Machine Learning

→ Supervised Learning

→ Unsupervised Learning

→ Reinforcement Learning

Example : GPT-5 (Generative Pre-Trained Transformer-5)

→ Predicts the next word for user input sequence (self-supervised learning via neural networks)

→ Creates a 'reward model' to align with human values (reinforcement learning)

Supervised Learning

→ Takes $\vec{x}$ and predicts y.
(blood pressure, glucose → has diabetes)

→ First developed by statisticians.
(Linear Regression, Decision Tree, Naive Bayes Logistic Regression)

→ Pioneered by a psychologist (Frank Rosenblatt).
(Perceptron)

Perceptron

x_1	x_2	y	
10	25	0	training set
15	30	1	
14	28	1	
11	14	0	
15	28	??	test instance

$$z = w_1 x_1 + w_2 x_2 + b \qquad (w_1 = 0.5,\ w = 0.2,\ b = 0.1)$$

→ randomly initialized

$$\hat{y} = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

Sample 1: $z = 0.5 \times 10 + 0.2 \times 25 + 0.1$

$= 10.1 \quad (\text{so, } \hat{y}_1 = 1)$

$$e = y_1 - \hat{y}_1 = 0 - 1 = -1$$

updating → $w_1 = w_1 + \eta e x_1$

$= 0.5 + 0.01 \times (-1) \times 10 = \boxed{0.4}$

$w_2 = w_2 + \eta e x_2$

$= 0.2 + 0.01 \times (-1) \times 25 = \boxed{-0.05}$

$b = b + \eta e = 0.1 + 0.01 \times (-1) = \boxed{0.09}$

Sample 2: $z = 0.4 \times 15 + (-0.05) \times 30 + 0.09$

$= 4.59 \quad (\text{so, } \hat{y}_2 = 1)$

$e = y_2 - \hat{y}_2 = 1 - 1 = 0$

Updating

$w_1 = w_1 + \eta e . x_1$

$= 0.4 + 0 = \boxed{0.4}$

$w_2 = w_2 + \eta e . x_2$

$= -0.05 + 0 = \boxed{-0.05}$

similarly, $b = \boxed{0.09}$

Logistic Regression

→ Perceptron with sigmoid activation function, and weight updates are made via gradient descent algorithm. Also uses binary cross-entropy loss.

Same example

Sample 1: $z = 0.5 \times 10 + 0.2 \times 25 + 0.1$

$= 10.1$

$\hat{y}_1 = \text{sigmoid}(10.1) = 0.99$

$L_1 = -y \log(\hat{y}) + (1-y) \log(1-\hat{y})$

$L_1 = -0 + (1-0) \log(1-0.99) = -2$

Sample 2: $z = 0.5 \times 15 + 0.2 \times 30 + 0.1$

$= 13.6$

$\hat{y}_2 = \text{sigmoid}(13.6) = 0.99$

$L_2 = -1 \log(0.99) + (1-1) \log(1-0.99)$

$= 0.004$

Sample 3:

$$z = 0.5 \times 14 + 0.2 \times 28 + 0.1 = 12.7$$

$$\hat{y} = \text{sigmoid}(12.7) = 0.99$$

$$L_3 = 0.004$$

Sample 4:

$$z = 0.5 \times 11 + 0.2 \times \overset{14}{\cancel{28}} + 0.1 = 8.4$$

$$\hat{y} = \text{sigmoid}(\overset{8.4}{\cancel{12.7}}) = 0.99$$

$$L_4 = -2$$

So, $L = \left(|-2| + |0.004| + |0.004| + |-2|\right)/4$

$$= 4.008 / 4 = 1.002$$

Updating

$$w = w - \alpha \times \frac{dL}{dw}$$

$$b = b - \alpha \times \frac{dL}{db}$$

$$\frac{dL}{dw_i} = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i) x_i}{n}$$

$$\frac{dL}{db} = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)}{n}$$

$$\frac{dL}{dw_1} = [(0.99-0)\times 10 + (0.99-1)\times 15 + (0.99-1)\times 14 + (0.99-0)\times 11]/4 = 5.125$$

$$\frac{dL}{dw_2} = [(0.99-0)\times 25 + (0.99-1)\times 30 + (0.99-1)\times 28 + (0.99-0)\times 14]/4 = 9.507$$

$$\frac{dL}{db} = \frac{(0.99-0)+(0.99-1)+(0.99-1)+(0.99-0)}{4}$$

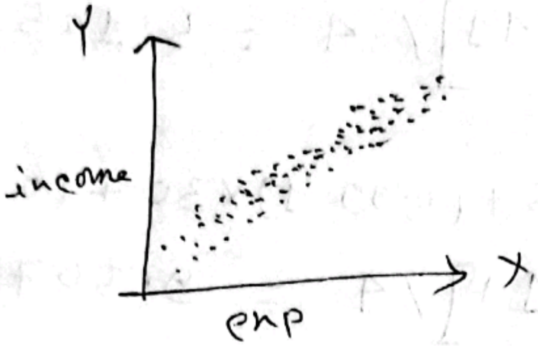
$$= 0.49$$

$$w_1 = 0.5 - (0.01 \times 5.125) = 0.44$$

$$w_2 = 0.2 - (0.01 \times 9.507) = 0.10$$

$$b = 0.1 - (0.01 \times 0.49) = 0.09$$

Linear Regression



can be represented as $y = mx + c$, or,

$$y = b_0 + b_1 x$$

This is simple linear regression. We learnt this in STA201.

