Logistic Regression

Logistic Regression Overview https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148
Gradient Derivation: https://medium.com/analytics-vidhya/derivative-of-log-loss-function-for-logistic-regression-9b832f025c2d

- used for classification
- y-pred has range (0, 1) $\hat{y} = \sigma(x) = \frac{1}{1+e^{-(X\vec{w}+\vec{b})}}$

  define descision boundary to classify

- loss function (cost function) $= \begin{cases} -\log(\hat{y}_i) & \text{if } y_i = 1 \\ -\log(1-\hat{y}_i) & \text{if } y_i = 0 \end{cases}$

  we need sth. that gives us a convex function we can optimize over

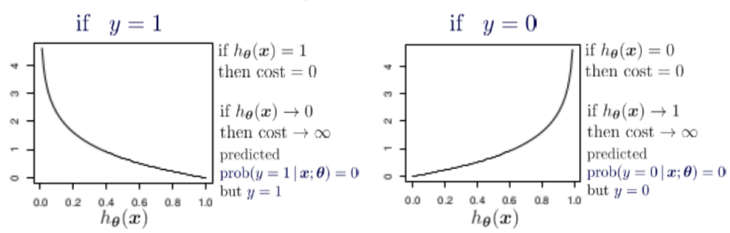
  combine

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

$$\text{cost}(h_\theta(x),\ y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1-h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$+ \log(\hat{y}_i)$

$- \log(1-y_i)$



- Gradient Calculation

  → Derivative of sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$

  $$\frac{\partial}{\partial x}\left(\frac{1}{1+e^{-x}}\right) = \frac{0 - 1 \times (-e^{-x})}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2}$$

  $$= \frac{1+e^{-x}}{(1+e^{-x})^2} - \frac{1}{(1+e^{-x})^2}$$

  $$= \sigma(x) - \sigma^2(x) = \sigma(x)(1-\sigma(x))$$

  $$\frac{\partial}{\partial x}\sigma(x) = \sigma(x)(1-\sigma(x))$$

→ Gradient of $\vec{w}$ wrt. loss

* $i \in [0, m)$ stands for # training input
$j \in [0, n)$ stands for features in X.
x_i is a row of the training input X.

$$l = \frac{1}{m} \sum_i \left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \right]$$

$$\frac{\partial l}{\partial w_j} = \frac{1}{m} \sum_i \left[\frac{y_i}{\hat{y}_i} \frac{\partial \hat{y}_i}{\partial w_j} - \frac{1-y_i}{1-\hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial w_j} \right]$$

$$\hat{y}_i = \sigma(x_i w + b)$$

→ use the sigmoid derivative & chain rule.

$$\frac{d\hat{y}_i}{d w_j} = \sigma(x_i w + b)\left(1 - \sigma(x_i w + b)\right) \cdot \frac{\partial}{\partial w_j}(x_i w + b)$$

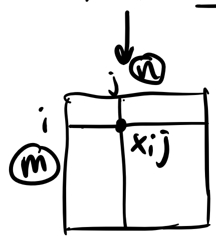
$$= \hat{y}_i (1-\hat{y}_i)\, x_{ij}$$

$$\frac{\partial l}{\partial w_j} = \frac{1}{m} \sum_i \left[\frac{y_i}{\cancel{\hat{y}_i}} \cancel{\hat{y}_i} (1-\hat{y}_i) x_{ij} - \frac{1-y_i}{\cancel{1-\hat{y}_i}} \hat{y}_i \cancel{(1-\hat{y}_i)} x_{ij} \right]$$

$$\frac{\partial l}{\partial w_j} = \frac{1}{m} \sum_i x_{ij} \left[y_i (1-\hat{y}_i) - \hat{y}_i (1-y_i) \right]$$

$$\frac{\partial l}{\partial w_j} = \frac{1}{m} \sum_i x_{ij} \underline{(y_i - \hat{y}_i)}$$

jth term of the gradient



i th term of error vector $\vec{e}$

$$\frac{\partial l}{\partial \vec{w}} = \frac{1}{m} X^T \vec{e}$$

- Gradients of logistic Regression is the <u>same</u> as gradients of linear regression! Thanks to the design of the loss function :)