Machine Learning - Week 4 - Neural Networks: Representation

T1 - Motivations

L1 - Non-linear Hypotheses

Non-linear Classification
- Using logistic regression can lead to there being way too many features ($O(n^2)$ features if quadratic, $O(n^3)$ features for cubic)

L2 - Neurons & the Brain

The "one learning algorithm" hypothesis

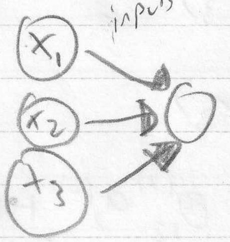
T2 - Neural Networks

L1 - Model Representation I

Look at how neurons in the brain work
- Pass messages from the axon of one neuron to the dendrites of the receiving neuron

Neuron model: Logistic unit

Singular



$\rightarrow h_\theta(x)$

$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$, $x = \begin{bmatrix} \cancel{x_0} \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$ $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$

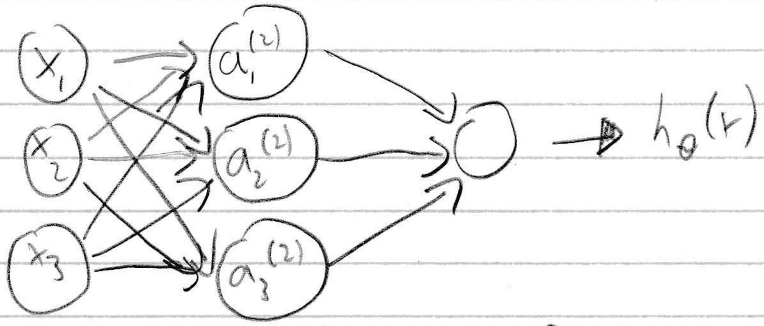
Note: x_0 input is the "bias unit", since $x_0 = 1$

Sigmoid (logistic) activation function

We have been calling θ the "parameters" of the model, but in neural networks they can be called "weights" of the model.

$\Rightarrow$

Neural Network



Layer 1 Input layer x

Layer 2 Hidden layer

Layer 3 Output layer y

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

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

In above nn example:

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

$$a_2^{(2)} = g(\Theta_{20}^{(1)} x_0 + \Theta_{21}^{(1)} x_1 + \Theta_{22}^{(1)} x_2 + \Theta_{23}^{(1)} x_3)$$

$$a_3^{(2)} = g(\Theta_{30}^{(1)} x_0 + \Theta_{31}^{(1)} x_1 + \Theta_{32}^{(1)} x_2 + \Theta_{33}^{(1)} x_3)$$

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

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

L2 - Model Representation 2

Recall from previous lecture the example with the long formula...

lets set.

$$a_1^{(2)} = g(z_1^{(2)})$$
$$a_2^{(2)} = g(z_2^{(2)})$$
$$a_3^{(2)} = g(z_3^{(2)})$$

Say

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \underbrace{z^{(2)}}_{\mathbb{R}^3} = \begin{bmatrix} z_1^{(2)} \\ z_2^{(2)} \\ z_3^{(2)} \end{bmatrix}$$

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

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

Now add $a_0^{(2)} = 1$

$\therefore a^{(2)} \in \mathbb{R}^4$

$$\therefore z^{(3)} = \Theta^{(2)} a^{(2)}$$
$$h_\Theta(x) = a^{(3)} = g(z^{(3)})$$

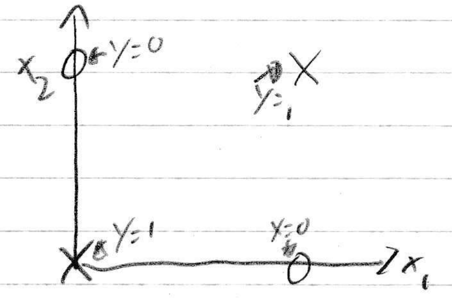
This whole process is called Forward propagation using vectorized implementation

T2 - Applications

L1 - Examples and Intuitions I

Non-linear classification example: XOR/XNOR

→ x_1, x_2 are binary (0 or 1)



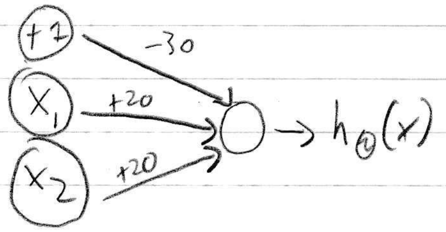
$y = x_1 \text{ XNOR } x_2$

$= \text{NOT}(x_1 \text{ XOR } x_2)$

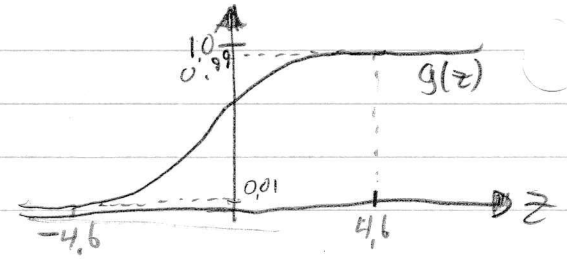
Simple example: AND

→ $x_1, x_2 \in \{0, 1\}$

→ $y = x_1 \text{ AND } x_2$



Recall:



