Derivative of Cost Function of Logistic Regression:

Odds:

$$odds = \frac{probability\ event\ occuring}{probability\ event\ NOT\ occuring} = \frac{p}{1-p} \tag{1}$$

- Log-Odds = Logit:

$$Logit(p) = log(odds) = \log\left(\frac{p}{1-p}\right) \tag{2}$$

- **linear equation**

$$\theta.x = \theta_0.1 + \theta_1.x_1 + \cdots\ \theta_n.x_n \tag{3}$$

- substitute

$$z = \theta.x \tag{3}$$

- **linear combination**:

$$Logit(p) = z$$

$$log\left(\frac{p}{1-p}\right) = z \tag{4}$$

- calculate logarithm:

$$\frac{p}{1-p} = e^z \tag{5}$$

$$p = e^z(1-p)$$

$$p = e^z - p.e^z \tag{6}$$

$$p + p.e^z = e^z$$

$$p(1+e^z) = e^z$$

$$p = \frac{\boldsymbol{e^z}}{\boldsymbol{1+e^z}} \tag{7}$$

$$p = \frac{\boldsymbol{e^z}}{\boldsymbol{1+e^z}} = \frac{\frac{\boldsymbol{e^z}}{\boldsymbol{e^z}}}{\frac{\boldsymbol{1+e^z}}{\boldsymbol{e^z}}} = \frac{\boldsymbol{1}}{\boldsymbol{1+e^{-z}}}$$

Probability Mass Function (PMF) of Bernoulli distribution:

$$P(y) = \begin{Bmatrix} p & if & y=1 \\ 1-p & if & y=0 \end{Bmatrix} \rightarrow P(y) = p^y.(1-p)^{1-y} \tag{8}$$

Likelihood function for logistic regression:

- for a set of independent Bernoulli trials with different probabilities p_i

- we want to find the maximum of this function!!!

$$\boldsymbol{L(\theta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}} \tag{9}$$

$$L(\theta) = \left[p_1^{y_1}(1-p_1)^{1-y_1}\right].\left[p_2^{y_2}(1-p_2)^{1-y_2}\right] \ldots\ldots\ldots \left[p_n^{y_n}(1-p_n)^{1-y_n}\right]$$

- algorithmize the equation (11):

- rule:

$$log(n^k) = k.log(n) \tag{10}$$

- applying the rule, we get the log likelihood function:

$$l(\theta) = \sum_{i=1}^{n} [y_i.log(p_i) + (1 - y_i).log(1 - p_i)] \tag{11}$$

Average Cost function (Objective function):

- used in machine learning
- **we want find minimum of this function!!!**

$$J(\theta) = -\frac{1}{n} l(\theta) \tag{12}$$

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} [y_i.log(p_i) + (1 - y_i).log(1 - p_i)] \tag{13}$$

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} [y_i.log(p_i) + (1 - y_i).log(1 - p_i)] \tag{14}$$

Derivative of the cost function:

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \left[-\frac{1}{n} \sum_{i=1}^{n} [y_i.log(p_i) + (1 - y_i).log(1 - p_i)] \right] \tag{15}$$

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i.\boldsymbol{\frac{\partial}{\partial \theta} log(p_i)} + (1 - y_i) \boldsymbol{\frac{\partial}{\partial \theta} log(1 - p_i)} \right] \tag{16}$$

- chain rule (outer and then inner function derivative):

$$\boldsymbol{\frac{\partial}{\partial \theta} log(p_i)} = \frac{\partial log(p_i)}{\partial \theta}.\frac{\partial p_i}{\partial \theta} = \frac{1}{p_i}.\frac{\partial p_i}{\partial \theta} \tag{16.1}$$

$$\boldsymbol{\frac{\partial}{\partial \theta} log(1 - p_i)} = \frac{\partial log(1 - p_i)}{\partial \theta}.\frac{\partial (1 - p_i)}{\partial \theta} = \frac{1}{(1 - p_i)}.\frac{\partial (1 - p_i)}{\partial \theta} = -\frac{1}{(1 - p_i)}.\frac{\partial p_i}{\partial \theta} \tag{16.2}$$

- plug both derivatives back to the equation:

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i}{p_i}.\frac{\partial p_i}{\partial \theta} - \frac{(1 - y_i)}{(1 - p_i)}.\frac{\partial p_i}{\partial \theta} \right]$$

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{(1 - y_i)}{(1 - p_i)}. \right] \boldsymbol{\frac{\partial p_i}{\partial \theta}} \tag{17}$$

- because p_i is basically a sigmoid function I can re-write it like this:

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i}{\sigma_{zi}} - \frac{(1 - y_i)}{(1 - \sigma_{zi})}. \right] \boldsymbol{\frac{\partial \sigma_{zi}}{\partial \theta}} \tag{18}$$

