0$ and $a = 1$ when $x > 0$. it perfectly works as a ReLU.

3.2.3

The number of auxiliary binary variables introduced corresponds to the number of neurons in the hidden layer of the network, which is the dimension of z_2 in the case given.

Therefore, If W_1 is a matrix of size $p \times d$, then there are p neurons in the hidden layer, and thus p binary variables a_i are introduced.

Q4

4.1

Start with the Definition of Total Variation Distance: Consider the total variation distance between the distributions of neighboring datasets $M(X)$ and $M\left(X^{\prime}\right)$:

$$d_{TV}\left(M(X), M\left(X^{\prime}\right)\right)=\frac{1}{2} \sum_{t \in Y}\left|\operatorname{Pr}[M(X)=t]-\operatorname{Pr}\left[M\left(X^{\prime}\right)=t\right]\right|$$

Using the definition of ε-differential privacy, we know that for each $t \in Y$:

$$\begin{aligned} \operatorname{Pr}(M(X)=t) & \leq e^{\varepsilon} \cdot \operatorname{Pr}\left(M\left(X^{\prime}\right)=t\right) \\ \operatorname{Pr}\left(M\left(X^{\prime}\right)=t\right) & \leq e^{\varepsilon} \cdot \operatorname{Pr}(M(X)=t) \end{aligned}$$

These inequalities imply that:

$$\left|\operatorname{Pr}(M(X)=t)-\operatorname{Pr}\left(M\left(X^{\prime}\right)=t\right)\right| \leq\left(e^{\varepsilon}-1\right) \cdot \max \left(\operatorname{Pr}(M(X)=t), \operatorname{Pr}\left(M\left(X^{\prime}\right)=t\right)\right)$$

Summing this inequality over all $t \in Y$ gives:

$$\textstyle\sum_{t \in Y}\left|\operatorname{Pr}(M(X)=t)-\operatorname{Pr}\left(M\left(X^{\prime}\right)=t\right)\right| \leq\left(e^{\varepsilon}-1\right) \cdot \sum_{t \in Y} \max (\operatorname{Pr}(M(X)=t), \operatorname{Pr}\left(M\left(X^{\prime}\right)=t\right))$$

Since the sum of probabilities over all possible outcomes t for a probability distribution is 1, we have:

$$\sum_{t \in Y} \max \left(\operatorname{Pr}(M(X)=t), \operatorname{Pr}\left(M\left(X^{\prime}\right)=t\right)\right) \leq 1$$

Therefore:

$$\frac{1}{2} \sum_{t \in Y}\left|\operatorname{Pr}(M(X)=t)-\operatorname{Pr}\left(M\left(X^{\prime}\right)=t\right)\right| \leq \frac{1}{2}\left(e^{\varepsilon}-1\right)$$

Since ϵ is small, then we have $e^{\epsilon} \approx 1+\epsilon$, therefore we can say that:

$$\begin{aligned} d_{TV}\left(M(X), M\left(X^{\prime}\right)\right) & =\frac{1}{2} \sum_{t \in Y}\left|\operatorname{Pr}[M(X)=t]-\operatorname{Pr}\left[M\left(X^{\prime}\right)=t\right]\right| \\ & \leq \frac{1}{2}\left(e^{\epsilon}-1\right) \\ & \leq \frac{1}{2} \epsilon \\ & \leq \epsilon \end{aligned}$$

4.2

For datasets X and X' differing in one position, differential privacy guarantees that for any $T \subseteq Y$:

$$\operatorname{Pr}(M(X) \in T) \leq e^{\varepsilon} \cdot \operatorname{Pr}\left(M\left(X'\right) \in T\right)$$

Consider datasets X and X' differing in k positions. We can think of transitioning from X to X' through k intermediate datasets $X_1, X_2, \ldots, X_{k-1}$, where each X_i differs from X_{i-1} in exactly one position (with $X_0 = X$ and $X_k = X'$).

Applying the differential privacy guarantee to each pair of neighboring datasets in the sequence, we get:

$$\operatorname{Pr}\left(M\left(X_{i-1}\right) \in T\right) \leq e^{\varepsilon} \cdot \operatorname{Pr}\left(M\left(X_i\right) \in T\right), \text{for } i=1,2, \ldots, k.$$

Chaining these inequalities together, we obtain:

$$\begin{aligned} & \operatorname{Pr}(M(X) \in T) \\ = & \operatorname{Pr}\left(M\left(X_0\right) \in T\right) \leq e^{\varepsilon} \cdot \operatorname{Pr}\left(M\left(X_1\right) \in T\right) \leq \ldots \leq e^{k \varepsilon} \cdot \operatorname{Pr}\left(M\left(X_k\right) \in T\right)=e^{k \varepsilon} \cdot \operatorname{Pr}\left(M\left(X'\right) \in T\right) \end{aligned}$$

