Homework 3

Spring 25, CS 442: Trustworthy Machine Learning
Due Fri. Apr. 25th at 23:59 CT

Instructor: Han Zhao

Instructions for submission All the homework submissions should be typeset in LaTeX. For all the questions, please clearly justify each step in your derivations or proofs.

1 Bayes Optimal Predictor for Regression [10pts]

Let μ be a joint distribution over $X \times Y$, where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. In this problem we are interested in deriving the Bayes optimal predictor for regression under the mean-squared error. More specifically, we are interested in finding a function $f^*(\cdot)$ that minimizes the following expected squared loss:

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mu} \left[(f(x) - y)^2\right],$$

where $\mathcal{F}$ is the function class that contains all possible functions from $\mathbb{R}^p$ to $\mathbb{R}$. Derive an analytical solution for f^*.

Solution:

$$\mathbb{E}\left[(f(X) - Y)^2\right] = \mathbb{E}\Big[\mathbb{E}\left[(f(X) - Y)^2 \mid X\right]\Big] = \mathbb{E}\Big[(f(X) - \mathbb{E}[Y \mid X])^2 + \mathrm{Var}\big(Y \mid X\big)\Big].$$

Since $\mathrm{Var}(Y \mid X)$ does not depend on f, minimizing the above expectation is equivalent to minimizing

$$\mathbb{E}\Big[(f(X) - \mathbb{E}[Y \mid X])^2\Big].$$

Therefore, for each fixed $x \in \mathbb{R}^p$,

$$f^*(x) = \arg\min_{a \in \mathbb{R}} \ \mathbb{E}\left[(a - Y)^2 \mid X = x\right] = \mathbb{E}[Y \mid X = x].$$

Thus,

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

2 Tradeoff between Robustness and Accuracy [30pts]

In this problem we will show the potential tradeoff between robustness and accuracy for linear classifiers under a constructed distribution that we have seen in the lectures. Consider a binary classification task as follows. To sample a data from the distribution, we first sample uniformly at random

$y \in \{+1, -1\}$, i.e., $\Pr(Y = 1) = \Pr(Y = -1) = 1/2$. Given the label of interest y, there are d features, namely $x_1, \ldots, x_d$, whose distributions are defined as follows:

$$x_1 = \begin{cases} +y, & \text{w.p. } p \\ -y, & \text{w.p. } 1-p \end{cases}, \quad x_2, \ldots, x_d \sim \mathcal{N}(2y/\sqrt{d}, 1),$$

where $0.5 < p \leq 0.8$, $d \geq 25$ and $\mathcal{N}(2y/\sqrt{d}, 1)$ is the univariate Gaussian distribution with mean $2y/\sqrt{d}$ and variance 1.

To demonstrate the impact of enforcing robustness, we focus on using linear classifiers for classification. More specifically, let $w \in \mathbb{R}^d$, the classifier we are going to use is given by

$$f_w(x) := \operatorname{sgn}(w^\top x).$$

2.1 [10pts]

Show that there exists a linear classifier $w_n \in \mathbb{R}^d$, such that the standard accuracy of $f_{w_n}(\cdot)$ on this problem is at least 0.85.

Hint: You can use the following tail bound for Gaussian distribution. Suppose $Z \sim \mathcal{N}(\mu, \sigma^2)$, then $\Pr(Z - u \leq -t\sigma) \leq \exp(-t^2/2)$ for any $t \geq 0$.

Solution:

Let's choose the weight vector

$$w_n = (0, 1, 1, \ldots, 1)^\top \in \mathbb{R}^d.$$

Then for $X, Y \sim \mu$,

$$w_n^\top X = \sum_{i=2}^{d} x_i,$$

and conditional on $Y = y$,

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

Hence

$$\Pr\big(f_{w_n}(X) \neq Y\big) = \Pr\big(Y\, w_n^\top X \leq 0\big) = \Pr\Big(Z - (d-1)\tfrac{2}{\sqrt{d}} \ \leq\ -\tfrac{(d-1)2}{\sqrt{d}}\Big),$$

where $Z \sim \mathcal{N}(0, d-1)$. Applying the Gaussian tail bound with

$$t = \frac{(d-1)2/\sqrt{d}}{\sqrt{d-1}},$$

we get

$$\Pr\big(f_{w_n}(X) \neq Y\big) \ \leq\ \exp(-t^2/2) = \exp\Big(-2\,(d-1)/d\Big).$$

