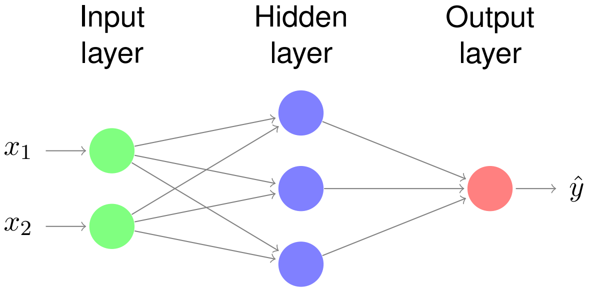
UCL Deep Learning Reading Group

Chapter 6 (part 2): Backpropagation

João M. Monteiro

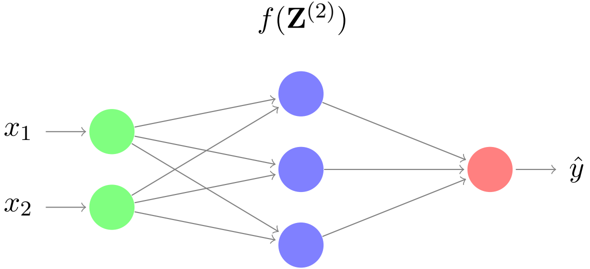
23/06/2016

Forward propagation



Forward propagation

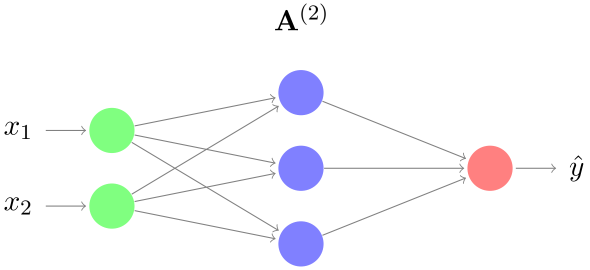
$\mathbf{X} \in \mathbb{R}^{N \times 2}$
$\mathbf{W}^{(1)} \in \mathbb{R}^{2 \times 3}$
$\mathbf{Z}^{(2)} = \mathbf{X}\mathbf{W}^{(1)} \quad , \quad \mathbf{Z}^{(2)} \in \mathbb{R}^{N \times 3}$



$$\mathbf{A}^{(2)} = f(\mathbf{Z}^{(2)})$$

Forward propagation

$\mathbf{A}^{(2)} \in \mathbb{R}^{N\times 3}$
$\mathbf{W}^{(2)} \in \mathbb{R}^{3\times 1}$
$\mathbf{Z}^{(3)} = \mathbf{A}^{(2)}\mathbf{W}^{(2)} \quad , \quad \mathbf{Z}^{(3)} \in \mathbb{R}^{N\times 1}$



$$f(\mathbf{Z}^{(3)}) = \mathbf{A}^{(3)} = \hat{y}$$

Backpropagation

$$\frac{\partial J}{\partial \mathbf{W}^{(1)}} \in \mathbb{R}^{2\times 3}$$

$$\frac{\partial J}{\partial \mathbf{W}^{(2)}} \in \mathbb{R}^{3\times 1}$$

The chain rule:

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y} \cdot \frac{\mathrm{d}y}{\mathrm{d}x}$$

Backpropagation

$$\frac{\partial J}{\partial \mathbf{W}^{(2)}}$$

$$\frac{\partial J}{\partial \mathbf{W}^{(2)}} = \frac{\partial \frac{1}{2}(y-\hat{y})^2}{\partial \mathbf{W}^{(2)}} = (y-\hat{y})\frac{\partial (y-\hat{y})}{\partial \mathbf{W}^{(2)}}$$

Note that: $\hat{y} = f(\mathbf{Z}^{(3)})$

$$(y-\hat{y})\frac{\partial (y-\hat{y})}{\partial \mathbf{W}^{(2)}} = -(y-\hat{y})\frac{\partial f(\mathbf{Z}^{(3)})}{\partial \mathbf{W}^{(2)}} = -(y-\hat{y})\frac{\partial f(\mathbf{Z}^{(3)})}{\partial \mathbf{Z}^{(3)}}\frac{\partial \mathbf{Z}^{(3)}}{\partial \mathbf{W}^{(2)}}$$

$$-(y-\hat{y})f^{'}(\mathbf{Z}^{(3)})\frac{\partial \mathbf{Z}^{(3)}}{\partial \mathbf{W}^{(2)}} = \boldsymbol{\delta}^{(3)}\frac{\partial \mathbf{Z}^{(3)}}{\partial \mathbf{W}^{(2)}} = \boldsymbol{\delta}^{(3)}\mathbf{A}^{(2)}$$

Backpropagation

$$\frac{\partial J}{\partial \mathbf{W}^{(1)}}$$

Same as before until: $\frac{\partial J}{\partial \mathbf{W}^{(1)}} = -(y - \hat{y}) f'(\mathbf{Z}^{(3)}) \frac{\partial \mathbf{Z}^{(3)}}{\partial \mathbf{W}^{(1)}}$

$$\frac{\partial J}{\partial \mathbf{W}^{(1)}} = \boldsymbol{\delta}^{(3)} \frac{\partial \mathbf{Z}^{(3)}}{\partial \mathbf{W}^{(1)}} = \boldsymbol{\delta}^{(3)} \frac{\partial \mathbf{Z}^{(3)}}{\partial \mathbf{A}^{(2)}} \frac{\partial \mathbf{A}^{(2)}}{\partial \mathbf{W}^{(1)}} = \boldsymbol{\delta}^{(3)} \frac{\partial \mathbf{Z}^{(3)}}{\partial \mathbf{A}^{(2)}} \frac{\partial \mathbf{A}^{(2)}}{\partial \mathbf{Z}^{(2)}} \frac{\partial \mathbf{Z}^{(2)}}{\partial \mathbf{W}^{(1)}}$$

$$\frac{\partial J}{\partial \mathbf{W}^{(1)}} = \boldsymbol{\delta}^{(3)} (\mathbf{W}^{(2)})^\intercal f'(\mathbf{Z}^{(2)}) \mathbf{X}$$

References

[1] Neural networks demystified. `http://lumiverse.io/series/neural-networks-demystified`. Accessed: 2016-06-19.

[2] Ian Goodfellow Yoshua Bengio and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2016.