第三讲
全连接网络
Lecture 3
Fully-connected Neural Networks

明玉瑞 Yurui Ming

yrming@gmail.com

声明
Disclaimer

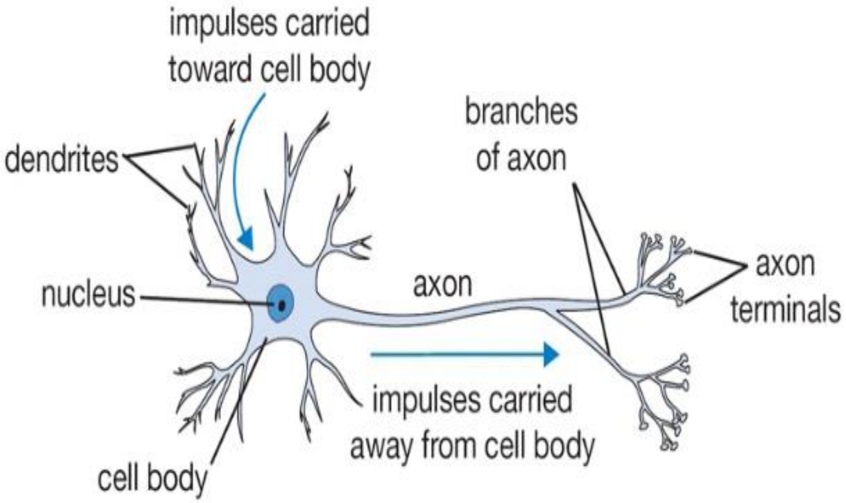
- 本讲义在准备过程中由于时间所限，所用材料来源并未规范标示引用来源。所引材料仅用于教学所用，作者无意侵犯原著者之知识版权，所引材料之知识版权均归原著者所有；若原著者介意之，请联系作者更正及删除。

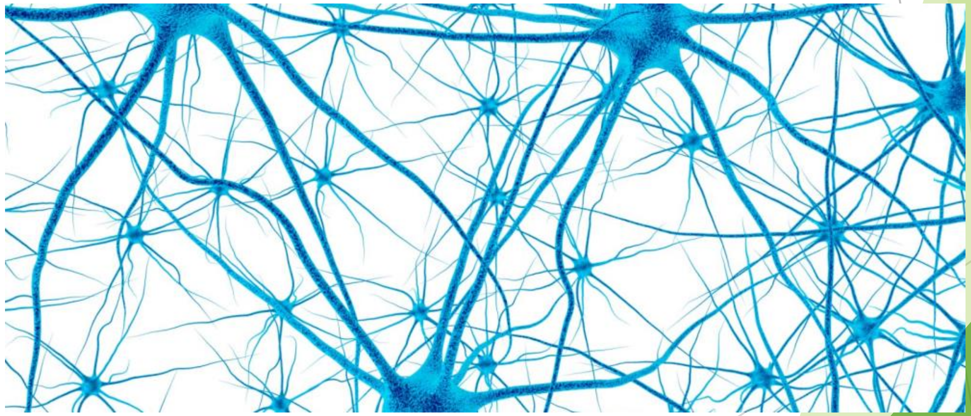
  The time limit during the preparation of these slides incurs the situation that not all the sources of the used materials (texts or images) are properly referenced or clearly manifested. However, all materials in these slides are solely for teaching and the author is with no intention to infringe the copyright bestowed on the original authors or manufacturers. All credits go to corresponding IP holders. Please address the author for any concern for remedy including deletion.

脑科学启发
Brain inspiration

- 全连接网络基于神经元处理信息的方式及其组网方式

  Fully-connected neural networks are inspired by the way neurons work and their connections into network


impulses carried toward cell body
dendrites
branches of axon
nucleus
axon
axon terminals
impulses carried away from cell body
cell body



