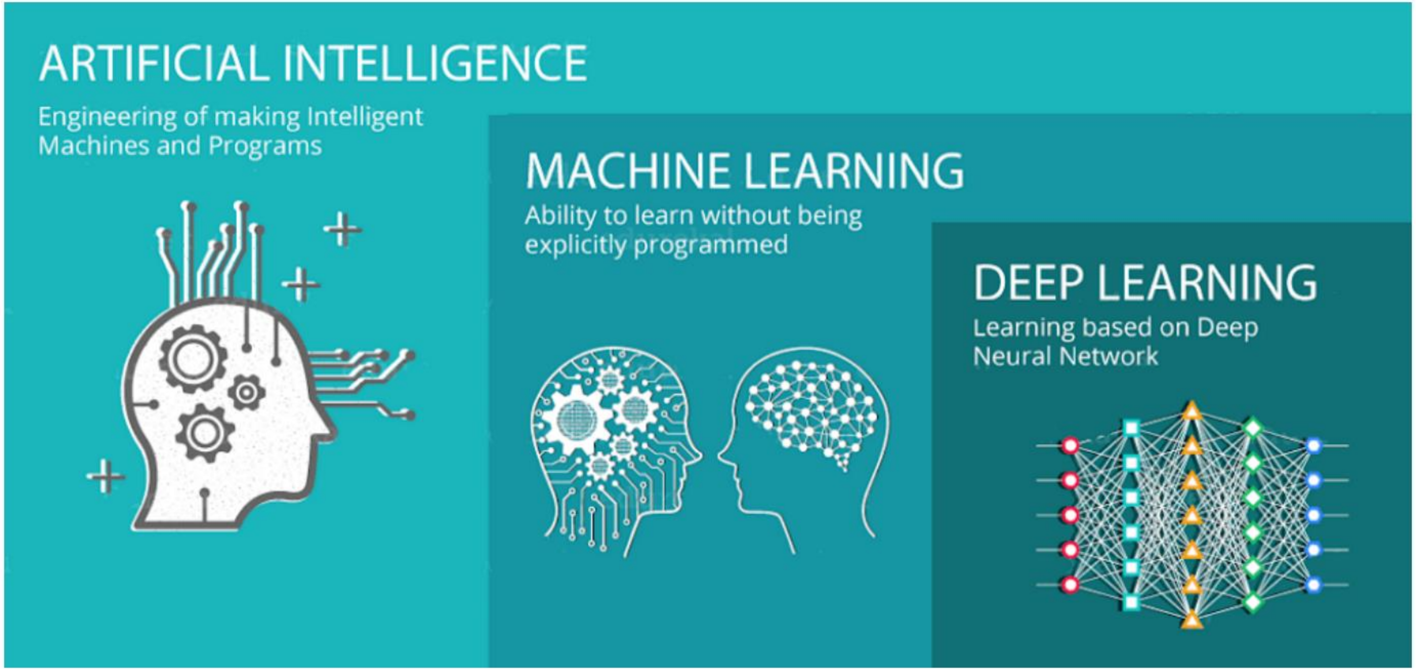
Artificial Neural Network

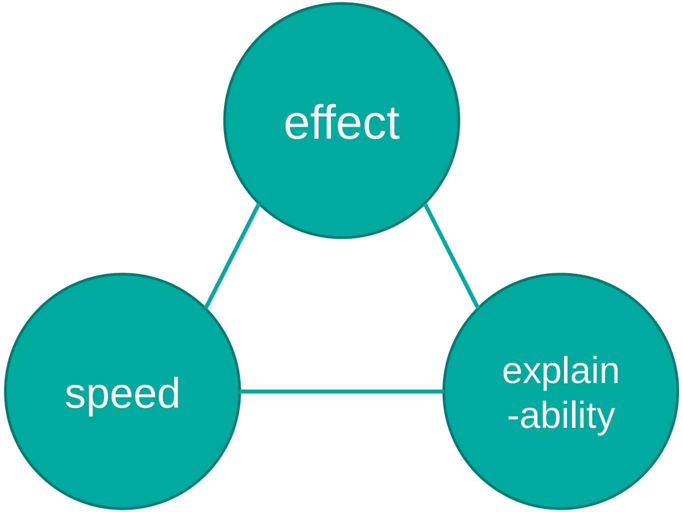
FMMS050 ML PhD course

Lecture 10

Position of Artificial Neural Network



Targets of Artificial Neural Network

