HW 6

Seth Hall

11/17/2024

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.*)

Gradient descent and stochastic gradient descent are two ways to train machine learning models by improving their predictions step by step. Gradient descent calculates the average of all the gradients for the entire dataset before taking a step to update the model. The update step is thetai+1 =thetai − f(thetai,X,Y) where is known as the step size or learning rate. Stochastic gradient descent on the other hand uses only one random example at a time instead of using all the data. The update step is thetai+1 =thetai − f(thetai,Xi ,Yi). Gradient descent is like gathering everyone's opinions before making a decision, so it's slower but very precise. SGD is like asking one random person, so it's faster but noisier. Over time, SGD averages out to a good solution.

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.*)

omega_t+1 = {k=1}^K (n_k/n) (omega_t − n F_k(omega_t)) –> omega_t+1 = (omega_t ({k=1}^K) (n_k/n)) - (−n {k=1}^K(n_k/n) F_k(_t)) –> omega_t+1 = omega_t−n {k=1}^K (n_k/n) F_k(omega_t)

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 is more intuitive because it has an update for each of the K clients that are averaged to give a global update. Each device or client updates its own parameters based on its local data (^k_(t+1)). The updates are then combined and averaged together (and weighted) to get global parameters. This makes sense because it shows collaboration. Everything does its part locally and then shares it with the server, which balances the contributions.

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

Randomized-response differential privacy flips the value of sensitive data with some probability to protect privacy. If the true value is x {0,1} we report y such that $^{=thetaP}$*+(1 − theta)theta and P^= (^−(1 − theta)theta)/theta. For randomized algorithm A it is −differentially private if and only if datasets D1, D2 differing in exactly one element and subsets S im(A) for For >0 (P[A(D_1) S])/(P[A(D_2) S]) e^. Randomized Response Differential Privacy is −Differentially Private where = ln(3). (P[A(Yes)=Yes])/P[A(No)=Yes] = (P[Output=Yes|Input=Yes]/(P[Output=Yes|Input=No] = (3/4)/(1/4) = 3 = e^(ln(3)). Similarly (P[A(No)=Yes])/(P[A(Yes)=Yes]) = 1/3. Thus (P[A(D1) S])/(P[A(D2) S]) 3 = e^(ln(3)). The randomized response is ln(3) − DP.*

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.*)

John Stuart Mill's harm principle says we should only restrict someone's freedom if their actions harm others. Right now, machine learning models lack agency because they don't make independent decisions or have intentions and they only follow the rules set by humans. However, ML models can still limit people's autonomy, such as by making unfair decisions like biased algorithms that make unfair parole decisions (COMPAS) or amplifying harmful behaviors.