DSC 140A - Discussion 02

Problem 1.

You are building a predictive model for a dataset representing measurements of houses. Each data point has three features: the number of bedrooms (x_1), the size in square feet (x_2), and the age of the house in years (x_3). The target is the house price (y).

The dataset is:

x_1	x_2	x_3	y
2	1200	20	250
3	1500	15	300
5	1700	10	400
1	800	25	200

We aim to fit a linear model of the form:

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$

using stochastic gradient descent (SGD) and the following loss function for each data point:

$$L(H(\vec{x}^{(i)}; \vec{w}), y_i) = (y_i - Aug(\vec{x}^{(i)}) \cdot \vec{w})^2.$$

Perform three iterations of SGD and report the final values of w_0, w_1, w_2 and w_3. As a reminder, the steps to take are:

1. Derive the gradients of the loss function with respect to each parameter w_0, w_1, w_2 and w_3.

Solution:

Expand the loss function:

$$L = (Aug(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2 = (w_0 + x_1^{(i)} w_1 + x_2^{(i)} w_2 + x_3^{(i)} w_3 - y_i)^2$$

By the Chain Rule,

$$\begin{aligned}
\frac{d}{dw_0} L &= 2(Aug(\vec{x}^{(i)}) \cdot \vec{w} - y_i) \cdot 1 \\
\frac{d}{dw_1} L &= 2(Aug(\vec{x}^{(i)}) \cdot \vec{w} - y_i) \cdot x_1^{(i)} \\
\frac{d}{dw_2} L &= 2(Aug(\vec{x}^{(i)}) \cdot \vec{w} - y_i) \cdot x_2^{(i)} \\
\frac{d}{dw_3} L &= 2(Aug(\vec{x}^{(i)}) \cdot \vec{w} - y_i) \cdot x_3^{(i)}
\end{aligned}$$

2. Perform stochastic gradient descent for three iterations with a batch size of 2 with the following settings:
   - Initial values: $w_0 = 0$, $w_1 = 0$, $w_2 = 0$, $w_3 = 0$
   - Learning rate: $\eta = 0.01$
3. Update the parameters step-by-step using the gradients.

4. Report the final values of w_0, w_1, w_2 and w_3.

Solution:

(a) Iteration #1 - Assume $B^{(1)} = \{1, 3\}$.

From (1), the gradient of the loss is

$$\frac{d}{d\vec{w}} L_i = 2(Aug(\vec{x}^{(i)}) \cdot \vec{w} - y_i) Aug(\vec{x}^{(i)})$$

So, when $i \in B = \{1, 3\}$ and $\vec{w}^{(0)} = \vec{0}$,

$$\frac{d}{d\vec{w}} L_1 = 2(Aug(\vec{x}^{(1)}) \cdot \vec{w}^{(0)} - y_1) Aug(\vec{x}^{(1)}) = 2 \cdot (0 - 250) \cdot \begin{bmatrix} 1 \\ 2 \\ 1200 \\ 20 \end{bmatrix} = 2 \begin{bmatrix} -250 \\ -500 \\ -300000 \\ -5000 \end{bmatrix}$$

$$\frac{d}{d\vec{w}} L_3 = 2(Aug(\vec{x}^{(3)}) \cdot \vec{w}^{(0)} - y_3) Aug(\vec{x}^{(3)}) = 2 \cdot (0 - 400) \cdot \begin{bmatrix} 1 \\ 5 \\ 1700 \\ 10 \end{bmatrix} = 2 \begin{bmatrix} -400 \\ -2000 \\ -680000 \\ -4000 \end{bmatrix}$$

Calculate the stochastic gradient,

$$\vec{g} = \frac{1}{2}(\frac{d}{d\vec{w}} L_1 + \frac{d}{d\vec{w}} L_3) = \begin{bmatrix} -650 \\ -2500 \\ -980000 \\ -9000 \end{bmatrix}$$

Update,

$$\vec{w}^{(1)} = \vec{w}^{(0)} - \eta \vec{g} = \vec{0} - 0.01 \begin{bmatrix} -650 \\ -2500 \\ -980000 \\ -9000 \end{bmatrix} = \begin{bmatrix} 6.5 \\ 25 \\ 9800 \\ 90 \end{bmatrix}$$

(b) Iteration #2 - Assume $B^{(2)} = \{2, 3\}$.