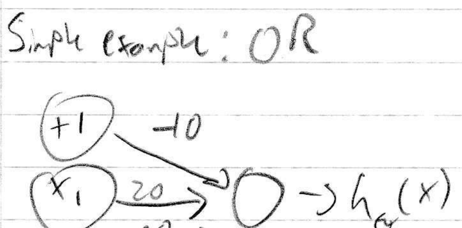
$$h_\Theta(x) = g(-30 + 20x_1 + 20x_2)$$

$\Theta^{(1)}_{10}$ $\Theta^{(1)}_{11}$ $\Theta^{(1)}_{12}$

x_1	x_2	$h_\Theta(x)$
0	0	$g(-30) \approx 0$
0	1	$g(-10) \approx 0$
1	0	$g(-10) \approx 0$
1	1	$g(10) \approx 1$

$\therefore h_\Theta(x) \approx x_1 \text{ AND } x_2$

Simple example: OR



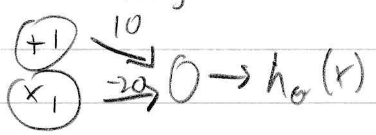
$$g(-10 + 20x_1 + 20x_2)$$

x_1	x_2	$h_\Theta(x)$
0	0	$g(-10) \approx 0$
0	1	$g(10) \approx 1$
1	0	$g(10) \approx 1$
1	1	$g(30) \approx 1$

$\therefore h_\Theta(x) \approx x_1 \text{ OR } x_2$

L2 - Examples and Intuitions 2

Example: Negation: NOT x_1

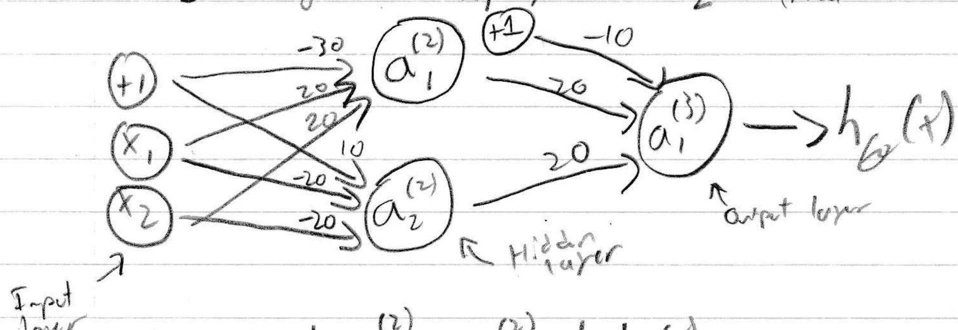


$$h_\theta(x) = g(10 - 20x_1)$$

x_1	$h_\theta(x)$
0	$g(10) \approx 1$
1	$g(-10) \approx 0$

$\therefore h_\theta(x) \approx$ NOT x_1

Putting it together: x_1 XNOR x_2 (recall from lecture 1)



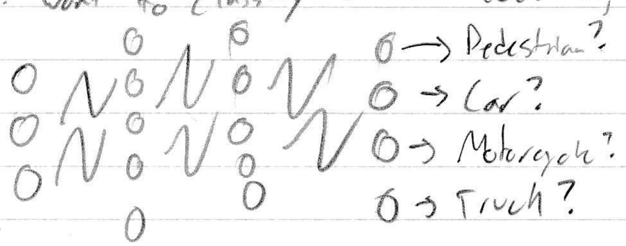
x_1	x_2	$a_1^{(2)}$	$a_2^{(2)}$	$h_\theta(x)$
0	0	0	1	1
0	1	0	0	0
1	0	0	0	0
1	1	1	0	1

$\therefore h_\theta(x) = x_1$ XNOR x_2

L3 - Multi-class classification

Multiple output units: One-vs-all

Ex. want to classify either: Pedestrian, Car, Motorcycle or Truck



$h_\theta(x) \in \mathbb{R}^4$

want $h_\theta(x) \approx \begin{bmatrix}1\\0\\0\\0\end{bmatrix}$, $h_\theta(x) \approx \begin{bmatrix}0\\1\\0\\0\end{bmatrix}$, $h_\theta(x) \approx \begin{bmatrix}0\\0\\1\\0\end{bmatrix}$, etc.

when pedestrian, when car, when motorcycle

Training set: $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$, $y^{(i)}$ is one of $\begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\1\end{bmatrix}$