HW 6

Lynn Check

1/21/2024

1

What is the difference between gradient descent and *stochastic* gradient descent as discussed in class? (*You need not give full details of each algorithm. Instead you can describe what each does and provide the update step for each. Make sure that in providing the update step for each algorithm you emphasize what is different and why.*)

The key difference between gradient descent and stochastic gradient descent lies in how the gradient is computed and applied to update the parameters of the model. Furthermore, the gradient ($\nabla f(\theta, X, Y)$) for gradient descent is computed using the entire dataset, making it computationally expansive for large datasets. Stochastic gradient descent, on the other hand, computes the gradient ($\nabla f(\theta, X_i, Y_i)$) using a single data point or a subset of the dataset, making it computationally smaller and faster. This difference is notated within the parentheses of the gradient where (X, Y) represents all the data being references while (X_i, Y_i) represents a subset of the radomly selected data.

More specifically, gradient descent computes the gradient of the loss function with respect with all of the training data points before updating the model parameters. The update step for gradient descent is $\theta_{i+1} = \theta_i - \alpha\nabla f(\theta_i, X, Y)$.

Stochastic gradient descent computes the gradient of the loss function with respect to a single randomly selected data point and immediately updates the parameters. The update step for stochastic gradient descent is $\theta_{i+1} = \theta_i - \alpha\nabla f(\theta_i, X_i, Y_i)$.

2

Consider the `FedAve` algorithm. In its most compact form we said the update step is $\omega_{t+1} = \omega_t - \eta\sum_{k=1}^{K}\frac{n_k}{n}\nabla F_k(\omega_t)$. However, we also emphasized a more intuitive, yet equivalent, formulation given by $\omega_{t+1}^k = \omega_t - \eta\nabla F_k(\omega_t); w_{t+1} = \sum_{k=1}^{K}\frac{n_k}{n}w_{t+1}^k$.

Prove that these two formulations are equivalent.
(*Hint: show that if you place ω_{t+1}^k from the first equation (of the second formulation) into the second equation (of the second formulation), this second formulation will reduce to exactly the first formulation.*)

1. First, substitute the local update $\omega_{t+1}^k = \omega_t - \eta\nabla F_k(\omega_t)$ into $w_{t+1} = \sum_{k=1}^{K}\frac{n_k}{n}w_{t+1}^k$:

$$w_{t+1} = \sum_{k=1}^{K}\frac{n_k}{n}(\omega_t - \eta\nabla F_k(\omega_t))$$

2. Then, distribute the $\sum_{k=1}^{K}\frac{n_k}{n}$:

$$w_{t+1} = \sum_{k=1}^{K}\frac{n_k}{n}\omega_i - \sum_{k=1}^{K}\frac{n_k}{n}\eta\nabla F_k(\omega_t)$$

3. Factorize the ω_i from the first term and the η from the second term:

$$w_{t+1} = \omega_i \sum_{k=1}^{K} \frac{n_k}{n} - \eta \sum_{k=1}^{K} \frac{n_k}{n} \nabla F_k(\omega_t)$$

4. Since $\sum_{k=1}^{K} \frac{n_k}{n} = 1$, the first term will simplify to:

$$w_{t+1} = \omega_i - \eta \sum_{k=1}^{K} \frac{n_k}{n} \nabla F_k(\omega_t)$$

Therefore, proving the two formulations to be equivalent to one another.

3

Now give a brief explanation as to why the second formulation is more intuitive. That is, you should be able to explain broadly what this update is doing.

The second formulation of the Federated Average (`FedAve`) algorithm is more intuitive because it explicitly separates the update step into easier and clearer steps: local updates and global aggregation. Each client independently updates its local model by applying stochiastic gradient descent to its data in the first step by using the rule $\omega_{t+1}^k = \omega_t - \eta \nabla F_k(\omega_t)$. This step reflects what a client would do if it were training its model independently and working with only its local data. The global step involves aggregating the updated local models to create a global model, which is weighed by the proportion of data contributed by each client to ensure that clients with a larger datasets have a proportionally greater influence on the global model, mirroring a decentralized system. By separating the local and global steps, the second formulation is not only intuitive, but also emphasizes the preservation of data, scalability, and data privacy all while optimizing the global model.

