ॐ AMMA

$$\hat{y} = w_1 x_1 + w_2 x_2 + \dots + w_n x_n + b$$

~~Given~~ $\cancel{(x_0, y_0)}$ . . . . . .

Given m datapoints $(\langle x_1^1 \dots x_n^1 \rangle, y^1)$ . . .
. . . . $(\langle x_1^m, \dots, x_n^m \rangle, y^m \rangle$

classification

→ This does not work for logistic regression

$$\hat{y} = \frac{1}{1 + e^{-(w_1 x_1 + \dots + w_n x_n + b)}}$$

$$= \sigma(w_1 x_1 + \dots + w_n x_n + b)$$

Sigmoid function

Cost function

$$\text{cost} = \frac{1}{m} \sum_{i=0}^{m} | y - \hat{y} |$$

This is suitable for Linear regression

In classification $y = 0$ or 1

↳ If we take the above formula, we may end up in local minima instead of local minima.

(Not proving this. Pls take it as a fact)

So, we use Binary Cross Entropy

$$\text{cost} = -\frac{1}{m} \left[y \log \hat{y} + (1-y) \log(1-\hat{y}) \right]$$

$$\hat{y} = \sigma(W^T x + b)$$

How is the above equation a cost function?
Let's try to answer that.

When y & $\hat{y}$ are not the same, cost should be high. Let's check that.

$$\text{Cost} = -\frac{1}{m}\sum\left[y \log \hat{y} + (1-y)\log(1-\hat{y})\right]$$

$y = 0 \Rightarrow$ & $\hat{y} \approx 0$ | 0 | $-1 \log (1-\varepsilon)$ close to 1, very small quantity

ε = epsilon

$y = 0$ & $\hat{y} \approx 1$ | 0 | $-1 \log (0+\varepsilon)$ close to zero, very high quantity

$y = 1$ & $\hat{y} \approx 0$ | $-\log(0+\varepsilon)$ close to 0, very high quantity | 0

$y = 1$, $\hat{y} \approx 1$ | $-\log(1-\varepsilon)$ close to 1, very small quantity | 0

Conclusion: The formula works well as cost function

Derivative of the cost.

$$\text{cost} = -\frac{1}{m}\sum_{i=1}^{m} y \log \hat{y} + (1-y)\log(1-\hat{y})$$

$$\hat{y} = \sigma(W^T \cancel{X} + b)$$

$$= \sigma(z) = \frac{1}{1+e^{-z}}$$

$$z = W^T x + b$$

$$W = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

To find $\frac{\partial \text{cost}}{\partial w}$ & $\frac{\partial \text{cost}}{\partial b}$

$$\begin{bmatrix} \frac{\partial \text{cost}}{\partial w_1} \\ \vdots \\ \frac{\partial \text{cost}}{\partial w_n} \end{bmatrix}$$

$$\frac{\partial cost}{\partial w_1} = \frac{\partial cost}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1}$$

$$= \frac{\partial cost}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w_1}$$

$$cost = -\frac{1}{m} \sum_{i=1}^{m} y \log \hat{y} + (1-y) \log (1-\hat{y})$$

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$z = w_1 x_1 + w_2 x_2 + \dots + w_n x_n + b$$

$$\frac{\partial cost}{\partial \hat{y}} = -\frac{y}{\hat{y}} \overset{(-)}{+} \frac{(1-y)}{(1-\hat{y})}$$

$$= \frac{-y + \cancel{y\hat{y}} + \hat{y} - \cancel{y\hat{y}}}{\hat{y}(1-\hat{y})}$$

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

$$\frac{\partial \hat{y}}{\partial z} = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$= \frac{(1-\hat{y})}{\cancel{\hat{y}}} \hat{y}^{\cancel{2}}$$

$$= (1-\hat{y})\hat{y}$$

$$\frac{\partial z}{\partial w_1} = x_1$$

$$\frac{\partial cost}{\partial w_1} = \frac{-y+\hat{y}}{\cancel{\hat{y}(1-\hat{y})}} \cancel{(1-\hat{y})\hat{y}} \cdot x$$

$$= (\hat{y} - y)\, x$$

$$\boxed{\begin{gathered} \frac{\partial x^n}{\partial x} = n x^{n-1} \\ \frac{\partial}{\partial x}\left(\frac{1}{x}\right) = -\frac{1}{x^2} \\ \frac{\partial}{\partial x}(\log x) = \frac{1}{x} \\ \frac{\partial}{\partial x} e^x = e^x \\ \frac{\partial}{\partial x}[f(x)]^n = n[f(x)]^{n-1} \times \frac{\partial}{\partial x} f(x) \end{gathered}}$$

$$\frac{\partial z}{\partial b} = \frac{\partial \, cost}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial b}$$

$$= (\hat{y} - y) \times 1 \qquad \text{since } \frac{\partial z}{\partial b} = \frac{\partial b}{\partial b} = 1$$

$$\therefore \; W = W - \alpha \cdot \frac{\partial \, cost}{\partial w}$$

$$b = b - \alpha \cdot \frac{\partial \, cost}{\partial b}$$

α = Learning rate.

<u>Algorithm</u>: Logistic Regression

$X = [[\;], [\;], \ldots [\;]]$ → m samples; n features

$y = [\;]$ → m labels 0/1

Learning-rate = 0.1

Initialize params $w_1, \ldots w_n, b$

Repeat 1000 (N) times

- Compute loss using Binary Cross Entropy
- Determine differential of cost w.r.t. w_i's & b
- update w_i's & b values with differential and Learning-rate.

$\boxed{\text{Implementation in github}}$