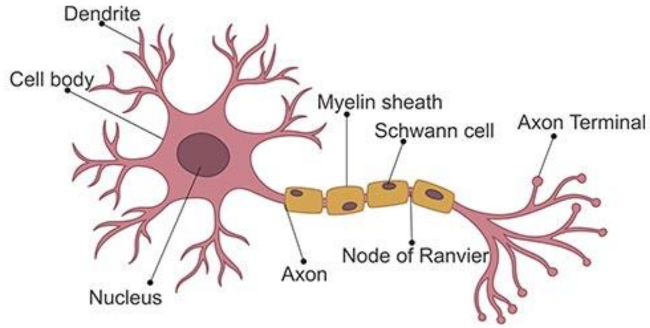


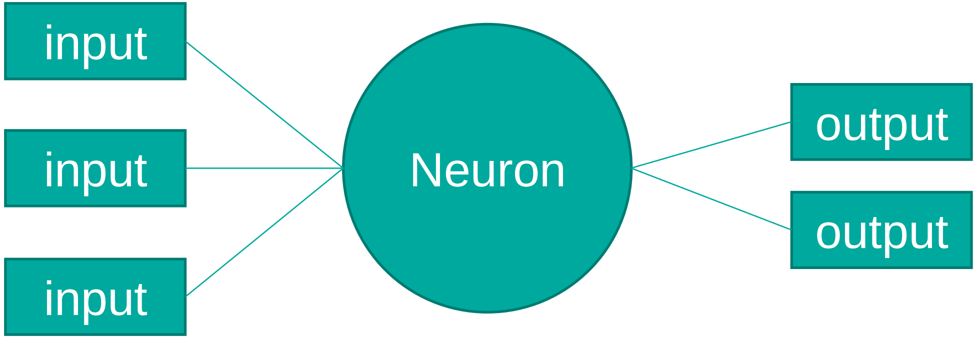
Artificial Neural Network (DL)

Artificial Neural Network (DL)

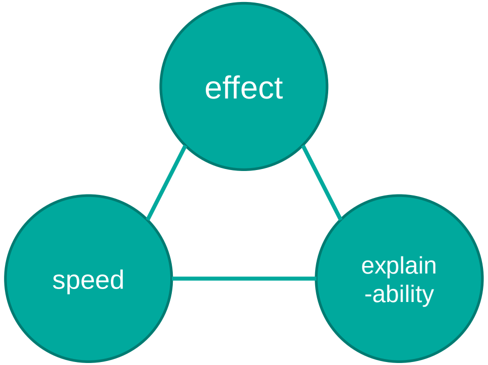
Artificial Neural Network (DL)

