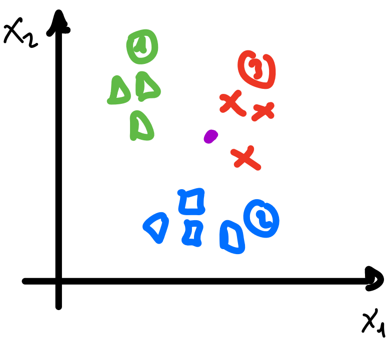
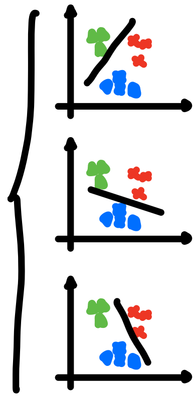
⑥ Supervised learning II

* Multi-classification

* One-vs-all :

$y \in \{1, 2, 3\}$





$g(\theta^T x) = P(y=1|x;\theta)$

$g(\theta^T x) = P(y=2|x;\theta)$

$g(\theta^T x) = P(y=3|x;\theta)$

* ¿ How is the cost?

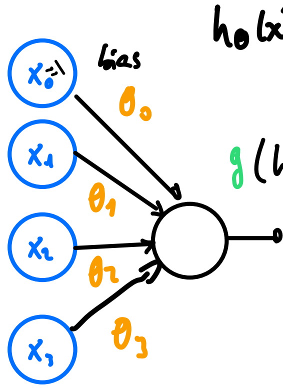
$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\Big[\sum_{k=1}^{K}(1-y^i_k)(-\log(1-g(h_\theta(x^{(i)}))_k) +$$
$$y^i_k(-\log(g(h_\theta(x^{(i)}))_k)) + \lambda\|\theta\|^2\Big]$$

$$y_1 = \begin{bmatrix}1\\0\\0\end{bmatrix};\quad y_2 = \begin{bmatrix}0\\1\\0\end{bmatrix};\ \ldots$$

• Neural model : Logistic unit

$$h_\theta(x) = X\theta = x_1\theta_1 + x_2\theta_2 + \underset{\overset{\uparrow}{1}}{x_0}\theta_0$$

Linear case



$g(h_\theta(x^{(i)}))$

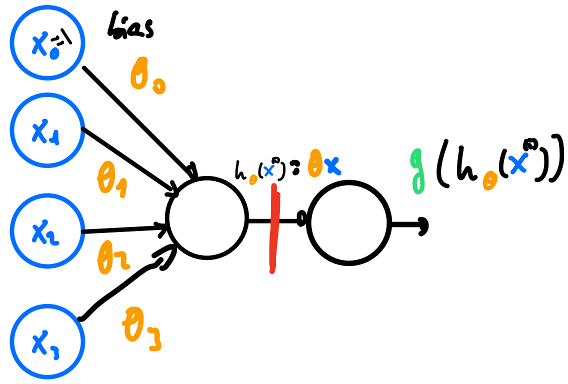
$x = [x_0, x_1, \ldots, x_3]$

$\theta = [\theta_0, \theta_1, \ldots, \theta_3]$

$g \equiv$ sigmoid activation function (logistic)

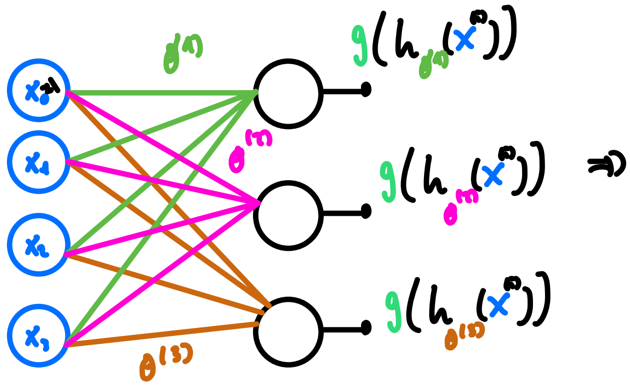
x_1^2

or more schematically



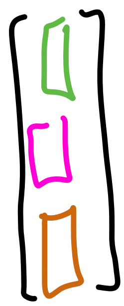
In <u>one vs all</u> :

$$\begin{bmatrix}1\\0\\0\end{bmatrix};\begin{bmatrix}0\\1\\0\end{bmatrix};\begin{bmatrix}0\\0\\1\end{bmatrix}$$



$$\Rightarrow \begin{bmatrix} g(h_{\theta^{(1)}}(x^{(i)})) \\ g(h_{\theta^{(2)}}(x^{(i)})) \\ g(h_{\theta^{(3)}}(x^{(i)})) \end{bmatrix} \overset{?}{\sim} \begin{bmatrix} y \end{bmatrix}$$

$$\theta = \left[\theta^{(1)}, \theta^{(2)}, \theta^{(3)}\right] \Rightarrow$$



$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\Big[\sum_{k=1}^{K}(1-y^i_k)(-\log(1-g(h_\theta(x^{(i)}))))_k + y^i_k(-\log(g(h_\theta(x^{(i)}))))_k\Big] + \lambda\|\theta\|^2$$

