Neural Networks: Non-linear hypotheses

typically have more features than just 2

if $n=100$, 2nd order terms $\to \approx 5000$ features $O(n^2) \approx \frac{n^2}{2}$

subset

3rd order terms $O(n^3)$ $\approx 170{,}000$ features

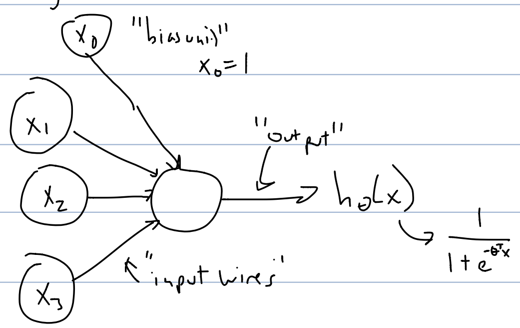
Neurons and the brain

algorithms to try to mimic the brain

"one learning algorithm" hypothesis

Model Representation

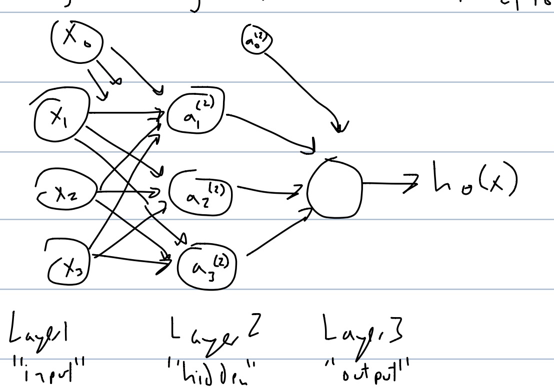
Logistic unit



$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \Theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

"weights"/params

Sigmoid (logistic) activation function



$a_i^{(j)}$ = "activation" of unit i in layer j

$\Theta^{(j)}$ = matrix of weights controlling function mapping from layer j to layer $j+1$

$$a_1^{(2)} = g(\Theta_{10}^{(1)} x_0 + \Theta_{11}^{(1)} x_1 + \Theta_{12}^{(1)} x_2 + \Theta_{13}^{(1)} x_3)$$

$a_2^{(2)} = g(\Theta_{20}^{(1)} x_0 \ldots\ldots$

$h_\Theta(x) = a_1^{(3)} = g(\Theta_{10}^{(2)} a_0^{(2)} + \Theta_{11}^{(2)} a_1^{(2)} + \Theta_{12}^{(2)} a_2^{(2)} + \Theta_{13}^{(2)} a_3^{(2)})$

If network has s_j units in layer j, s_{j+1} units in layer $j+1$, then $\Theta^{(j)}$ will be of dimension

$s_{j+1} \times (s_j + 1)$

Model Representation II

$X = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad z^{(2)} = \begin{bmatrix} z_1^{(2)} \\ z_2^{(2)} \\ z_3^{(2)} \end{bmatrix}$

$a_1^{(2)} = g(z_1^{(2)})$

$a_2^{(2)} = g(z_2^{(2)})$

$a_3^{(2)} = g(z_3^{(2)})$

$z^{(2)} = \Theta^{(1)} x$

$\underbrace{a^{(2)}}_{\in \mathbb{R}^3} = g(z^{(2)})$

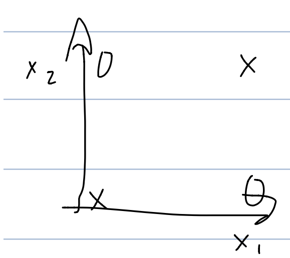
Add $a_0^{(2)} = 1 \rightarrow a^{(2)} \in \mathbb{R}^4$

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

$h_\Theta(x) = a^{(3)} = g(z^{(3)})$

Examples and Intuitions

x_1, x_2 are binary (0 or 1)

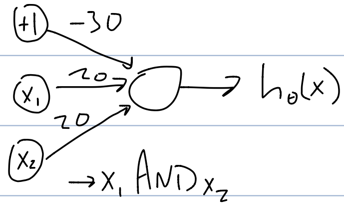


$y = x_1$ XOR x_2

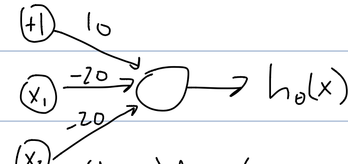
→ x_1 XNOR x_2

→ NOT (x_1 XOR x_2)

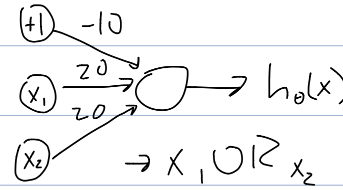
Examples and Intuitions II



$\rightarrow x_1$ AND x_2



$\rightarrow$ (NOT x_1) AND (NOT x_2)



$\rightarrow x_1$ OR x_2

Multi-class Classification

One vs. all