Gradient Descent Mini-Quiz

TA: Raguvir Kunani **DS100 Spring 2020**

The questions are meant to increase in difficulty, with a challenge question at the end. The challenge question is at least as hard, if not harder, than a hard exam question.

1. **What is the purpose of gradient descent (i.e. what goal does it accomplish for us that is relevant to the modeling process)?**

2. **Recall the gradient descent update rule:**

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \rho \left(\nabla L(\theta)\big|_{\theta=\theta^{(t)}}\right)$$

   (a) **Could gradient descent still find a minimum if the minus sign was changed to a plus sign (no other changes)?**

   (b) **Could gradient descent still find a minimum if the ρ term was removed (no other changes)?**

   (c) **When does gradient descent stop?** *Hint:* See when $\theta^{(t+1)} = \theta^{(t)}$ in the update rule.

3. **Are there any functions for which gradient descent is not guaranteed to find the global minimum? If so, give an example and explain.**

4. **Which values of ρ are guaranteed to produce incorrect results for gradient descent?** Assume the initial guess $\theta^{(0)}$ is not the minimum. Select all that apply.

   A. 1

   B. $-\frac{1}{2}$

   C. 0

   D. 100

   E. -1

5. **We often replace the ρ term in the gradient descent update rule with a $\rho(t)$ term. This allows each iteration of gradient descent to have its own value of ρ. Give an example of one appropriate selection of $\rho(t)$ and explain why your choice is appropriate.**

6. **Challenge Question**

   Recall the gradient descent update rule:

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \rho \left(\nabla L(\theta)\big|_{\theta=\theta^{(t)}}\right)$$

   For this question, **assume that $L(\theta)$ is MSE (mean squared error) and the model is $f_\theta(x)$.**

   (a) **Rewrite the gradient descent update rule replacing the gradient with a summation of terms, one for each of n data points. The final form should look like: $\theta^{(t+1)} \leftarrow \theta^{(t)} - \rho\frac{1}{n}\sum_{i=1}^{n} g(x_i, y_i, \theta^{(t)})$, where you define the g function.**

   (b) **Gradient descent requires the computation of ______________ gradients during each iteration.**
   Hint: Use the result from the previous problem.

(c) (**Hard**) Since our datasets are large, we want to avoid computing so many gradients each iteration. Let's see if we can take advantage of our old friend, random sampling. **Show that the average gradient of the loss function evaluated at B randomly chosen data points is an unbiased estimator of the average gradient of the loss function evaluated at all n data points.** Let $l(x_i, y_i, \theta)$ denote the loss of the model $f_\theta(x)$ on a single data point (x_i, y_i) and assume $B < n$. ***Hint:*** **Z is an unbiased estimator of Y if $E[Z] = E[Y]$.**

(d) This is great news! We can compute B gradients at each iteration and still find the minimum. **Update the gradient descent update rule you wrote in part (a) to reflect the change that we are only computing B gradients at each iteration.**

(e) **For a fixed B, how does the performance of the updated gradient descent (which computes only B gradients per iteration) relate to the original gradient descent? Select all that apply.**

A. On average, finds a minimum in less iterations than original gradient descent

B. On average, finds a minimum in more iterations than original gradient descent

C. Will find the same minimum as original gradient descent

D. On average, finds a minimum in less time than original gradient descent

E. On average, finds a minimum in more time than original gradient

(f) **How does increasing B affect the performance of the updated version of gradient descent? Select all that apply.**

A. Finds a minimum in less iterations

B. Finds a minimum in less time

C. Finds a minimum in more iterations

D. Finds a minimum in more time

(g) **The updated gradient descent is also known as ______________ gradient descent.**