HW 6

Sam Pell

4/10/2024

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.*)

Student Input

Gradient descent is a technique that iteratively moves a model down its gradient towards some objective function minima, generating a new model that approaches the population-level model. Stochastic gradient descent is similar, but when moving a model, instead of iterating on some fully predictable value, the descent adds a bit of random noise by selecting one observation to use at random, which helps shield data and prevent memory overload.

In gradient descent, the function must be differentiable and convex, and this holds true for stochastic gradient descent. Typically, this would be done in a multivariate setting given the focus is to provide the best model, meaning the current position is a vector, but it still applies in the single-variable case. The formula is such:

Gradient descent:

$$p_{n+1} = p_n - \eta \nabla f(p_n)$$

Stochastic gradient descent:

$$p_{n+1} = p_n - \eta \nabla f_i(p_n)$$

Note that here we are only changing one thing: selecting one random observation instead of the entire set. This gives the algorithm a more winding path as it follows the gradient.

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.)

Student Input

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^{k}$$

$$\omega_{t+1}^{k} = \omega_t - \eta \nabla F_k(\omega_t)$$

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

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

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

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

$$\sum_{k=1}^{K} \frac{n_k}{n} = 1$$

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

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.

Student Input

The second formulation is better because it gives some idea of how the model is iterating along the gradient, and gives more indication, in my opinion, as to how the individual weights are calculated. However, I honestly feel like I'm missing something here, and would love to know the full correct answer.

Explain how the harm principle places a constraint on personal autonomy. 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 states that people maintain autonomy, up to the limit that any action they take cannot harm any other person. In considering current machine learning models, there is an inherent trade-off between performance and privacy, where individual data points may be re-identified using model outputs in MIAs. While this is uncommon, it still presents the issue that in many machine learning models, there is a lack of sufficient security for individuals who provide consent, though rarely fully informed consent, to their data. This can cause harm to the individuals by exposing their personal, and sometimes sensitive, information. It would be interesting to discuss further how, when using approaches like differential privacy and data federation, people could still be harmed, but there is a clear trade-off, and there may be more benefits from a utilitarian perspective to favor some level of performance from models, especially when the model may provide some significant benefit to the individuals who provide its sample data.