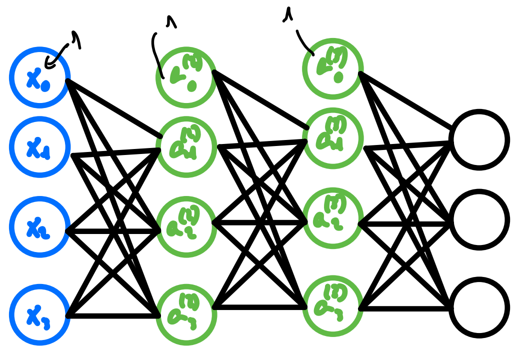
$$= J_1(\theta_1) + J_2(\theta_2) + J_3(\theta_3) + \lambda\|\theta\|^2$$

$$\nabla_{\theta_1} J = \frac{1}{m}\sum_{i=1}^{m}\left[g(h_{\theta_1}(x^{(i)})) - y^i_1\right]x^{(i)}$$

$$\nabla_{\theta_2} J = \frac{1}{m}\sum_{i=1}^{m}\left[g(h_{\theta_2}(x^{(i)})) - y^i_2\right]x^{(i)}$$

$$\nabla_{\theta_3} J = \frac{1}{m}\sum_{i=1}^{m}\left[g(h_{\theta_3}(x^{(i)})) - y^i_3\right]x^{(i)}$$

* Neural network (classification)



$\{(x^{(1)},y^{(1)})\,(x^{(2)},y^{(2)})\,(x^{(3)},y^{(3)})\,(x^{(m)},y^{(m)})\}$

L = total number of Layers

s_l = number of units in layer l

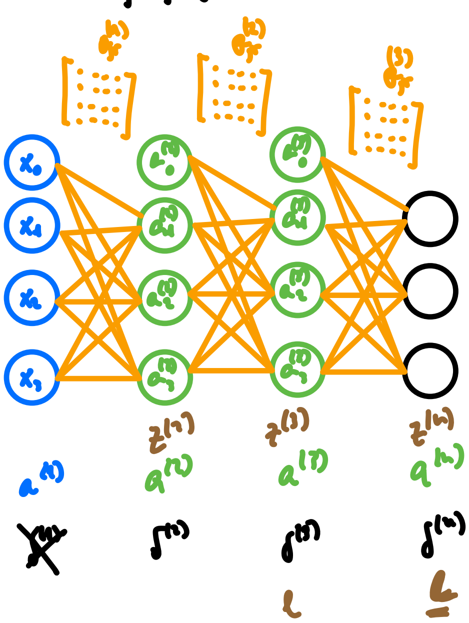
* Cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\Big[\sum_{k=1}^{K}(1-y_k^i)\big(-\log(1-g(h_\theta(g(h_\theta(\cdots(x^{(i)})\cdots)))_k\big) + y_k^i\big(-\log(g(\underbrace{h_\theta(x^{(i)})}_{h_\theta(x)}))_k\big)\Big]$$

$$+\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{r=1}^{s_l}\sum_{j}^{s_{l+1}}\left(\theta_{jr}^{(l)}\right)^2$$

but not 0

* Forward propagation



$$\left\{\begin{aligned}
&a^{(1)} = x\\
&z^{(2)} = \theta^{(1)}a^{(1)}\\
&a^{(2)} = g(z^{(2)})\\
&z^{(3)} = \theta^{(2)}a^{(2)}\\
&a^{(3)} = g(z^{(3)})\\
&z^{(4)} = \theta^{(3)}a^{(3)}\\
&a^{(4)} = g(z^{(4)})
\end{aligned}\right.$$

Remember:

$$\left\{\begin{aligned}
&g(z) = \frac{1}{1+e^{-z}}\\
&g'(z) = g(z)(1-g(z))
\end{aligned}\right.$$

* Gradient.