For $d \geq 25$, $(d-1)/d \geq 24/25$, so

$$\Pr\big(f_{w_n}(X) \neq Y\big) \ \leq\ \exp(-48/25) \ <\ 0.15,$$

and thus the standard accuracy is

$$1 - \Pr\big(f_{w_n}(X) \neq Y\big) \ \geq\ 0.85.$$

2.2 [15pts]

Now let's consider the ℓ_∞ norm attack with budget $\epsilon = \frac{4}{\sqrt{d}}$ for this problem. In particular, we are interested in finding a linear classifier $w_r \in \mathbb{R}^d$ that minimizes the following robust error under the ℓ_∞ ball:

$$w_r = \arg\min_w \ell_r(w) := \arg\min_w \mathbb{E}\left[\max_{\|\Delta x\|_\infty \le \frac{4}{\sqrt{d}}} \ell_{01}(f_w(x+\Delta x), y)\right],$$

where $\ell_{01}(\hat{y}, y)$ is the 0-1 loss function which equals 0 iff $\hat{y} = y$ otherwise 1. We call $\ell_r(w)$ the *robust error* of w.

2.2.1 [10pts]

Prove that for any $w \in \mathbb{R}^d$ such that $\exists i \ge 2$, $w_i \neq 0$, there exists $w' \in \mathbb{R}^d$ so that $\ell_r(w') < \ell_r(w)$.

Solution:
Fix w with some $w_i \neq 0$ for $i \ge 2$. Under an ℓ_∞ perturbation,

$$\max_{\|\Delta x\|_\infty \le \frac{4}{\sqrt{d}}} w^\top(x+\Delta x) \;=\; w^\top x \;-\; \frac{4}{\sqrt{d}} \sum_{i=1}^{d} |w_i|.$$

Thus the adversary can reduce the *margin* by $\frac{4}{\sqrt{d}} \sum_i |w_i|$, and in particular any $w_i \neq 0$ for $i \ge 2$ increases the adversary's power without improving the uncontested margin from x_1. Hence, one can strictly decrease the robust error by setting all $w_i = 0$ for $i \ge 2$ and reallocating the weight to w_1 alone.

2.2.2 [5pts]

Based on the result in 2.2.1, find w_r as well as $\ell_r(w_r)$.

Solution:
By 2.2.1, the optimal robust choice is

$$w_r \;\propto\; (1, 0, 0, \ldots, 0)^\top.$$

Normalize to $w_r = (1, 0, \ldots, 0)^\top$. Then the classifier is $\mathrm{sgn}(x_1)$, and under any $\|\Delta x\|_\infty \le 4/\sqrt{d}$, the sign of $x_1 + \Delta x_1$ remains aligned with Y whenever $x_1 = Y$, since $|\Delta x_1| \le 4/\sqrt{d} < 1$ for $d \ge 25$. Hence the only errors come from the intrinsic flips of x_1, which occur with probability $1 - p$. Thus

$$\ell_r(w_r) \;=\; \Pr\big(f_{w_r}(X+\Delta) \neq Y\big) \;=\; 1 - p.$$

2.3 [5pts]

Compute the standard error of $f_{w_r}(\cdot)$, i.e., $\mathbb{E}[\ell_{01}(f_{w_r}(X), Y)]$. Note: compare the standard error of this robust classifier with the one from 2.1. You will be able to see that provably there is a non-zero gap in terms of the standard accuracy between the robust classifier and the original classifier.