When $i \in B = \{2, 3\}$ and $\vec{w}^{(1)} = \begin{bmatrix} 6.5 \\ 25 \\ 9800 \\ 90 \end{bmatrix}$,

$$\frac{d}{d\vec{w}} L_2 = 2(Aug(\vec{x}^{(2)}) \cdot \vec{w}^{(1)} - y_2) Aug(\vec{x}^{(2)}) = 2 \cdot (\begin{bmatrix} 1 \\ 3 \\ 1500 \\ 15 \end{bmatrix} \cdot \begin{bmatrix} 6.5 \\ 25 \\ 9800 \\ 90 \end{bmatrix} - 300) \cdot \begin{bmatrix} 1 \\ 3 \\ 1500 \\ 15 \end{bmatrix} = 2 \begin{bmatrix} 14701131.5 \\ 44103394.5 \\ 22051697200 \\ 220516972 \end{bmatrix}$$

$$\frac{d}{d\vec{w}} L_3 = 2(Aug(\vec{x}^{(3)}) \cdot \vec{w}^{(1)} - y_3) Aug(\vec{x}^{(3)}) = 2 \cdot (\begin{bmatrix} 1 \\ 5 \\ 1700 \\ 10 \end{bmatrix} \cdot \begin{bmatrix} 6.5 \\ 25 \\ 9800 \\ 90 \end{bmatrix} - 400) \cdot \begin{bmatrix} 1 \\ 5 \\ 1700 \\ 10 \end{bmatrix} = 2 \begin{bmatrix} 16660631.5 \\ 83303157.5 \\ 28323073600 \\ 166606315 \end{bmatrix}$$

Calculate the stochastic gradient,

$$\vec{g} = \frac{1}{2}(\frac{d}{d\vec{w}}L_2 + \frac{d}{d\vec{w}}L_3) = \begin{bmatrix} 31361763 \\ 127406552 \\ 50374770800 \\ 387123288 \end{bmatrix}$$

Update,

$$\vec{w}^{(2)} = \vec{w}^{(1)} - \eta\vec{g} = \begin{bmatrix} 6.5 \\ 25 \\ 9800 \\ 90 \end{bmatrix} - 0.01 \begin{bmatrix} 31361763 \\ 127406552 \\ 50374770800 \\ 387123288 \end{bmatrix} = \begin{bmatrix} -313611.130 \\ -1274040.52 \\ -503737908 \\ -3871142.88 \end{bmatrix}$$

(c) Iteration #3 - Assume $B^{(3)} = \{1, 4\}$.

When $i \in B = \{1, 4\}$ and $\vec{w}^{(2)} = \begin{bmatrix} -313611.130 \\ -1274040.52 \\ -503737908 \\ -3871142.88 \end{bmatrix}$,

$$\frac{d}{d\vec{w}}L_1 = 2(Aug(\vec{x}^{(1)}) \cdot \vec{w}^{(2)} - y_1)Aug(\vec{x}^{(1)}) = 2 \cdot (\begin{bmatrix} 1 \\ 2 \\ 1200 \\ 20 \end{bmatrix} \cdot \begin{bmatrix} -313611.130 \\ -1274040.52 \\ -503737908 \\ -3871142.88 \end{bmatrix} - 250) \cdot \begin{bmatrix} 1 \\ 2 \\ 1200 \\ 20 \end{bmatrix}$$

$$= 2 \begin{bmatrix} -6.04565774e + 11 \\ -1.20913155e + 12 \\ -7.25478929e + 14 \\ -1.20913155e + 13 \end{bmatrix}$$

$$\frac{d}{d\vec{w}}L_4 = 2(Aug(\vec{x}^{(4)}) \cdot \vec{w}^{(2)} - y_4)Aug(\vec{x}^{(4)}) = 2 \cdot (\begin{bmatrix} 1 \\ 1 \\ 800 \\ 25 \end{bmatrix} \cdot \begin{bmatrix} -313611.130 \\ -1274040.52 \\ -503737908 \\ -3871142.88 \end{bmatrix} - 200) \cdot \begin{bmatrix} 1 \\ 1 \\ 800 \\ 25 \end{bmatrix}$$

$$= 2 \begin{bmatrix} -4.03088693e + 11 \\ -4.03088693e + 11 \\ -3.22470954e + 14 \\ -1.00772173e + 13 \end{bmatrix}$$