<u>Example:</u>

x	y
2	4
4	6
6	8

Update weights and bias.

Solution: $w = 0.1$, $b = 0.3$ $(w = m = b_1)$
$(b = b_0 = c)$

Let, our equation: $y = mx + c$

then, $m = 0.1$, $c = 0.3$

Sample 1

SSR= $\sum_{i=1}^{n} (y_i - \hat{y})^2$

$y = m \times 2 + c = 2m + c$

$L_1 = [4 - (2m + c)]^2$

Sample 2

$y = m \times 4 + c = 4m + c$

$L_2 = [6 - (4m + c)]^2$

Sample 3

$y = m \times 6 + c = 6m + c$

$L_3 = [8 - (6m + c)]^2$

$L = [4 - (2m + c)]^2 + [6 - (4m + c)]^2 + [8 - (6m + c)]^2$

Here, L is our loss function. Our goal:

→ Find the optimal values for m and c so that this function gets its lowest value.

We call it, "minimizing the loss function."

How to minimize a function

Remember, loss functions are usually convex functions.

Fig: Convex function

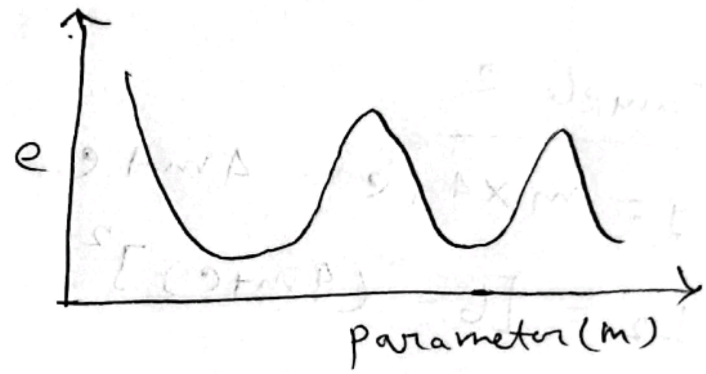


Fig: Non-convex function

* Gradient descent work well for convex functions. For non-convex, it may not find the optimal parameters.

Geometric definition of Differentiation

The derivative of a function at any given point is the slope of tangent line on the point, and the tangent line goes through the point of tangency.

⊛ Assume, our loss function; $y = f(x) = 3x^2 + 4x$

Or precisely: $e = f(m) = 3m^2 + 4m$

$$\text{for, } m = 0.1, \quad e = 3(0.1)^2 + 4(0.1)$$
$$= 0.43$$

$$\text{point of tangency } (\underset{x_0}{0.1}, \underset{y_0}{0.43})$$

Equation of tangent line : $y - y_0 = m(x - x_0)$

$$\Rightarrow y - 0.43 = m(x - 0.1)$$

$$m = f'(0.1)$$

$$f'(x) = 6x + 4$$

$$f'(0.1) = 4.6$$

$$y - 0.43 = 4.6(x - 0.1)$$

$y = 3x^2 + 4x$ (our loss func.)

$x = 0.1$

$x' = x -$ something

$= x - (\alpha \times f'(x))$

$= x - (\alpha \times \frac{dy}{dx})$

if: $y = L$, $x = m$, then,

$$\boxed{m' = m - (\alpha \times \frac{dL}{dm})}$$

$0.1 - 4.6\alpha$
$0.1 - (-4.6)\alpha$

α = learning rate.

for $\alpha = 0.01$: $m' = 0.1 - (0.01 \times 4.6)$
$= 0.054$

for $\alpha = 0.3$: $m' = 0.1 - (0.3 \times 4.6)$
$= -1.28$

$$L = [4 - (2m+c)]^2 + [6 - (4m+c)]^2 + [8 - (6m+c)]^2$$

$$\frac{dL}{dm} = 2[4 - (2m+c)] \cdot -2$$
$$+ 2[6 - (4m+c)] \cdot -4$$
$$+ 2[8 - (6m+c)] \cdot -6$$
$$= -14 - 42.4 - 85.4 = -141.8$$

$$m' = 0.1 - (0.01 \times -141.8) = \boxed{1.518}$$

$$\frac{dL}{dc} = 2[4 - (2m+c)] \cdot -1$$
$$+ 2[6 - (4m+c)] \cdot -1$$
$$+ 2[8 - (6m+c)] \cdot -1$$
$$= -7 - 10.6 - 14.2 = -31.8$$

$$c' = 0.3 - (0.01 \times -31.8) = \boxed{0.618}$$