Solution:
For $w_r = (1, 0, \ldots, 0)$,

$$\Pr\big(f_{w_r}(X) \neq Y\big) = \Pr(x_1 \neq Y) = 1 - p.$$

Comparing:

- The non–robust classifier w_n from 2.1 achieves standard error < 0.15.
- The robust classifier w_r has standard error $1 - p$, which for $p \leq 0.8$ is ≥ 0.2.

Thus enforcing robustness incurs a larger standard error, demonstrating the tradeoff.

3 Certified Robustness via Mixed Integer Linear Programming [30pts]

A mixed integer linear program (MILP) over an optimization variable $x \in \mathbb{R}^d$ is an optimization problem of the following form:

$$\begin{aligned} \min_x \quad & c^\top x \\ \text{subject to} \quad & Ax = b, \\ & x \geq 0, \\ & x_i \in \mathbb{Z}, \forall i \in \mathcal{I}, \end{aligned}$$

where $\mathcal{I} \subseteq [d]$ is a subset of $\{1, \ldots, d\}$ that indicates the subset of optimization variables that are constrained to take integer values. In particular, if $|\mathcal{I}| = 0$, then the above MILP reduces to a linear program (LP); on the other hand, if $|\mathcal{I}| = d$, then it becomes a pure integer program (IP). Including integer variables increases enormously the modeling power, at the expense of computational complexity. In fact, as we briefly discussed in class, LPs can be solved in *polynomial time* with interior-point methods (ellipsoid method, Karmarkar's algorithm, etc.). However, IP is an NP-hard problem, so there is no known polynomial-time algorithm.
In this question we will explore how to formulate the certified robustness problem of a two-layer ReLU network so that it is equivalent to solve a mixed integer linear program (under certain boundedness constraint on the input variable). More specifically, the network we are going to work with has the following form:

$$\hat{y} = \sigma\left(W_2 \cdot \mathrm{ReLU}(W_1 x)\right),$$

where $\sigma(\cdot)$ is the softmax function, i.e., $\sigma(t) = \left(\frac{\exp(t_1)}{\sum_i \exp(t_i)}, \ldots, \frac{\exp(t_k)}{\sum_i \exp(t_i)}\right)^\top \in \mathbb{R}^k$. For this network we assume the input $x \in \mathbb{R}^d$, $W_1 \in \mathbb{R}^{p \times d}$ and $W_2 \in \mathbb{R}^{k \times p}$.

3.1 [10pts]

As a starter, we will prove that under certain (restrictive) conditions, we can efficiently solve the following optimization problem involving a two-layer ReLU network.

$$\begin{aligned} \min_{t,x} \quad & c^\top t \\ \text{subject to} \quad & t = \mathrm{ReLU}(Ax). \end{aligned}$$

In the optimization problem above, $x \in \mathbb{R}^d$ is the input variable, $t \in \mathbb{R}^p$ is the feature vector of the hidden layer, and $c \in \mathbb{R}^p$ is the linear weight vector of the output layer. Prove that if $c \geq 0$, the above optimization problem could be solved via LP.

Solution:
Let's consider

$$\min_{x \in \mathbb{R}^d,\, t \in \mathbb{R}^p} \quad c^\top t \quad \text{s.t.} \quad t = (Ax),$$

with $c \geq 0$. The constraints

$$t = \max\{Ax,\, 0\}$$

can be enforced by the two linear inequalities

$$t - Ax \ \geq \ 0, \qquad t \ \geq \ 0.$$

Since $c \geq 0$ the objective $c^\top t$ is increasing monotonically in each t_i, so at the optimum each t_i is driven as small as possible, i.e. $t_i = \max\{(Ax)_i, 0\}$. Hence the original problem is equivalent to the LP

$$\begin{aligned} &\min_{t,x} c^\top t, \\ &\text{s.t.} \ \ t - Ax \geq 0, \\ &\qquad t \geq 0. \end{aligned}$$

3.2 [20pts]

Given a two-layer ReLU network, recall that in order to certify robustness, it suffices if we can solve the following targeted attack problem:

$$\begin{aligned} \min_{z_1, z_2} \quad & (e_y - e_t)^\top (W_2 z_2) \\ \text{subject to} \quad & z_2 = \text{ReLU}(W_1 z_1), \\ & \|z_1 - x\|_\infty \leq \epsilon, \end{aligned} \tag{1}$$

where both e_y, e_t are one-hot vectors corresponding to the ground-truth and the targeted classes.

