Q2) $P(y=1|x,w) = \frac{1}{2}\left(1 + \frac{w^T x}{\sqrt{1+(w^T x)^2}}\right)$

$P(y=0|x,w) = 1 - P(y=1|x,w)$

$J(w)$ = Cost function

$$= -\left\{\sum_{i=1}^{n} y^i \log P(y=1|x_i,w) + (1-y^i)\log P(y=0|x_i,w)\right\}$$

*Neglecting the -ve sign for now and forming a generalized equation without Σ, will add Σ later

$$J(w)^* = y^i \log P(y=1|x_i,w) + (1-y^i)\log P(y=0|x_i,w)$$

$$= y^i \log \frac{1}{2}\left(1+\frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right) + (1-y^i)\log\left(1 - \frac{1}{2}\left(1+\frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right)\right)$$

$$= y^i \log \frac{1}{2}\left(1+\frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right) + (1-y^i)\log \frac{1}{2}\left(1-\frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right)$$

$\Rightarrow \log a \times b = \log a + \log b \quad a = \frac{1}{2} ; b = \left(1+\frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right)$ or $\left(1-\frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right)$

Since $y^i \log a$ and $(1-y^i)\log a$ are gonna be constant when we derive/differentiate $J(w)^*$ with respect to w_{ij} there is no need to solve it and we neglect the terms.

$$\frac{\partial J(w)^{*}}{\partial w_{ij}} = y^i \frac{\log\left(1 + \frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right)}{\partial w_{ij}} + (1-y^i)\frac{\log\left(1 - \frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right)}{\partial w_{ij}}$$

$$\Rightarrow \frac{\log x}{\partial x'} = \frac{1}{x} \times \frac{\delta x}{\partial x'} \quad ; \quad \frac{\delta a \times b}{\partial x} = b\frac{\partial a}{x} + \frac{a\,\partial b}{x}$$ where a and b are functions of x

Using the rules

$$= \frac{y^i \sqrt{1+(w^T x_i)^2}}{\sqrt{1+(w^T x_i)^2} + w^T x_i}\left(\frac{x_{ij}}{\sqrt{1+(w^T x_i)^2}} - \frac{\cancel{2}}{\cancel{2}}\frac{(w^T x)^2 \times x_{ij}}{\left(1+(w^T x_i)^2\right)^{3/2}}\right)$$

$$+ \frac{(1-y_i)\sqrt{1+(w^T x_i)^2}}{\sqrt{1+(w^T x_i)^2} - w^T x_i}\left(-\frac{x_{ij}}{\sqrt{1+(w^T x_i)^2}} + \frac{\cancel{2}}{\cancel{2}}\frac{(w^T x)^2 \times x_{ij}}{\left(1+(w^T x_i)^2\right)^{3/2}}\right)$$

$$= \frac{y^i \cancel{\sqrt{1+(w^T x_i)^2}}}{\left\{\sqrt{1+(w^T x_i)^2} + (w^T x_i)\right\} \times \cancel{\sqrt{1+(w^T x_i)^2}}}\left(x_{ij} - \frac{(w^T x_i)^2 x_{ij}}{1+(w^T x_i)^2}\right)$$

$$+ \frac{(1-y_i)\cancel{\sqrt{1+(w^T x_i)^2}}}{\left\{\sqrt{1+(w^T x_i)^2} + (w^T x_i)\right\} \times \cancel{\sqrt{1+(w^T x_i)^2}}}\left(-x_{ij} + \frac{(w^T x_i)^2 x_{ij}}{1+(w^T x_i)^2}\right)$$

$$= \frac{y^i x_{ij}}{\sqrt{1+(w^T x_i)^2}+(w^T x_i)} \left(\frac{1 + \cancel{(w^T x_i)^2} - \cancel{(w^T x_i)^2}}{1+(w^T x_i)^2} \right)$$

$$+ \frac{(1-y^i)\, x_{ij}}{\sqrt{1+(w^T x)^2}-(w^T x_i)} \left(\frac{-1 - \cancel{(w^T x_i)^2} + \cancel{(w^T x_i)^2}}{1+(w^T x_i)^2} \right)$$

$$= \frac{x_{ij}}{((w^T x_i)^2+1)} \left\{ \frac{y_i}{\left(\sqrt{1+(w^T x_i)^2}+(w^T x_i)\right)} \times 1 + \frac{(y_i - 1)}{\left(\sqrt{1+(w^T x_i)^2}-(w^T x_i)\right)} \right\}$$

$$= \frac{x_{ij}}{(w^T x_i)^2+1} \left(\frac{y_i\left(\sqrt{1+(w^T x_i)^2} - w^T x_i\right) + (y_i - 1)\left(\sqrt{1+(w^T x_i)^2} + w^T x_i\right)}{\left(\sqrt{1+(w^T x_i)^2}\right)^2 - (w^T x_i)^2} \right)$$

$$\Rightarrow (a+b)(a-b) = a^2 - b^2$$

$$= \frac{x_{ij}}{(w^T x_i)^2+1} \left(2y_i \sqrt{1+(w^T x_i)^2} - \sqrt{1+(w^T x_i)^2} - w^T x_i \right)$$

$$= x_{ij} \left(\frac{2y_i}{\sqrt{1+(w^T x_i)^2}} - \frac{1}{\sqrt{1+(w^T x_i)^2}} - \frac{w^T x_i}{1+(w^T x_i)^2} \right)$$

So $$\frac{\delta J(\omega)}{\delta w_{ij}} = x_{ij}\left(\frac{2y_i}{\sqrt{p}} - \frac{1}{\sqrt{p}} - \frac{w^T x}{p}\right)$$

where $p = (1 + w^T x_i)^2$

This is the gradient descent with respect to weight j. Since we have n samples $i \in n$. We can use batch or stochastic gradient descent to calculate the optimal weights.

$$w^j_{k+1} = w^j_k - \alpha \frac{\delta J(w)}{\delta w_j} \quad \text{(iterative approach)}$$

$$\frac{\delta J(\omega)}{\delta w_j} = \left(\sum_{i=1}^{n} \frac{\delta J(\omega)}{\delta w_{ij}}\right) \Big/ n$$; Since earlier I had omitted the -ve sign from cost function which directly impacts the gradient descent function since it is a derivation of $J(\omega)$ cost function. I will add the -ve sign direct to step rule of iterative approach of my model. So updated rule looks like

$$w^j_{k+1} = w^j_k + \alpha \frac{\delta J(\omega)}{\delta w_j}$$

where w^j is the jth feature w_k is the k the iteration

& ~~&~~ α is the step size.