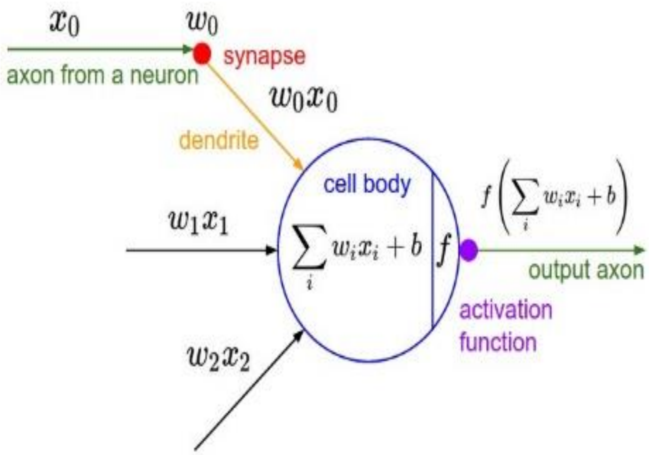
数学模型
Mathematical Modelling

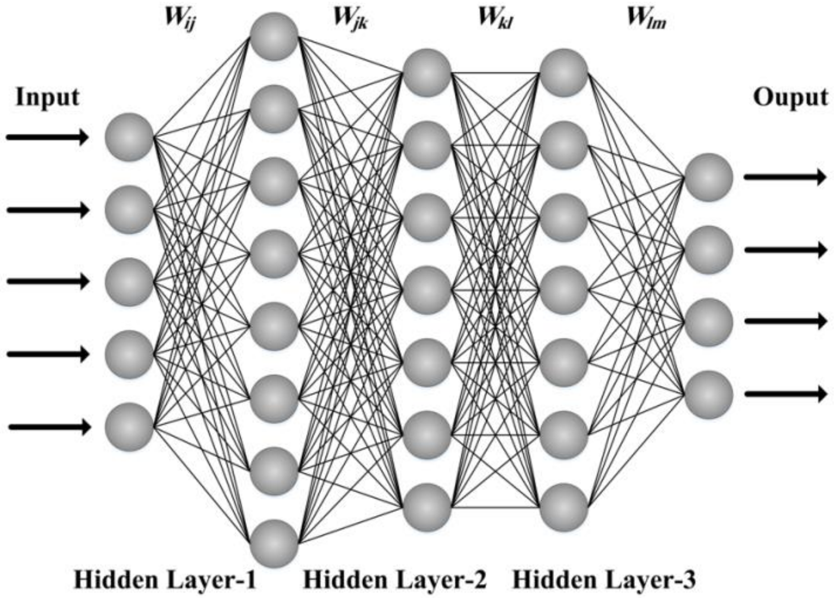
- 对神经元及其构成网络的数学建模

  Mathematical modelling of the neurons and corresponding networks

- 为建模方便，引入了额外的术语，如输入层，隐含层，输出层等

  Terminologies are introduced for modelling, such as input layer, hidden layer, output layer, etc.





记号
Denotation

- 第l层的第i个神经元输出记为$y_{l,i}$

  The output of i-th neuron at level l is denoted by $y_{l,i}$

- 第l层的第i个神经元到第$l+1$层的第j个神经元记为$w_{l,ji}$

  The weight from the i-th neuron at level l to the j-th neuron at level $l+1$ is denoted by $w_{l,ji}$

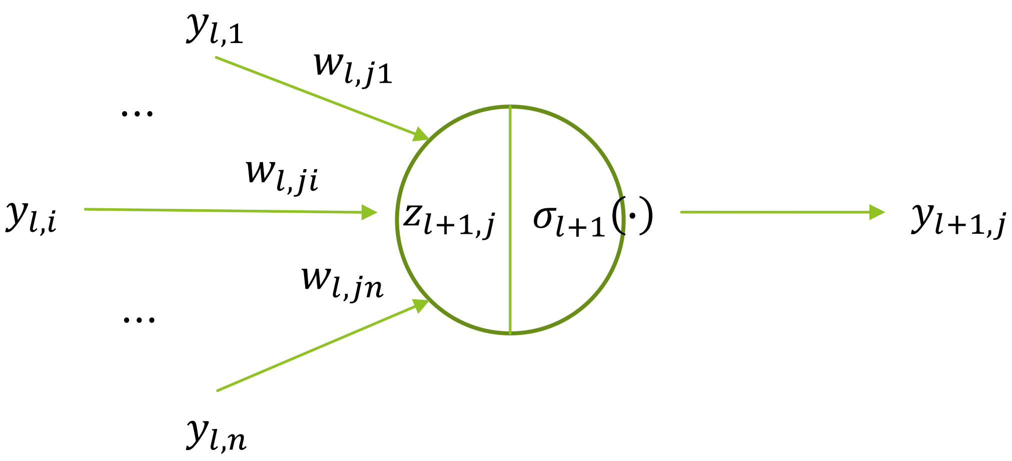
- 由第l层的神经元到第$l+1$层的第j个神经元的加权和记为$z_{l+1,j}$

  The intermediate result for the j-th neuron at level $l+1$ is denoted by $z_{l+1,j}$

记号
Denotation

- 用于建模的数学式子展示

  Illustration of mathematical modelling formulas



$$z_{l+1,j} = \sum_{i=1}^{n} y_{l,i} \cdot w_{l,ji}$$

$$y_{l+1,j} = \sigma_{l+1}\left(z_{l+1,j}\right) = \frac{1}{1 + e^{-z_{l+1,j}}}$$

(sigmoid function)

前向传播
feed-forward propagation

- 可以逐节点算

  Node-based computation

$$y_6 = x_6 \qquad\qquad y_7 = x_7$$

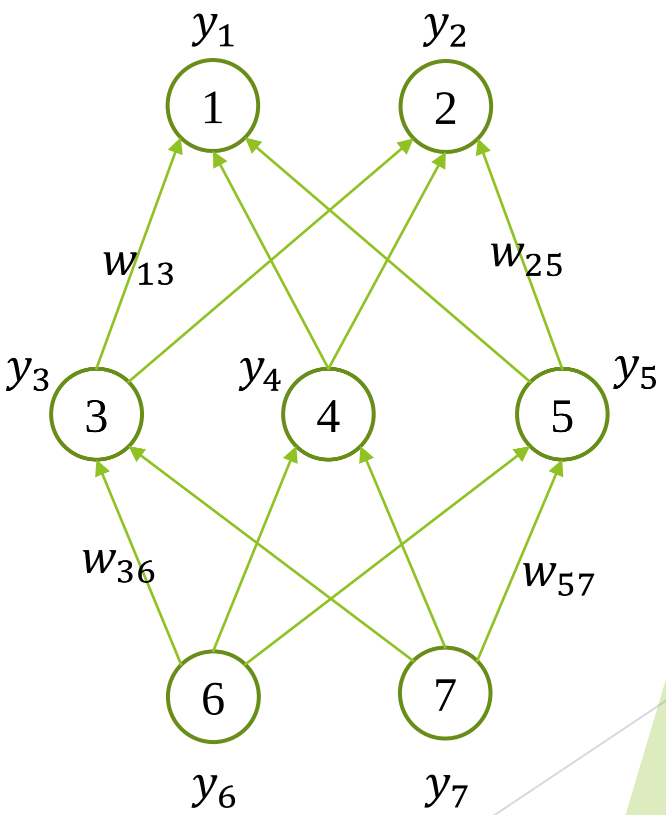
$$y_3 = \sigma(w_{36} \cdot y_6 + w_{37} \cdot y_7)$$