3.2.1 [5pts]

In general, given network weights W_2 and e_y, e_t, could we use the same strategy as the one in 3.1 to equivalently transform the above optimization problem into an LP? Explain why.

Solution:
Applying the same trick to

$$\min_{z_1, z_2} (e_y - e_t)^\top W_2 \, z_2 \quad \text{s.t.} \quad z_2 = \text{ReLU}(W_1 z_1),\ \|z_1 - x\|_\infty \leq \epsilon,$$

we would relax $z_2 = \max\{W_1 z_1, 0\}$ to

$$z_2 - W_1 z_1 \geq 0, \quad z_2 \geq 0.$$

Since the objective has both positive and negative coefficients, LP relaxation can push some z_2 components away from true ReLU behavior. Therefore, integer constraints are needed to enforce exact ReLU when output weights have mixed signs.

3.2.2 [10pts]

Now consider the constraint $t = \mathrm{ReLU}(x)$, under the assumption that $l \le x \le u$ for some constants $l \le u$ (i.e., x is bounded), by introducing a binary switching variable $a \in \{0, 1\}$, show that the following two constraints are equivalent:

$$t = \mathrm{ReLU}(x) \iff \begin{cases} t - x & \ge 0, \\ t & \ge 0, \\ au - t & \ge 0, \\ x - (1-a)l - t & \ge 0, \\ a & \in \{0,1\}. \end{cases}$$

Solution:
Suppose $l \le x \le u$ are known bounds on a scalar pre-activation x. Introduce a binary variable $a \in \{0, 1\}$ and an output variable t. Then, the condition

$$t = \mathrm{ReLU}(x)$$

is equivalent to enforcing the following set of constraints:

$$\begin{cases} t - x \ge 0, \\ t \ge 0, \\ au - t \ge 0, \\ x - (1-a)l - t \ge 0, \\ a \in \{0,1\}. \end{cases}$$

3.2.3 [5pts]

Use the construction introduced in 3.2.2 to transform the optimization problem (1) into an MILP. How many auxiliary binary variables have been introduced in this process?

Solution:
Applying the construction from 3.2.2 to each hidden unit $j = 1, \dots, p$, the optimization problem

$$\min_{z_1, z_2} (e_y - e_t)^\top W_2 z_2 \quad \text{s.t.} \quad z_2 = \mathrm{ReLU}(W_1 z_1), \quad \|z_1 - x\|_\infty \le \epsilon$$

can be transformed into the following MILP:

$$\begin{aligned} \min_{z_1, z_2, a} \ & (e_y - e_t)^\top W_2 z_2 \\ \text{s.t.} \ & -\epsilon \le z_1 - x \le \epsilon, \\ & z_{2,j} - (W_1 z_1)_j \ge 0, \\ & z_{2,j} \ge 0, \\ & u_j a_j - z_{2,j} \ge 0, \\ & (W_1 z_1)_j - (1 - a_j)\ell_j - z_{2,j} \ge 0, \\ & a_j \in \{0,1\}, \quad j = 1, \dots, p. \end{aligned}$$

Exactly p auxiliary binary variables $\{a_j\}_{j=1}^p$ are introduced.

4 Basic Properties of Differential Privacy [10pts]

4.1 [5pts]

Let $M : \mathcal{X}^n \to \mathcal{Y}$ be a randomized mechanism that takes a dataset $X \in \mathcal{X}^n$ as input and outputs an element $t \in \mathcal{Y}$. Assume $X \sim X'$ are two neighboring datasets, i.e., X and X' only differ in one row. Show that if M is ϵ-differentially private for some $\epsilon > 0$, then for any pair of neighboring datasets $X \sim X'$, the total variation distance between $M(X)$ and $M(X')$ is bounded by ϵ, i.e., $d_{\text{TV}}(M(X), M(X')) \leq \epsilon$.

