Logistic Regression – Loss function & Gradient Descent

$$J(w,b) = -\frac{1}{m}\sum_i L\left(a^i, y^i\right)$$

$$= -\frac{1}{m}\sum_i y^i \log a^i + \left(1-y^i\right)\log\left(1-a^i\right)$$

where $a^i = \hat{y}^i = 1/\left(1+e^{-w^T x^i}\right)$.

$J(w,b)$ is the cost/loss function; while $-J(w,b)$ is the log likelihood function (can be derived from Bernoulli). To minimize the cost/loss function is equivalent to maximize the likelihood function. The gradient descent method can be used to minimize the cost/loss function; while Newton's method can be used to find the maximum arguments for parameters by solving the 1st derivative of $-J(w,b)$.

$$\frac{\partial a}{\partial w_i} = -\frac{1}{\left(1+e^{-w^T x}\right)^2}\cdot e^{-w^T x}\cdot(-x_i) = a^2\cdot e^{-w^T x}\cdot x_i$$

$$= a^2\cdot\frac{1-a}{a}\cdot x_i = a\cdot(1-a)\cdot x_i$$

where we have $e^{-w^T x} = \frac{1-a}{a}$

$$\frac{\partial a}{\partial b} = a\cdot(1-a)$$

$$\frac{\partial J}{\partial w_i} = -\frac{1}{m}\sum_j\left(y\cdot\frac{1}{a}\cdot\frac{\partial a}{\partial w_i} + (1-y)\cdot\frac{1}{1-a}\cdot\left(-\frac{\partial a}{\partial w_i}\right)\right)$$

$$= -\frac{1}{m}\sum_j\left(y\cdot\frac{1}{a}\cdot a(1-a)x_i + (1-y)\cdot\frac{1}{1-a}\cdot(-a(1-a)x_i)\right)$$

$$= -\frac{1}{m}\sum_j (y(1-a)x_i - (1-y)ax_i)$$

$$= -\frac{1}{m}\sum_j (-yx_i + yax_i + ax_i - yax_i)$$

$$= \frac{1}{m}\sum_j (a-y)x_i$$

where a, y and x_i are vectors. Also the constant term is

$$\frac{\partial J}{\partial b} = \frac{1}{m}\sum_j (a-y)$$