$$y_4 = \sigma(w_{46} \cdot y_6 + w_{47} \cdot y_7)$$

$$y_5 = \sigma(w_{56} \cdot y_6 + w_{57} \cdot y_7)$$

$$y_1 = \sigma(w_{13} \cdot y_3 + w_{14} \cdot y_4 + w_{15} \cdot y_5)$$

$$y_2 = \sigma(w_{23} \cdot y_3 + w_{24} \cdot y_4 + w_{25} \cdot y_5)$$



前向传播
feed-forward propagation

- 实践中会是矩阵的形式

  Matrix form in practice

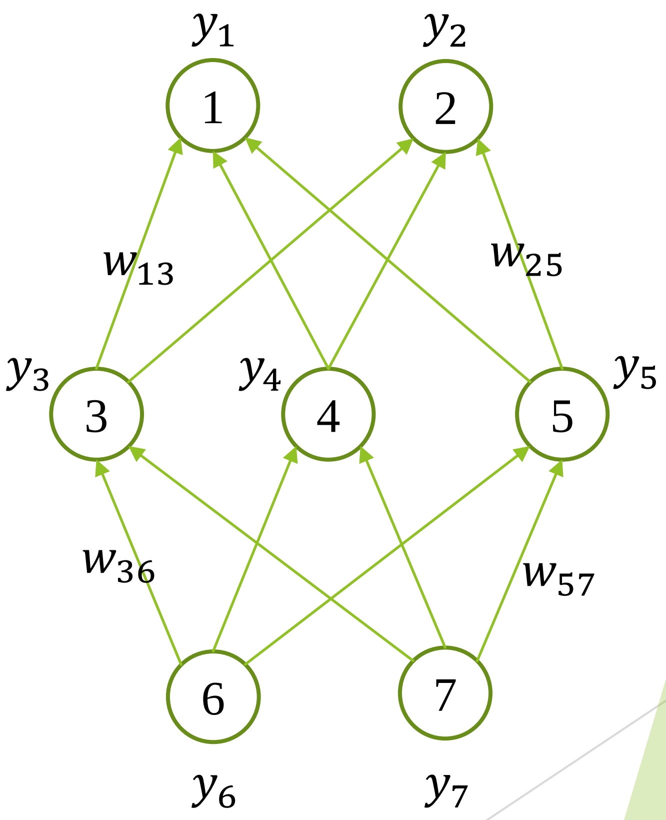
$$y_3 = \sigma(w_{36} \cdot y_6 + w_{37} \cdot y_7)$$

$$y_4 = \sigma(w_{46} \cdot y_6 + w_{47} \cdot y_7)$$

$$y_5 = \sigma(w_{56} \cdot y_6 + w_{57} \cdot y_7)$$

$$\begin{bmatrix} y_3 \\ y_4 \\ y_5 \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{36} & w_{37} \\ w_{37} & w_{37} \\ w_{37} & w_{37} \end{bmatrix} \begin{bmatrix} y_6 \\ y_7 \end{bmatrix}\right)$$

$$\vec{y}_{l+1} = \sigma(W\vec{y}_l)$$



前向传播
feed-forward propagation

- 实践中会是矩阵的形式

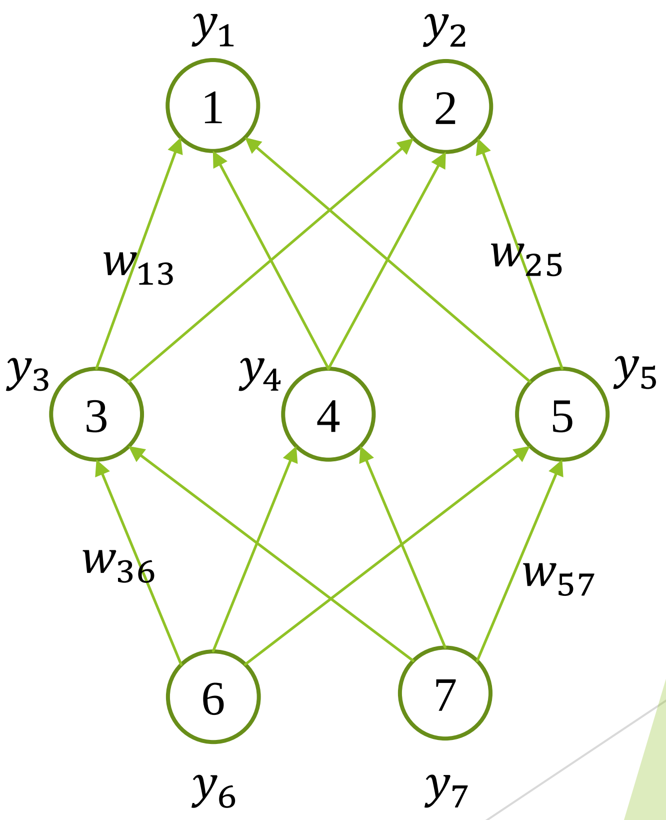
  Matrix form in practice

$$y_1 = \sigma(w_{13} \cdot y_3 + w_{14} \cdot y_4 + w_{15} \cdot y_5)$$

$$y_2 = \sigma(w_{23} \cdot y_3 + w_{24} \cdot y_4 + w_{25} \cdot y_5)$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{13} & w_{14} & w_{15} \\ w_{23} & w_{24} & w_{25} \end{bmatrix} \begin{bmatrix} y_3 \\ y_4 \\ y_5 \end{bmatrix} \right)$$