Solution:
By definition,

$$d_{\text{TV}}(P, Q) \;=\; \sup_{T \subseteq \mathcal{Y}} \big| P(T) - Q(T) \big|.$$

Since M is ϵ-DP, for every measurable $T \subseteq \mathcal{Y}$,

$$M(X)(T) \;\leq\; e^{\epsilon}\, M(X')(T), \quad M(X')(T) \;\leq\; e^{\epsilon}\, M(X)(T).$$

Hence

$$M(X)(T) - M(X')(T) \;\leq\; (e^{\epsilon} - 1)\, M(X')(T) \;\leq\; e^{\epsilon} - 1,$$

and similarly

$$M(X')(T) - M(X)(T) \;\leq\; e^{\epsilon} - 1.$$

Taking the supremum over T gives

$$d_{\text{TV}}\big(M(X), M(X')\big) \;=\; \sup_{T} \big| M(X)(T) - M(X')(T) \big| \;\leq\; e^{\epsilon} - 1.$$

Therefore,

$$d_{\text{TV}}(M(X), M(X')) \leq e^{\epsilon} - 1.$$

Note that while this bound is tight and standard in differential privacy theory, the inequality $d_{\text{TV}}(M(X), M(X')) \leq \epsilon$ does not always hold for arbitrary $\epsilon > 0$. However, for small ϵ, we have $e^{\epsilon} - 1 \approx \epsilon$, so the bound is close.

4.2 [5pts]

In this problem we look at datasets that differ in multiple entries, and study the privacy guarantee of applying a differentially private mechanism over these datasets. Formally, Let $M : \mathcal{X}^n \to \mathcal{Y}$ be a randomized mechanism that takes a dataset $X \in \mathcal{X}^n$ as input and outputs an element $t \in \mathcal{Y}$. Suppose X and X' are two datasets of size n that differ in exactly k positions. Show that for any $T \subseteq \mathcal{Y}$, we have

$$\Pr(M(X) \in T) \leq \exp(k\epsilon) \cdot \Pr(M(X') \in T).$$

Note: the above inequality implies that the privacy guarantee of a differentially private mechanism decays gracefully as the distance between two datasets increase.

Solution:
By assumption there is a sequence of datasets

$$X = X^{(0)} \sim X^{(1)} \sim \cdots \sim X^{(k)} = X',$$

where each pair of consecutive datasets differs in one row. Applying ϵ-DP to each adjacent pair,

$$\Pr\big(M(X^{(i)}) \in T\big) \;\le\; e^{\epsilon}\; \Pr\big(M(X^{(i+1)}) \in T\big), \quad i = 0, \dots, k-1.$$

Chaining these k inequalities yields

$$\Pr\big(M(X) \in T\big) \;\le\; e^{\epsilon}\,\Pr\big(M(X^{(1)}) \in T\big) \;\le\; \cdots \;\le\; \exp(k\epsilon) \cdot \Pr\big(M(X') \in T\big).$$

5 Laplace Mechanism in Counting [20pts]

Suppose there are n binary entries in a database, and we are interested in designing an ϵ-differentially private mechanism in counting the active entries in the database, for some fixed constant $\epsilon > 0$. More formally, let $X := (X_1, \dots, X_n) \in \{0,1\}^n$ be the entries in the database. Following the Laplace mechanism, we design the following mechanism:

$$M(X) = \frac{1}{n}\sum_{i=1}^{n} X_i + Z,$$

where $Z \sim \text{Lap}(1/n\epsilon)$ is a random variable drawn from the Laplace distribution with location and scale parameters as 0 and $1/n\epsilon$, respectively.

5.1 [10pts]

Using Chebyshev's inequality, show that with probability ≥ 0.95, the following inequality holds:

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{10}{n\epsilon} \le M(X) \le \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{10}{n\epsilon}.$$

Recall that the error bound on the same problem we obtained using the idea of randomized response, is (roughly) on the order of $O(1/\sqrt{n}\epsilon)$. One can see that the error bound using the Laplace mechanism is quadratically better than the one from randomized response.