Therefore, we have shown that:

$$\operatorname{Pr}(M(X) \in T) \leq \exp(k \varepsilon) \cdot \operatorname{Pr}\left(M\left(X'\right) \in T\right)$$

Q5

5.1

Since $Z \sim \text{Lap}(\frac{1}{n\epsilon})$ with location 0 and scale parameter $\frac{1}{n\epsilon}$, by definition we have:

$$\text{Var}(Z) = \frac{2}{(n\epsilon)^2} \Rightarrow \sigma = \sqrt{\frac{2}{(n\epsilon)^2}}$$

Since we are essentially trying to bound the Z noise between $\frac{10}{n\epsilon}$, That is to say, we want to find a bound such that the probability is at least 0.95, or conversely, the probability of deviation beyond this bound is at most 0.05. By Chebyshev's inequality, we know that for $Z \sim \text{Lap}(0, \frac{1}{n\epsilon})$ [1]:

$$\Pr(|Z| \geq k\sigma) = \frac{1}{k^2} \leq 0.05$$

Setting $\frac{1}{k^2} = 0.05$ we solve for k:

$$\frac{1}{k^2} = 0.05$$
$$k = \sqrt{20}$$

We can then apply this k to the standard deviation of Z:

$$\begin{aligned} k\sigma &= \sqrt{20} \times \sqrt{\frac{2}{(n\epsilon)^2}} \\ &= \sqrt{40} \times \frac{1}{n\epsilon} \end{aligned}$$

Since $\frac{-10}{n\epsilon} < \frac{-\sqrt{40}}{n\epsilon} < \frac{\sqrt{40}}{n\epsilon} < \frac{10}{n\epsilon}$, this suggests that the bound provided in the question is within the range determined by Chebyshev's inequality. Therefore, the inequality holds with a probability of at least 0.95.

5.2

For the Laplace distribution $Z \sim \text{Lap}(0, \frac{1}{n\epsilon})$ we can write the corresponding PDF as [2]:

$$f\left(z \mid 0, \frac{1}{n\epsilon}\right) = n\epsilon \exp(-n\epsilon|z|)/2$$

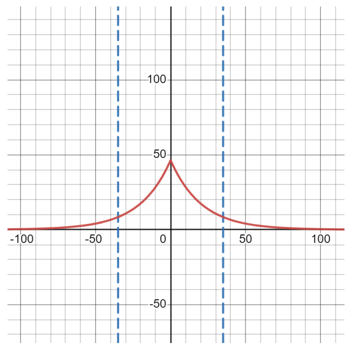

Figure 3: Laplace PDF

Due to the fact that it is symmetric along the line $z = 0$, the region for $Pr(|Z| \geq \frac{t}{n\epsilon})$ is thus:

$$\begin{aligned}
Pr(|Z| \geq \frac{t}{n\epsilon}) &= 2 \times Pr(Z \geq \frac{t}{n\epsilon}) \\
&= 2\int_{\frac{t}{n\epsilon}}^{\infty} n\epsilon \exp(-n\epsilon z)/2dz \\
&= 2\left[-\frac{1}{2}\exp(n\epsilon z)\right]_{\frac{t}{n\epsilon}}^{\infty} \\
&= 2\left[0 - \left(-\frac{1}{2}\exp(-t)\right)\right] \\
&= \exp(-t)
\end{aligned}$$

5.3

Since we know that: $Pr(|Z| \geq \frac{t}{n\epsilon}) = \exp(-t)$ from 5.2, to prove the inequality given, we are essentially looking for the probability of $\frac{-t}{n\epsilon} \leq Z \leq \frac{t}{n\epsilon}$ is at least 0.95:

$$\begin{aligned}
Pr(|Z| \geq \frac{t}{n\epsilon}) &\leq 0.05 \\
\exp(-t) &\leq 0.05 \\
t &\geq -\ln(0.05)
\end{aligned}$$

Since $-\ln(0.05) < 3$, Therefore, we can say that with probability at least 0.95, the noise Z will be within $\pm\frac{3}{n\epsilon}$.
Adding the upper/lower bounds of Z back thus recovers the inequality we would like to verify, hence completes the proof.

References

[1] Wikipedia contributors. Chebyshev's inequality — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Chebyshev%27s_inequality&oldid=1182723320, 2023. [Online; accessed 13-November-2023].

[2] Wikipedia contributors. Laplace distribution — Wikipedia, the free encyclopedia, 2023. [Online; accessed 13-November-2023].