$$\vec{y}_{l+1} = \sigma(W\vec{y}_l)$$



前向传播
feed-forward propagation

- 进一步考虑阈限值

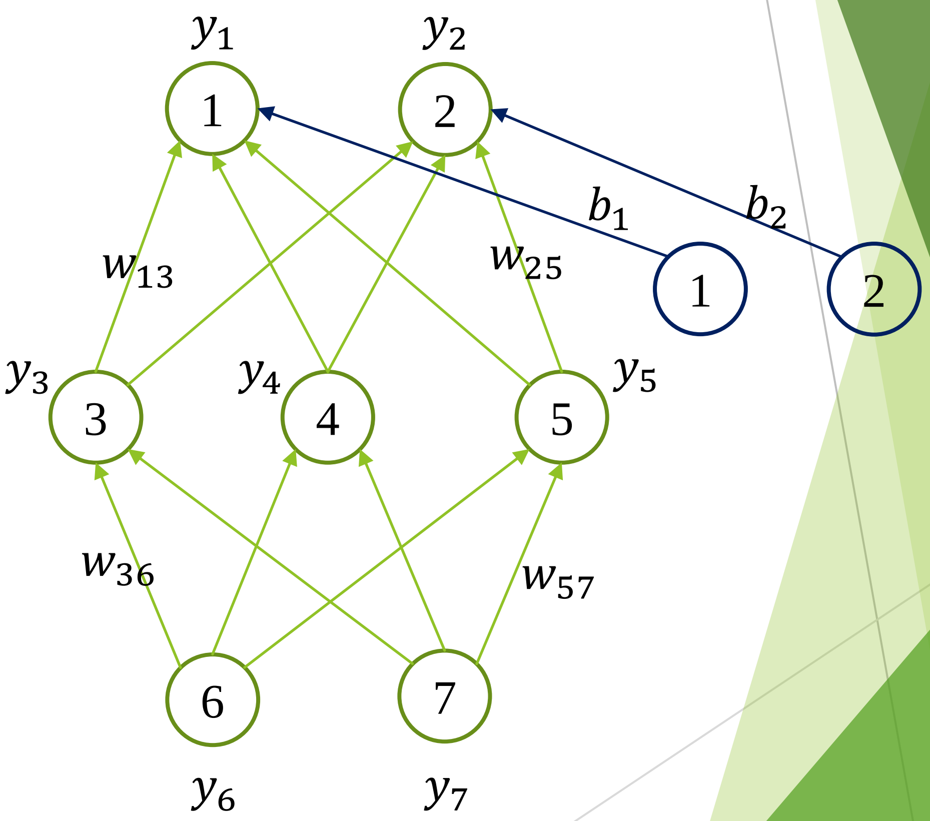
  Put bias into consideration

$$y_1 = \sigma(w_{13} \cdot y_3 + w_{14} \cdot y_4 + w_{15} \cdot y_5 + b_1)$$

$$y_2 = \sigma(w_{23} \cdot y_3 + w_{24} \cdot y_4 + w_{25} \cdot y_5 + b_2)$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{13} & w_{14} & w_{15} \\ w_{23} & w_{24} & w_{25} \end{bmatrix}\begin{bmatrix} y_3 \\ y_4 \\ y_5 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\right)$$

$$\vec{y}_{l+1} = \sigma(W\vec{y}_l + \vec{b})$$



前向传播
feed-forward propagation

- 进一步考虑批量的情况

  Put batch into consideration

$$y_1^1 = \sigma\left(w_{13} \cdot y_3^1 + w_{14} \cdot y_4^1 + w_{15} \cdot y_5^1 + b_1\right)$$
$$y_1^2 = \sigma\left(w_{13} \cdot y_3^2 + w_{14} \cdot y_4^2 + w_{15} \cdot y_5^2 + b_1\right)$$
$$\dots$$
$$y_1^N = \sigma\left(w_{13} \cdot y_3^N + w_{14} \cdot y_4^N + w_{15} \cdot y_5^N + b_1\right)$$

$$y_2^1 = \sigma\left(w_{23} \cdot y_3^1 + w_{24} \cdot y_4^1 + w_{25} \cdot y_5^1 + b_2\right)$$
$$y_2^2 = \sigma\left(w_{23} \cdot y_3^2 + w_{24} \cdot y_4^2 + w_{25} \cdot y_5^2 + b_2\right)$$
$$\dots$$
$$y_2^N = \sigma\left(w_{23} \cdot y_3^N + w_{24} \cdot y_4^N + w_{25} \cdot y_5^N + b_2\right)$$

$$\begin{bmatrix} y_1^1 & y_1^2 & y_1^3 & y_1^4 \\ y_2^1 & y_2^2 & y_2^3 & y_2^4 \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{13} & w_{14} & w_{15} \\ w_{23} & w_{24} & w_{25} \end{bmatrix}\begin{bmatrix} y_3^1 & y_3^2 & y_3^3 & y_3^4 \\ y_4^1 & y_4^2 & y_4^3 & y_4^4 \\ y_5^1 & y_5^2 & y_5^3 & y_5^4 \end{bmatrix} + \begin{bmatrix} b_1 & b_1 & b_1 & b_1 \\ b_2 & b_2 & b_2 & b_2 \end{bmatrix}\right)$$

$$Y_{l+1} = \sigma(WY_l + B)$$

前向传播
feed-forward propagation

- 工程习惯中，批量大小用N表示，作为第一个维度；特征列大小用C表示，作为第二个维度，因此需要转置一下

  From the engineering convention perspective, the batch size is denoted by N as the first dimension, while the size of feature column is denoted by C as the second dimension, so the formula deducted in the last slide needs to be transposed.