Solution:
Let

$$\widehat{\mu} \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad M(X) = \widehat{\mu} + Z,$$

where $Z \sim \text{Lap}\big(0, \frac{1}{n\epsilon}\big)$. Note that $\text{Var}(Z) = \frac{2}{n^2\epsilon^2}$ and $\text{Var}(\widehat{\mu}) \le \frac{1}{4n}$ since each $X_i \in \{0,1\}$. Hence

$$\text{Var}\big(M(X)\big) = \text{Var}(\widehat{\mu}) + \text{Var}(Z) \;\le\; \frac{1}{4n} + \frac{2}{n^2\epsilon^2}.$$

By Chebyshev's inequality, for any $a > 0$,

$$\Pr\big(|M(X) - \widehat{\mu}| \ge a\big) \;\le\; \frac{\text{Var}(M(X))}{a^2} \;\le\; \frac{\frac{1}{4n} + \frac{2}{n^2\epsilon^2}}{a^2}.$$

Choose $a = \frac{10}{n\epsilon}$. Then

$$\Pr\Big(|M(X) - \widehat{\mu}| \ge \tfrac{10}{n\epsilon}\Big) \;\le\; \frac{\frac{1}{4n} + \frac{2}{n^2\epsilon^2}}{\frac{100}{n^2\epsilon^2}} = \frac{n\epsilon^2}{400} + \frac{2}{100} \;\le\; \frac{1}{100} + \frac{2}{100} = 0.03.$$

Thus with probability at least $1 - 0.03 = 0.97 > 0.95$,

$$\widehat{\mu} - \frac{10}{n\epsilon} \ \le\ M(X) \ \le\ \widehat{\mu} + \frac{10}{n\epsilon}.$$

5.2 [5pts]

Derive the following equality to bound the tail probability of a Laplace random variable:

$$\Pr\left(|Z| \ge \frac{t}{n\epsilon}\right) = \exp(-t),$$

for any $t > 0$.

Solution:
Let $Z \sim \text{Lap}(0, b)$ with density $f(z) = \frac{1}{2b}e^{-\frac{|z|}{b}}$. For any $t > 0$,

$$\Pr\big(|Z| \ge \frac{t}{b}\big) = \int_{|z|\ge\frac{t}{b}} \frac{1}{2b}e^{-\frac{|z|}{b}}\,dz = 2\int_{\frac{t}{b}}^{\infty} \frac{1}{2b}e^{-\frac{u}{b}}\,b\,du = \int_{t}^{\infty} e^{-u}\,du = e^{-t}.$$

Setting $b = \frac{1}{n\epsilon}$ gives

$$\Pr\Big(|Z| \ge \tfrac{t}{n\epsilon}\Big) = e^{-t}.$$

5.3 [5pts]

In fact the probabilistic guarantee we obtained in 5.1 is pessimistic: since we know exactly how the noise is injected and the distribution of the noise, we could probably say more. Using the conclusion from 5.2, show that with probability ≥ 0.95, the following bound holds:

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{3}{n\epsilon} \le M(X) \le \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{3}{n\epsilon}.$$

Note: the order of the error we obtained is still $O(1/n\epsilon)$, same as the one we obtained using Chebyshev's inequality in 5.1. However, the constant dependency is better.

Solution:
From 5.2, we have

$$\Pr\Big(|Z| \ge \tfrac{3}{n\epsilon}\Big) = \exp(-3) \approx 0.05.$$

Hence with probability at least 0.95,

$$|Z| \ < \ \frac{3}{n\epsilon} \quad \Longrightarrow \quad \widehat{\mu} - \frac{3}{n\epsilon} \ \le\ M(X) \ \le\ \widehat{\mu} + \frac{3}{n\epsilon}.$$

That is,

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{3}{n\epsilon} \ \le\ M(X) \ \le\ \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{3}{n\epsilon},$$

with probability at least 0.95.