4

Prove that randomized-response differential privacy is ϵ-differentially private.

To prove that randomized-response differential privacy is ϵ-differentially private, we are going to demonstrate this by verifying the definition of ϵ-differential privacy.

For the randomized response mechanism:

- If the true answer is "Yes" ($x = \text{Yes}$):

$$P(\text{Output} = \text{Yes} \mid x = \text{Yes}) = \frac{3}{4}, \quad P(\text{Output} = \text{No} \mid x = \text{Yes}) = \frac{1}{4}.$$

- If the true answer is "No" ($x = \text{No}$):

$$P(\text{Output} = \text{Yes} \mid x = \text{No}) = \frac{1}{4}, \quad P(\text{Output} = \text{No} \mid x = \text{No}) = \frac{3}{4}.$$

The definition of ϵ-Differential Privacy is a mechanism that satisfies ϵ-differential privacy if, for any output $y \in \{\text{Yes}, \text{No}\}$ and for any pair of inputs $x, x' \in \{\text{Yes}, \text{No}\}$:

$$\frac{P(\text{Output} = y \mid x)}{P(\text{Output} = y \mid x')} \leq e^{\epsilon}$$

We are going to verify this condition for both outputs where Case 1 is where Output = "Yes" and Case 2 is where Output = "No"

Case 1: Output = "Yes" For y = Yes:

$$\frac{P(\text{Output} = \text{Yes} \mid x = \text{Yes})}{P(\text{Output} = \text{Yes} \mid x = \text{No})} = \frac{\frac{3}{4}}{\frac{1}{4}} = 3.$$

Thus, the ratio is $3 \leq e^{\epsilon}$. For this to hold, we require:

$$e^{\epsilon} \geq 3.$$

Case 2: Output = "No" For y = No:

$$\frac{P(\text{Output} = \text{No} \mid x = \text{Yes})}{P(\text{Output} = \text{No} \mid x = \text{No})} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

Thus, the ratio is $\frac{1}{3} \leq e^{\epsilon}$. For this to hold, we require:

$$e^{-\epsilon} \leq \frac{1}{3}.$$

In conclusion, after combining the results from both cases, we have $\frac{1}{3} \leq e^{\epsilon} \leq 3$, demonstrating that the randomized response mechanism satisfies ϵ-differential privacy for ϵ such that $e^{\epsilon} = 3$. The proof shows that the mechanism is bounded within the required constraints for ϵ-differential privacy.

5

Define the harm principle. Then, discuss whether the harm principle is *currently* applicable to machine learning models. (*Hint: recall our discussions in the moral philosophy primer as to what grounds agency. You should in effect be arguing whether ML models have achieved agency enough to limit the autonomy of the users of said algorithms.*)

The harm principle is defined as when personal autonomy can be justifiably restricted if the exercise of it causes harm to others or moral agents as defined by John Stuart Mill. What this means is that people are free to do as they please until their actions cause harm to other people. Moral agency involves independent reasoning, intentionality, and accountability which are qualities that machine learning models currently lack. Therefore, the harm principle is currently not applicable to machine learning algorithms because the models operate based on predefined objectives, patterns, and data provided by human developers, rather than exercising autonomy in a way that could be morally evaluative. The developers of the algorithm have full control and free will to determine and shape the model to work the way that they want the algorithm to work. In other words, it is the developers' moral responsibility to ensure the algorithm does not produce bias predictions that can put others in harm's way. The responsibility would rest on the developers. With that being said, machine learning models themselves are not moral agents, but their outputs and application can cause harm, such as perpetuating bias, violating privacy, or spreading misinformation. Connecting to earlier points, these harms stem from human decisions in the design, deployment, and usage of machine learning systems and not from the models' own autonomy. Therefore, the harm principle does not directly apply to machine learning models. It can, however, serve as a guideline for the ethical responsibilities of those who develop and use them, ensuring that their actions minimize harm to others.