$$\begin{bmatrix} y_1^1 & y_2^2 & y_2^3 & y_2^4 \\ y_2^1 & y_2^2 & y_2^3 & y_2^4 \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{13} & w_{14} & w_{15} \\ w_{23} & w_{24} & w_{25} \end{bmatrix}\begin{bmatrix} y_3^1 & y_3^2 & y_3^3 & y_3^4 \\ y_4^1 & y_4^2 & y_4^3 & y_4^4 \\ y_5^1 & y_5^2 & y_5^3 & y_5^4 \end{bmatrix} + \begin{bmatrix} b_1 & b_1 & b_1 & b_1 \\ b_2 & b_2 & b_2 & b_2 \end{bmatrix}\right)$$

$$\begin{bmatrix} y_1^1 & y_2^1 \\ y_1^2 & y_2^2 \\ y_1^3 & y_2^3 \\ y_1^4 & y_2^4 \end{bmatrix} = \sigma\left(\begin{bmatrix} y_3^1 & y_4^1 & y_5^1 \\ y_3^2 & y_4^2 & y_5^2 \\ y_3^3 & y_4^3 & y_5^3 \\ y_3^4 & y_4^4 & y_5^4 \end{bmatrix}\begin{bmatrix} w_{13} & w_{23} \\ w_{14} & w_{24} \\ w_{15} & w_{25} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 \\ b_1 & b_2 \\ b_1 & b_2 \\ b_1 & b_2 \end{bmatrix}\right)$$

$$Y_{l+1} = \sigma(Y_l W + B)$$

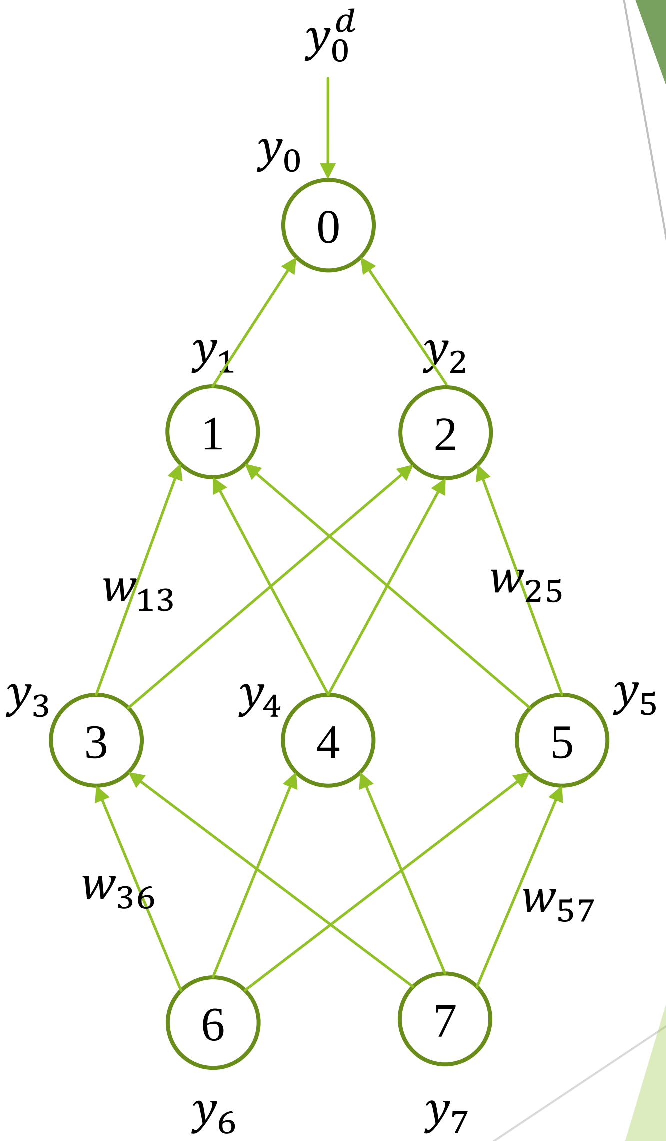
前向传播
feed-forward propagation

- 计算损失，这里以回归问题为例

  Loss computation，exemplified by regression problem

$$E = \frac{1}{2}\left(y_0 - y_0^d\right)^2$$

$$E = \frac{1}{2N}\sum_{i=1}^{N}\left(y_0^i - y_0^d\right)^2$$



反向传播算法
Back-propagation algorithm

- 反向传播算法又称错误传播或delta学习法则

  An alternative for back propagation is called error propagation or delta learning rule

- 为方便推导，做下面一些假设

  The following assumptions are made to ease the deduction

  - 若层不存在歧义，则$w_{l,ji}$简记为w_{ji}

    If no ambiguity from the layer perspective, $w_{l,ji}$ is denoted by w_{ji}

  - 对于输出层的第i个神经元，其对应的真实标签记为y_i^d

    For output from the i-th neurons at the final layer L, the ground-truth is denoted by $y_{L,i}^d$, or y_i^d