Before : one layer architecture

$$h = \theta x$$

$$\frac{\partial J(\theta)}{\partial \theta_j^k} = \frac{1}{m}\sum_{i=1}^{m}(1-y_k^i)\,\frac{-1}{\cancel{(1-g(h_\theta(x^{(i)})))}}\;-(\overbrace{g(h_\theta(x^{(i)}))\cancel{(1-g(h_\theta(x^{(i)})))}}^{g'(h_\theta(x^{(i)}))})\,\overset{\frac{\partial h}{\partial \theta}}{x^{(i)}}$$

$$+\,(y_k^i)\,\frac{-1}{\cancel{(g(h_\theta(x^{(i)})))}}\;\overbrace{\cancel{(g(h_\theta(x^{(i)}))}(1-g(h_\theta(x^{(i)})))}^{g'(h_\theta(x^{(i)}))}\,x^{(i)}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[g(h_\theta(x^{(i)})) - y_k^i\right]x^{(i)}$$

Now: several layer architecture

* For the last layer: $(L-1)$

$$\frac{\partial J(\theta)}{\partial \theta^{L-1}} = \frac{1}{m}\sum_{i=1}^{m}(1-y^i)\,\frac{-1}{\cancel{(1-g(z^{(L)}))}}\;-(\overbrace{g(z^L)\cancel{(1-g(z^L))}}^{g'(z^L)})\,\frac{\partial z^L}{\partial \theta}$$

$$+\,(y^i)\,\frac{-1}{\cancel{(g(z^L))}}\;\overbrace{\cancel{(g(z^L))}\,(1-g(z^L))}^{g'(z^L)}\,\frac{\partial z^L}{\partial \theta}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left[\underbrace{g(z^L)-y^i}_{\delta^L}\right]\frac{\partial z^L}{\partial \theta} = \frac{1}{m}\sum \delta^L\frac{\partial z^L}{\partial \theta}$$

$$z^{(L)} = \theta^{(L-1)}a^{(L-1)}$$

$$\frac{\partial J(\theta)}{\partial \theta^{L-1}} = \frac{1}{m}\sum_{i=1}^{m}\delta^L\frac{\partial z^L}{\partial \theta^{L-1}} = \frac{1}{m}\sum_{i=1}^{m}\delta^L a^{(L-1)}$$

$$\boxed{\frac{\partial J(\theta)}{\partial \theta^{L-1}} = \frac{1}{m}\sum_{i=1}^{m}\delta^L a^{(L-1)}} \quad \text{with} \quad \boxed{\delta^L = g(z^L) - y^i}$$

* For one before the last layers: $(L-2)$

$$\frac{\partial J(\theta)}{\partial \theta^{L-2}} = \frac{1}{m}\sum_{i=1}^{m} \delta^{L} \frac{\partial z^{L}}{\partial \theta^{L-2}} \overset{z^{(L)} = \theta^{(L-1)} a^{(L-1)}}{=} \frac{1}{m}\sum_{i=1}^{m} \delta^{L} \theta^{(L-1)} \frac{\partial a^{(L-1)}}{\partial \theta^{L-2}}$$

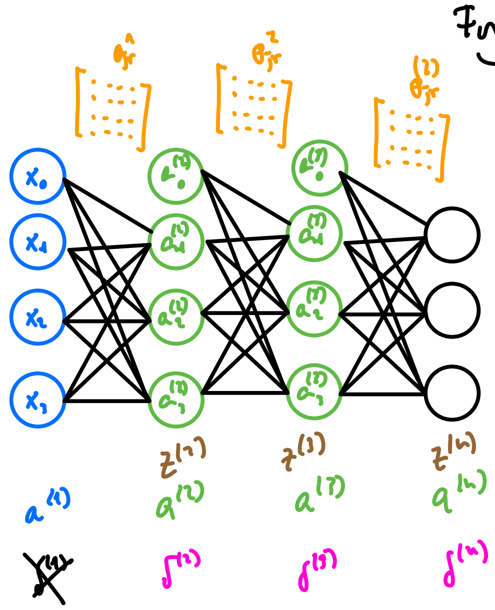
$$a^{(L-1)} = g(z^{(L-1)})$$

$$= \frac{1}{m}\sum_{i=1}^{m} \delta^{L} \theta^{(L-1)} \overbrace{\frac{\partial a^{(L-1)}}{\partial z^{(L-1)}}}^{g(z^{(L-1)})(1-g(z^{(L-1)}))} \overbrace{\frac{\partial z^{(L-1)}}{\partial \theta^{L-2}}}^{a^{(L-2)}} = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\delta^{L} \theta^{(L-1)} g(z^{(L-1)})(1-g(z^{(L-1)}))}_{\delta^{L-1}} a^{(L-2)}$$

$$\boxed{\frac{\partial J(\theta)}{\partial \theta^{L-2}} = \frac{1}{m}\sum_{i=1}^{m} \delta^{L-1} a^{(L-2)}} \quad \text{with} \quad \boxed{\delta^{L-1} = \delta^{L} \theta^{(L-1)} g(z^{(L-1)})(1-g(z^{(L-1)}))}$$