Then,

$$\vec{g} = \frac{1}{2}(\frac{d}{d\vec{w}}L_1 + \frac{d}{d\vec{w}}L_4) = \begin{bmatrix} -1.00765447e + 12 \\ -1.61222024e + 12 \\ -1.04794988e + 15 \\ -2.21685328e + 13 \end{bmatrix}$$

Update,

$$\vec{w}^{(3)} = \vec{w}^{(2)} - \eta\vec{g} = \begin{bmatrix} 1.00762311e + 10 \\ 1.61209284e + 10 \\ 1.04789951e + 13 \\ 2.21681457e + 11 \end{bmatrix}$$

which are the final values of w_0, w_1, w_2 and w_3 after the three SGD iterations.

Problem 2.

Suppose you are working on a machine learning model where underestimating the target variable (y) is penalized more than overestimating it. The custom loss function is defined as:

$$L(y, H(x)) = \begin{cases} 2(y - H(x))^2, & \text{if } y > H(x), \\ (y - H(x))^2, & \text{otherwise.} \end{cases}$$

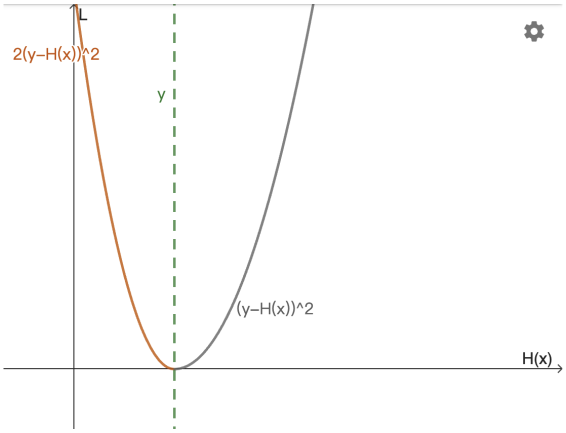
The model predicts the target variable using the equation:

$$H(x) = w_0 + w_1 x$$

where w_1 is the weight, w_0 is the bias, and x is the input feature.

a) Sketch a graph of the loss function $L(y, H(x))$ as a function of $H(x)$, for a fixed value of y.

Solution:



b) Derive the gradients of the custom loss function with respect to the parameters w_1 and w_0.

Solution:

Derive the gradients piecewise:

$$\frac{d}{d\vec{w}} L = \begin{bmatrix} \frac{d}{dw_0} L \\ \frac{d}{dw_1} L \end{bmatrix} = \begin{cases} \begin{bmatrix} 4(w_0 + w_1 x - y) \\ 4x(w_0 + w_1 x - y) \end{bmatrix}, & \text{if } y > H(x), \\ \\ \begin{bmatrix} 2(w_0 + w_1 x - y) \\ 2x(w_0 + w_1 x - y) \end{bmatrix}, & \text{if } y < H(x). \end{cases}$$

However, if $y = H(x)$ (at the point of intersection), we need to check if L is both continuous and differentiable; otherwise, no gradient exists at that point of intersection.

Let $f_1(H(x)) = 2(y - H(x))^2$ and $f_2(H(x)) = (y - H(x))^2$, then

- L is continuous if $f_1(H(x)) = f_2(H(x))$.
- L is differentiable if $f_1'(H(x)) = f_2'(H(x))$.

When $H(x) = y$, check both conditions,

- $f_1(H(x)) = f_2(H(x)) = 0$, so L is continuous.
- $f_1'(H(x)) = f_2'(H(x)) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, so L is differentiable.

So, there exists a gradient at the point of intersection $H(x) = y$, and the final answer is

$$\frac{d}{d\vec{w}} L = \begin{bmatrix} \frac{d}{dw_0} L \\ \frac{d}{dw_1} L \end{bmatrix} = \begin{cases} \begin{bmatrix} 4(w_0 + w_1 x - y) \\ 4x(w_0 + w_1 x - y) \end{bmatrix}, & \text{if } y > H(x), \\ \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, & \text{if } y = H(x), \\ \\ \begin{bmatrix} 2(w_0 + w_1 x - y) \\ 2x(w_0 + w_1 x - y) \end{bmatrix}, & \text{if } y < H(x). \end{cases}$$