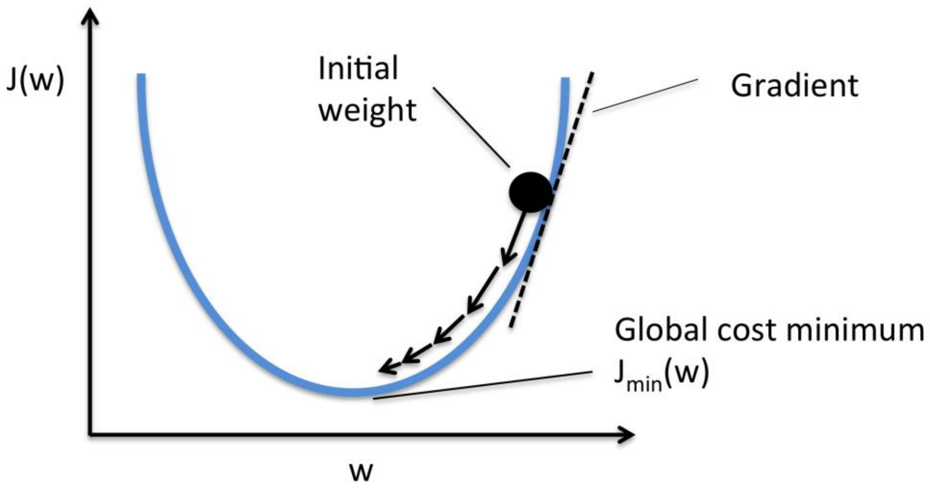
反向传播算法
Back-propagation algorithm

- 反向传播算法基本思想基于梯度下降

  Back-propagation algorithm is based on gradient descent algorithm

- 反向传播算法的实施依赖于求导的链式法则

  Implementation of back-propagation algorithm relies on the chain rules of derivatives



$$E = \frac{1}{2}\left(y_i - y_i^d\right)^2$$

$$w_{ij}(t+1) = w_{ij}(t) + \eta\left(-\frac{\partial E}{\partial w_{ij}}\right)$$

$$\frac{\partial E}{\partial w_{ij}} = \left(y_i - y_i^d\right)\frac{\partial y_i}{\partial w_{ij}} = \left(y_i - y_i^d\right)\frac{\partial y_i}{\partial z_i}\frac{\partial z_i}{\partial w_{ij}} = \left(y_i - y_i^d\right)y_i' y_j$$

$$w_{ij}(t+1) = w_{ij}(t) - \eta\left(y_i - y_i^d\right)y_i' y_j$$

反向传播算法
Back-propagation algorithm

- 简单起见，我们顺序标识神经元节点进行推导

  For simplicity, neurons are labelled sequentially

$$E = \frac{1}{2}\left(y_1 - y^d\right)^2 \qquad \Rightarrow \frac{\partial E}{\partial y_1} = \left(y_1 - y^d\right)$$

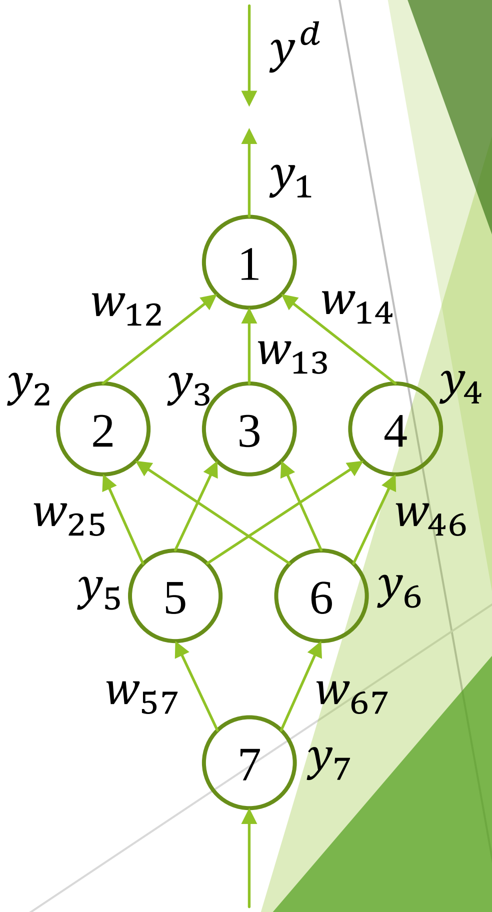
$$\frac{\partial E}{\partial w_{12}} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial w_{12}} = \left(y_1 - y^d\right)\frac{\partial y_1}{\partial z_1}\frac{\partial z_1}{\partial w_{12}} = \left(y_i - y_i^d\right)y_1' y_2$$

$$\delta_0 \triangleq \left(y_1 - y^d\right) \qquad \Rightarrow \frac{\partial E}{\partial w_{12}} = \delta_0 y_1' y_2$$

$$\delta_1 \triangleq \delta_0 y_1' \qquad \Rightarrow w_{12}(t+1) = w_{12}(t) - \eta\delta_1 y_2$$

$$w_{13}(t+1) = w_{13}(t) - \eta\delta_1 y_3$$

$$w_{14}(t+1) = w_{14}(t) - \eta\delta_1 y_4$$



反向传播算法
Back-propagation algorithm

$$\frac{\partial E}{\partial w_{25}} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial w_{25}} = (y_1 - y^d)\frac{\partial y_1}{\partial z_1}\frac{\partial z_1}{\partial w_{25}} = (y_1 - y_1^d)y_1' \times$$

$$\frac{\partial}{\partial w_{25}}[w_{12}y_2 + w_{13}y_3 + w_{14}y_4] = \delta_1 w_{12}\frac{\partial y_2}{\partial w_{25}}$$

