DS 102 Homework 6

If you are handwriting your solution please make sure your answers are legible, as you may lose points otherwise.
Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you do discuss the homework with others, please include their names on your submission.
Due on Gradescope by 9:29am, Thursday April 30, 2020

1. (30 points) **Neural Networks with ReLU Activations**

   In this problem, we will see how a neural net with just one hidden layer and ReLU activations can express a piece-wise linear decision boundary, which can provide a good decision boundary for datasets that are not linearly separable.

   To be more precise, we consider two-dimensional data points $\mathbf{x} \in \mathbb{R}^2$ that have labels $y \in \{-1, 1\}$. Given a training set of such points, we want to find a model of the form

   $$\hat{y}_{\vec{w},b,\alpha}(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j \max\{\langle \vec{w}_j, \mathbf{x} \rangle + b_j, 0\} \tag{1}$$

   that fits the data. For each hidden unit $j = 1, \ldots, k$, where k is the total number of hidden units, the coefficients α_j and b_j are real-valued scalars, and $\vec{w}_j$ is a two-dimensional vector of coefficients. The subscripts of $\hat{y}_{\vec{w},b,\alpha}$ are a notational reminder that the prediction depends on the coefficients of the neural network.

   (a) (5 points) Consider a dataset in $\mathbb{R}^2$ consisting of all 9 possible data points (a, b) with integer values $a, b \in \{0, 1, 2\}$ (*i.e.*, $(0,0), (0,1), (0,2), (1,0), \ldots, (2,2)$). A data point (a, b) is labeled $+1$ if $\max\{a, b\} \geq 2$, and is labeled -1 otherwise.

   The **decision boundary** of a classifier is the boundary separating the region where data points are classified as $+1$, from the region where data points are classified as -1. The **margin** of a classifier is the distance from its decision boundary to the nearest data point, and a **maximum-margin** classifier is one whose margin is maximized (among all classifiers).

   (i) With a as the x-axis and b as the y-axis, 1) plot this data set, labeling each data point with its label, and 2) draw the decision boundary of the maximum-margin piece-wise linear classifier (*i.e.*, the piece-wise linear classifier whose margin is maximized, among all classifiers with a piece-wise linear decision boundary).

   (ii) Explain why the decision boundary you drew maximizes the margin (you don't need to formally prove this).

   (b) (10 points) (i) Find a value of k and values of coefficients $\vec{w}_1, \ldots, \vec{w}_k, b_1, \ldots, b_k, \alpha_1, \ldots \alpha_k$ such that the ReLU NN defined in Equation (1) achieves zero loss, in terms of the **hinge loss** function

   $$\frac{1}{n} \sum_{i=1}^{n} \max\{1 - y_i \cdot \hat{y}_{w,b,\alpha}(\mathbf{x}_i), 0\}. \tag{2}$$

(ii) On your plot from Part (a), plot the decision boundary of the ReLU NN with the coefficient values you chose. (The decision boundary of a ReLU NN is the set $\{\mathbf{x} \mid \hat{y}_{\mathbf{w},b,\alpha}(\mathbf{x}) = 0\}$).

(c) (10 points) Suppose we want to use stochastic gradient descent to find the coefficients $\vec{w}_j$, b_j and α_j that minimize the regularized hinge-loss loss F:

$$F(\vec{w}_1, \ldots, \vec{w}_k, b_1, \ldots, b_k, \alpha_1, \ldots \alpha_k) = \frac{1}{n} \sum_{i=1}^{n} \left[(1 - y_i \cdot \hat{y}_{\vec{w},b,\alpha}(\mathbf{x}_i))_+ + \frac{\lambda}{2} \sum_{j=1}^{k} (\|\vec{w}_j\|_2^2 + \alpha_j^2) \right], \tag{3}$$

where λ gives the amount of regularization on $\vec{w}_j$ and α_j, and we use the notation $(\gamma)_+ = \max\{\gamma, 0\}$ for any real-valued γ.

For each summand in F, denote

$$f_i(\vec{w}, b, \alpha) = (1 - y_i \cdot \hat{y}_{\vec{w},b,\alpha}(\mathbf{x}_i))_+ + \frac{\lambda}{2} \sum_{j=1}^{k} (\|\vec{w}_j\|_2^2 + \alpha_j^2). \tag{4}$$

Given one data point $(\mathbf{x}_i, y_i)$, compute the gradients $\frac{\partial f_i}{\partial \alpha_j}$, $\frac{\partial f_i}{\partial b_j}$ and $\nabla_{\vec{w}_j} f_i$.

Hint: to compute gradients through the $(\gamma)_+$ operator, consider different cases.

(d) (5 points) Given an arbitrary dataset in $\mathbb{R}^2$, explain why a ReLU NN can perfectly fit the data if the number of hidden units k is large enough. Hint: think about how each hidden unit contributes to the decision boundary of a ReLU NN, like the one you plotted in Part (b.ii).

2. (30 points) **Differential Privacy via Randomized Responses**

When evaluating the frequency of illegal or embarrassing activities, the **randomized response** approach is a simple yet effective way of achieving differential privacy. In this problem, you'll explore and show some useful properties of randomized responses.