- chain rule (derivative of the outer function multiplied by the derivative of the inner function):

$$\frac{\partial}{\partial x} f(g) = \frac{\partial f(g)}{\partial g} . \frac{\partial g}{\partial x} \tag{19}$$

- applying a chain rule

$$\frac{\partial \sigma_{zi}}{\partial \theta_j} = \frac{\boldsymbol{\partial \sigma_{zi}}}{\boldsymbol{\partial z_i}} . \frac{\boldsymbol{\partial z_i}}{\boldsymbol{\partial \theta_j}} \tag{20}$$

- second part

$$\frac{\boldsymbol{\partial z_i}}{\boldsymbol{\partial \theta_j}} = \frac{\partial \left(\theta_j . x_{ij}\right)}{\partial \theta_j} = x_{ij} \tag{21}$$

- first part

$$\frac{\boldsymbol{\partial \sigma_{zi}}}{\boldsymbol{\partial z_i}} = \frac{\partial}{\partial z_i}\left(\frac{1}{1+e^{-z}}\right) = \frac{\partial}{\partial z_i}(1+e^{-z})^{-1} \tag{22}$$

- substitute:

$$u = 1+e^{-z} \tag{23}$$

- substitute and apply chain rule:

$$\begin{aligned}
\frac{\boldsymbol{\partial \sigma_{zi}}}{\boldsymbol{\partial z_i}} &= \frac{\partial (u)^{-1}}{\partial u} . \frac{\partial u}{\partial z_i} = -1 . u^{-2} . \frac{\partial u}{\partial z_i} = -(1+e^{-z})^{-2} . \frac{\partial (1+e^{-z})}{\partial z_i} = -(1+e^{-z})^{-2} . [0 - 1 . e^{-z}] \\
&= (1+e^{-z})^{-2} . e^{-z} = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1-1}{(1+e^{-z})^2} + \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1-1+e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1}{(1+e^{-z})} . \left[\frac{1-1+e^{-z}}{(1+e^{-z})}\right] = \frac{1}{(1+e^{-z})} . \left[\frac{1-(1+e^{-z})}{(1+e^{-z})}\right] = \frac{1}{(1+e^{-z})} . \left[\frac{-1+(1+e^{-z})}{(1+e^{-z})}\right] = \\
&= \frac{1}{(1+e^{-z})} . \left[-\frac{1}{(1+e^{-z})} + \frac{\cancel{(1+e^{-z})}}{\cancel{(1+e^{-z})}}\right] = \frac{1}{(1+e^{-z})} . \left[1 - \frac{1}{(1+e^{-z})}\right] = \sigma_{zi} . (1-\sigma_{zi})
\end{aligned} \tag{24}$$

- plug both part in the main equation:

$$\frac{\boldsymbol{\partial \sigma_{zi}}}{\boldsymbol{\partial \theta_j}} = \frac{\boldsymbol{\partial \sigma_{zi}}}{\boldsymbol{\partial z_i}} . \frac{\boldsymbol{\partial z_i}}{\boldsymbol{\partial \theta_j}} = [\sigma_{zi} . (1-\sigma_{zi})] . x_{ij} \tag{25}$$

- final equation:

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{n} \sum_{i=1}^{n} \left[\frac{\boldsymbol{y_i}}{\boldsymbol{\sigma_{zi}}} - \frac{\boldsymbol{(1-y_i)}}{\boldsymbol{(1-\sigma_{zi})}}\right] . [\sigma_{zi} . (1-\sigma_{zi}) . x_{ij}] \tag{26}$$

- simplify the first part by finding a common denominator:

$$\frac{\boldsymbol{y_i}}{\boldsymbol{\sigma_{zi}}} - \frac{\boldsymbol{(1-y_i)}}{\boldsymbol{(1-\sigma_{zi})}} = \frac{y_i . (1-\sigma_{zi}) - (1-y_i) . \sigma_{zi}}{\sigma_{zi} . (1-\sigma_{zi})} = \frac{y_i - \cancel{y_i \sigma_{zi}} - \sigma_{zi} + \cancel{y_i . \sigma_{zi}}}{\sigma_{zi} . (1-\sigma_{zi})} = \frac{y_i - \sigma_{zi}}{\sigma_{zi} . (1-\sigma_{zi})} \tag{27}$$

- plug back into the main equation:

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta} &= -\frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i - \sigma_{zi}}{\cancel{\sigma_{zi} . (1-\sigma_{zi})}}\right] . [\cancel{\sigma_{zi} . (1-\sigma_{zi})} . x_{ij}] \\
\frac{\partial J(\theta)}{\partial \theta} &= -\frac{1}{n} \sum_{i=1}^{n} [(y_i - \sigma_{zi}) . x_{ij}]
\end{aligned} \tag{28}$$