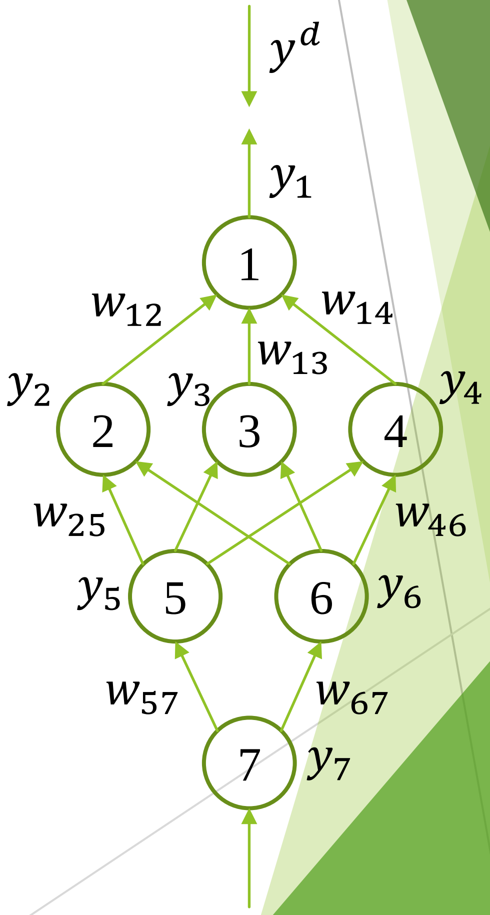
$$= \delta_1 w_{12}\frac{\partial y_2}{\partial z_2}\frac{\partial z_2}{\partial w_{25}} = \delta_1 w_{12} y_2' \frac{\partial z_2}{\partial w_{25}}$$

$$= \delta_1 w_{12}\frac{\partial y_2}{\partial z_2}\frac{\partial}{\partial w_{25}}[w_{25}y_5 + w_{26}y_6] = \delta_1 w_{12} y_2' y_5$$

$$\delta_2 \triangleq \delta_1 w_{12} y_2' \qquad \Rightarrow \frac{\partial E}{\partial w_{25}} = \delta_2 y_5$$

$$w_{25}(t+1) = w_{25}(t) - \eta\delta_2 y_5$$

$$w_{26}(t+1) = w_{26}(t) - \eta\delta_2 y_6$$



反向传播算法
Back-propagation algorithm

- Similarly, we have

$$\delta_3 \triangleq \delta_1 w_{13} y_3'$$

$$\delta_4 \triangleq \delta_1 w_{14} y_4'$$

$$\delta_5 \triangleq (\delta_2 w_{25} + \delta_3 w_{35} + \delta_4 w_{45}) y_5'$$

$$\delta_6 \triangleq (\delta_2 w_{26} + \delta_3 w_{36} + \delta_4 w_{46}) y_6'$$

$$w_{35}(t+1) = w_{35}(t) - \eta \delta_3 y_5$$

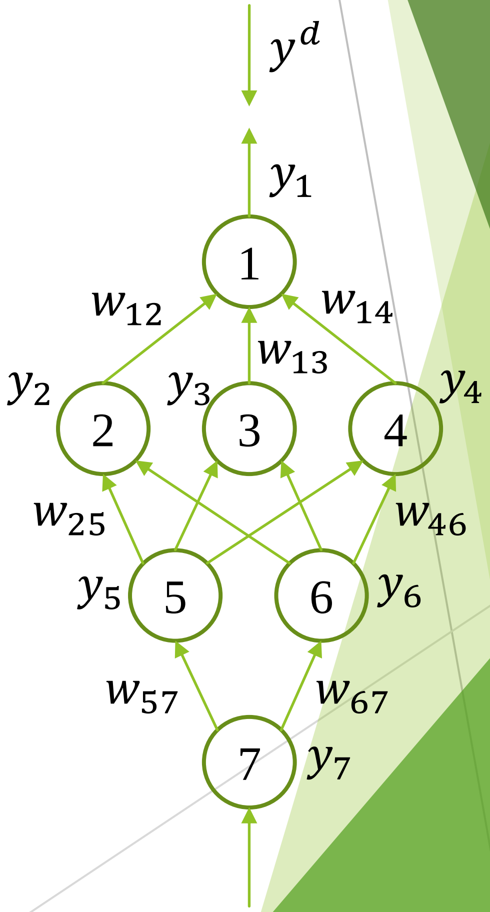
$$w_{36}(t+1) = w_{36}(t) - \eta \delta_3 y_6$$