Suppose we're interested in evaluating how effective shelter-in-place measures are. We conduct a survey to estimate the fraction of the population that participates in large social gatherings during shelter-in-place. The survey participant is asked, "In the past week, have you participated in an in-person social gathering of more than 10 people?" To achieve randomized responses, the participant is then instructed to respond as follows:

1. Flip a fair coin (*i.e.*, 50% chance of heads and 50% chance of tails).
2. If tails, then respond truthfully.
3. If heads, then flip the coin again. Respond "Yes" if heads, and "No" if tails.

(a) (5 points) For a given participant, suppose $b \in \{0, 1\}$ is the value of the true answer to the question, where $b = 1$ is "Yes" and $b = 0$ is "No". As the surveyors, we don't observe b. Instead, due to the randomized response we observe $b' \in \{0, 1\}$, where

$b' = 1$ means the participant responded "Yes" and $b' = 0$ means the participant responded "No" (note that both $b \in \{0, 1\}$ and $b' \in \{0, 1\}$).

i) Suppose that for a given participant, $b = 1$. Write down the distribution that their survey response b' is drawn from. Explain your reasoning.

ii) Alternatively, suppose $b = 0$. Write down the distribution that their survey response b' is drawn from. Explain your reasoning.

iii) Write down the distribution that the survey response b' is drawn from. Your solution should involve the true answer b, such that when you plug in $b = 1$ or $b = 0$, you get the same answers as Parts (i) and (ii).

(b) (5 points) Randomized responses encourage participants to respond truthfully, because if they respond "Yes" to the incriminating question, there's always plausible deniability (*i.e.*, even if they responded "Yes", the true answer could be "No").

(i) To formalize this idea, let $q \in [0, 1]$ denote the true fraction of the population that has participated in large social gatherings in the last week. Compute the probability $\mathbb{P}(b = 1 \mid b' = 1)$ that the true answer is "Yes", given that the survey response is "Yes". Your solution should be in terms of q and constants.

(ii) For what values of q is the probability $\mathbb{P}(b = 1 \mid b' = 1)$ less than a half? (For these values of q, if a survey participant responds "Yes", they can safely claim that the probability that the true answer is "Yes" is actually less than a half. That is, given that they responded "Yes", it's actually more likely that they're law-abiding and the true answer is "No".)

(c) (10 points) Despite the fact that each participant's response may or may not be truthful, the neat thing is that we can still use the responses in aggregate to estimate q.

(i) Let

$$b_i \sim \text{Bernoulli}(q)$$

for $i = 1, \ldots, n$ denote the true answers of n survey participants we randomly sampled from the population. Let $b'_i, i = 1, \ldots, n$ denote the participants' survey responses, which are also random (based on b_i, following the distribution you derived in Part (a.iii)). Suppose we compute the sample mean of the n survey responses, $\frac{1}{n}\sum_{i=1}^{n} b'_i$. Show that

$$\mathbb{E}_{\{b_i, b'_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^{n} b'_i\right] = 1/4 + (1/2)q, \tag{5}$$

where the expectation is over both the b_i and the b'_i.

ii) Suppose you take the sample mean of the survey responses and get a value $A = \frac{1}{n}\sum_{i=1}^{n} b'_i$. Using the fact shown in Part (i), propose an estimator for q, the true fraction of the population that participated in large social gatherings. Your solution should be in terms of A and constants.

(d) (10 points) Performing the survey with randomized responses, then computing the sample mean, can be thought of as a function $\mathcal{A} : \{0, 1\}^n \to \mathbb{R}$ that we call the

"randomized response algorithm". The function $\mathcal{A}$ takes as input a dataset of true answers $\mathcal{D} = \{b_1, \ldots, b_n\}$ and outputs a random value (the sample mean $\frac{1}{n}\sum_{i=1}^{n} b'_i$, where the b'_i have the distribution you derived in Part (a.iii)).

The output $\mathcal{A}(\mathcal{D})$ can take on any value in the set $\mathcal{R} = \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. Reviewing the definition of ϵ-differential privacy, the randomized response algorithm $\mathcal{A}$ is called ϵ-**differentially private** if

$$\mathbb{P}(\mathcal{A}(\mathcal{D}_1) \in S) \leq \exp(\epsilon)\mathbb{P}(\mathcal{A}(\mathcal{D}_1) \in S) \tag{6}$$

for any set $S \subseteq \mathcal{R}$, and any two datasets $\mathcal{D}_1, \mathcal{D}_2$ of size n that differ by only one datapoint (*i.e.*, they contain all the same true answers b_i except for one, which has the value 0 in one dataset but 1 in the other). Note that the probability is over the randomness in the algorithm (*i.e.*, the randomness in the survey responses given fixed $\mathcal{D}$), not randomness in $\mathcal{D}$ which we consider fixed.

For a dataset of size $n = 1$ (a single survey participant), show that the randomized response algorithm is $(\log 3)$-differentially private.

Hint: For $n = 1$, $\mathcal{R} = \{0, 1\}$, which has three possible subsets.