* For one before the last layers: $(L-3)$ works similarly

Backpropagation



Forward propagation

$a^{(1)} = x$

$z^{(2)} = \theta^{(1)} a^{(1)}$

$a^{(2)} = g(z^{(2)})$

$z^{(3)} = \theta^{(2)} a^{(2)}$

$a^{(3)} = g(z^{(3)})$

$z^{(4)} = \theta^{(3)} a^{(3)}$

$a^{(4)} = g(z^{(4)})$

Backward propagation

error on layer 4

$\delta^{(4)} = a^{(4)} - y$ vectorial: $\delta^{(4)}_j = a^{(4)}_j - y_j$

$\delta^{(3)} = \theta^{(3)} \delta^{(4)} .* \, g'(z^{(3)})$ — $g(z^{(3)})(1-g(z^{(3)}))$

$\delta^{(2)} = \theta^{(2)} \delta^{(3)} .* \, g'(z^{(2)})$ — point-wise

No $\delta^{(1)}$

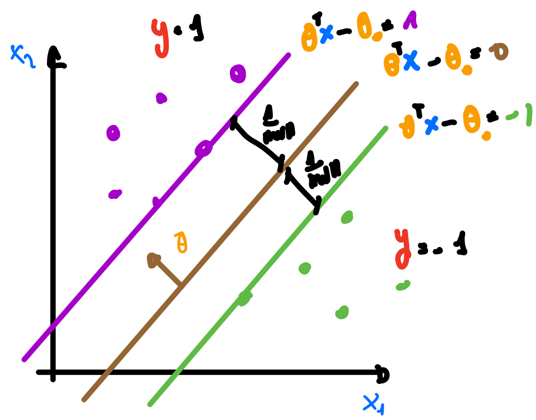
"We propagate the error"

$$\frac{\partial J(\theta)}{\partial \theta^{(l)}_{jr}} = \frac{1}{m}\sum_{i=1}^{m} a^{(l)}_j \delta^{(l+1)}_r$$

* Remarks:
  * Now people use automatic differentiation
  * Also use regularization

Support vector machines

SVM (Supp. vect. machine)



Linearly separable:

$$\begin{cases} \theta^T x - \theta_0 \geq 1 & \text{when } y_i = 1 \\ \theta^T x - \theta_0 \leq -1 & \text{when } y_i = -1 \end{cases}$$

$$\to \quad y_i(\theta^T x_i - \theta_0) \geq 1$$

$$\begin{cases} \max\limits_{\theta, \theta_0} \frac{2}{\|\theta\|} \\ \text{s.t. } y_i(\theta^T x_i - \theta_0) \geq 1 \end{cases}$$

$$\to \begin{cases} \min\limits_{w, b} \|\theta\|^2 \\ \text{s.t. } y_i(\theta^T x_i - \theta_0) \geq 1 \end{cases}$$

Non-linearly separable

Primal

$$\begin{cases} \min\limits_{\theta\, \theta_0} \|\theta\| \\ \text{s.t. } \underbrace{y_i(\theta^T x_i - \theta_0) - 1 \geq 0}\ \ \forall i \end{cases}$$

$$f_i(\theta) \leq 0 \to 1 - y_i(\theta^T x_i - \theta_0) \leq 0 \quad \forall i$$

Primal (relaxed)

$$\begin{cases} \min\limits_{\xi_i, \theta, \theta_0} \frac{1}{n}\sum\limits_{i=1} \xi_i + \mu\|\theta\|^2 \\ \text{s.t. } \quad y_i(w^T x_i - b) \geq 1 - \xi_i \quad (\lambda_i) \end{cases}$$

Dual

$$\begin{cases} \max\limits_{c} \sum \lambda_i - \frac{1}{2}\sum\limits_{i=1}\sum\limits_{j=1} \lambda_i y_i (x_i^T x_j) y_j \lambda_j \\ \text{s.t. } \sum \lambda_i y_i = 0 \quad \text{with } 0 \leq \lambda_i \leq \frac{1}{2n\mu} \end{cases}$$