$$w_{45}(t+1) = w_{45}(t) - \eta \delta_4 y_5$$

$$w_{46}(t+1) = w_{46}(t) - \eta \delta_4 y_6$$

$$w_{57}(t+1) = w_{57}(t) - \eta \delta_5 y_7$$

$$w_{67}(t+1) = w_{67}(t) - \eta \delta_6 y_7$$



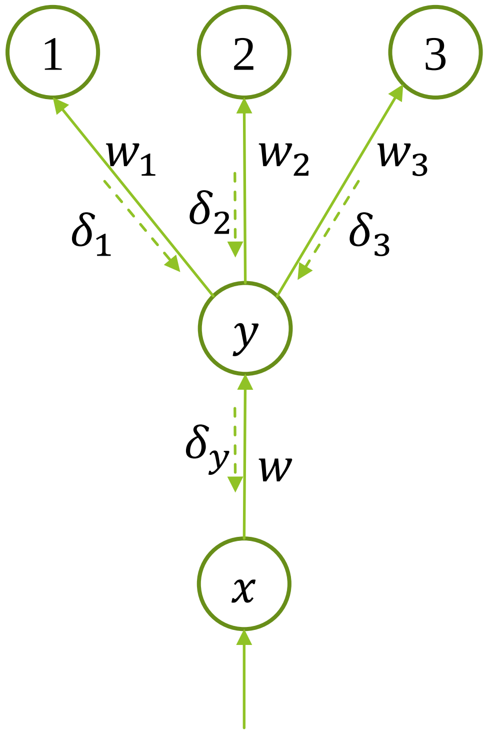
反向传播算法
Back-propagation algorithm

- 一般性规则

  General rule

  $$\delta_y \triangleq (\delta_1 w_1 + \delta_2 w_2 + \delta_3 w_3) y'$$

  $$w(t+1) = w(t) - \eta \delta_y x$$



实践过程
Practice

- 交替执行以下过程

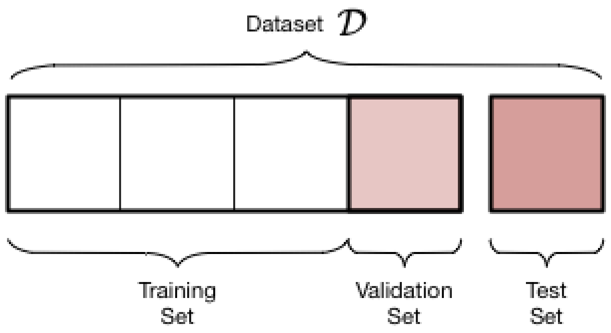
  Perform the following procedures in an iterative way:

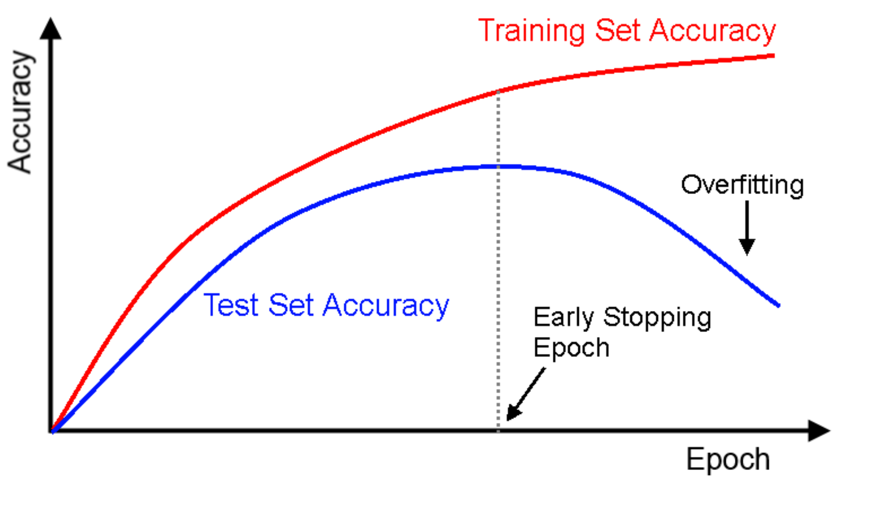
  - 调优网络配置，训练网络

    Decide the neural network configuration and train the neural network

  - 进行（交叉）验证，评估网络设计及性能

    (Cross-)validate to evaluate the design and performance





习题
Problems

1. 阅读相关文献，解释如何根据在训练过程中，根据模型在训练集与验证集上的性能来判断欠拟合与过拟合。

   Explain how to interpret the underfitting and overfitting phenomena during the training process by inspecting the loss and/or accuracy on training set and test set, by doing literature survey.

2. 请阅读文献，查阅异或（XOR）问题的历史，了解其对神经网络发展的影响，并设计一个全连接网络解决这个问题。

   By doing literature review to understand the XOR problems in the history of neural network evolvement, and its impact on the declining of neural networks. Design a multilayer fully-connected neural network to solve the XOR problem.

习题
Problems

3. 在Slide 4的公式中，σ称为激活函数

   σ is called activation function in the slide 4

   - 请举例说明你所知道的激活函数；

     Enumerate some examples of activation functions.

   - 请证明若激活函数为恒等变换（Identity），即$\sigma(x) = x$，则多层神经网络等价于线性变换；

     Proof that if identity function is adopted as activation function, then multi-perceptron network is equivalent to linear transformation.

   - 对于一些特定的激活函数，例如Sigmoid函数，在实际应用中可能会有梯度消失（Gradient vanishing）问题，请解释，并证明其最大梯度值为0.25；

     For some activation functions, such as sigmoid function, there could exist gradient vanishing problem in practice. Please explain it and prove that the maximal gradient is 0.25

习题
Problems

- 单调性是通常认为是激活函数的一个必要条件，请阅读最新文献，说明这个条件非必要，即非单调函数有可能在某些情况下表现更好；并尝试从神经科学的角度解释这件事。

  Monotone is attributed to the necessity for the candidature of activation functions. By doing literature survey to illustrate that non-monotonic function could have better performance, which dis-prove the proclamation. Find explanations from the neuroscientific perspective.

4. 偏差和方差是机器学习模型的两个重要属性。偏差是为简化模型优化过程而做出的假设，方差是目标函数可以输出的值的范围的度量。请说明如果偏差与方差过高，都有哪些危害。

   Bias and variance are two factors measuring the quality of ML models. Bias is introduced due to assumption for model simplification; while variance measures ranges of model outputs. By doing literature review to clarify the harmfulness of high bias and high variance.