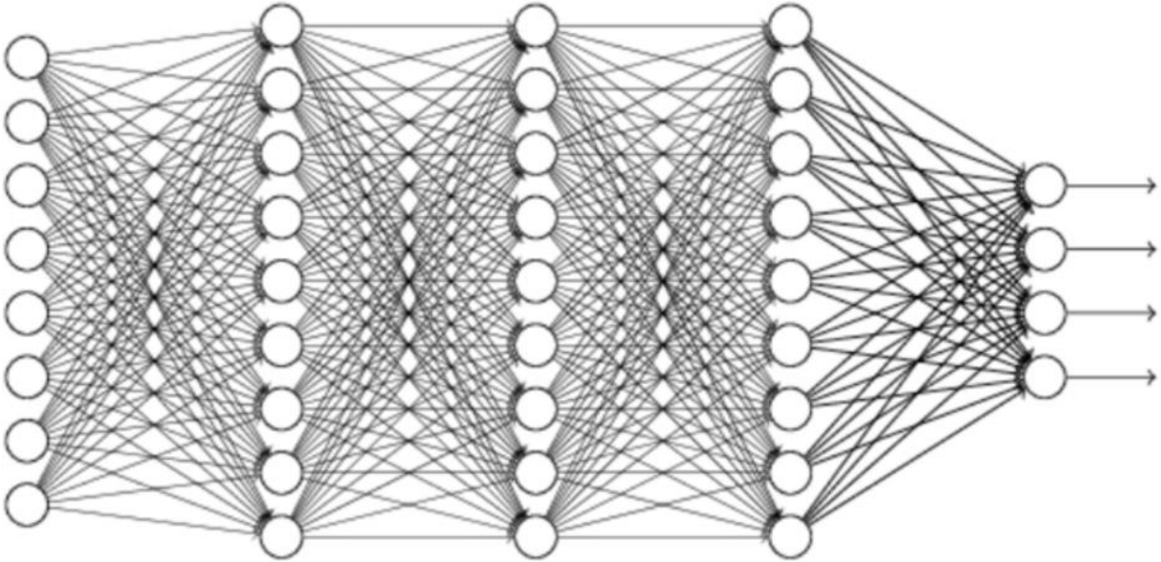
Artificial Neuron





Artificial Neural Network





Artificial Neural Network

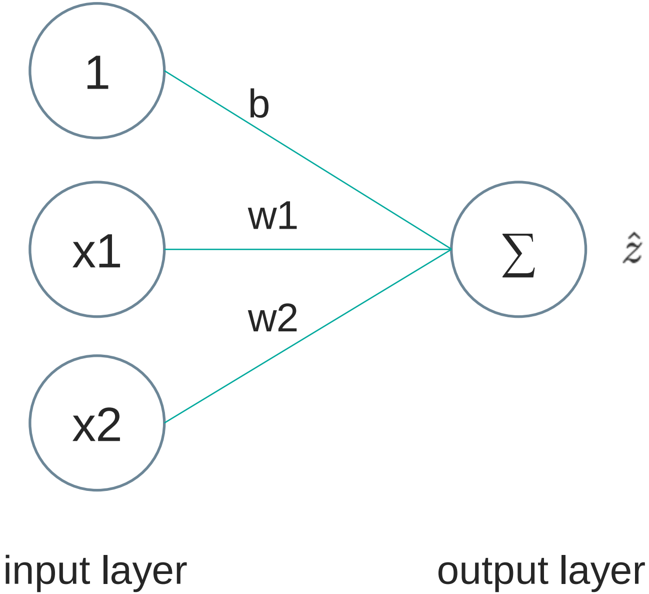
- Boltzmann machine (RBM)
- Multilayer Perceptron Classifier (MLPClassifier)
- Multilayer Perceptron Regressor (MLPRegressor)

Single layer neural network (regression)

$$\hat{z} = \underbrace{b}_{\text{(intercept,bias)}} + x_1 w_1 + x_2 w_2$$

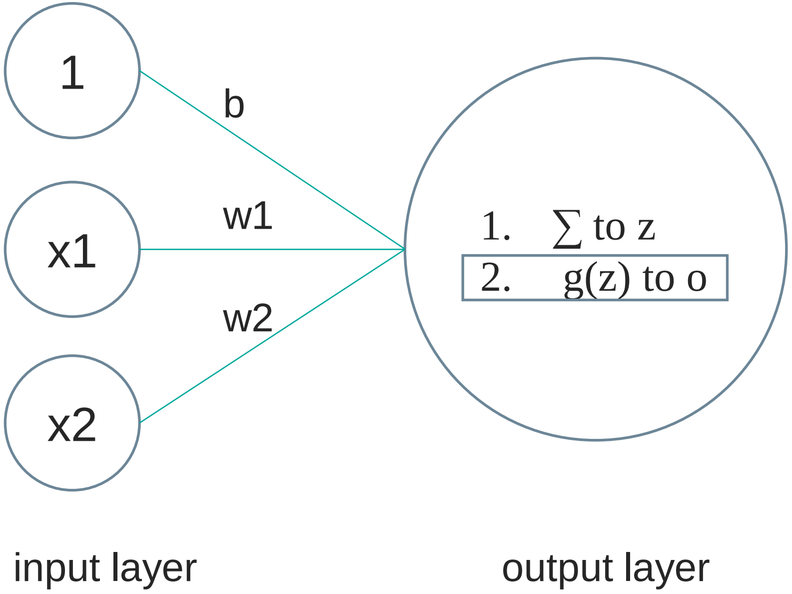
(intercept,bias) (weights)

Single layer neural network



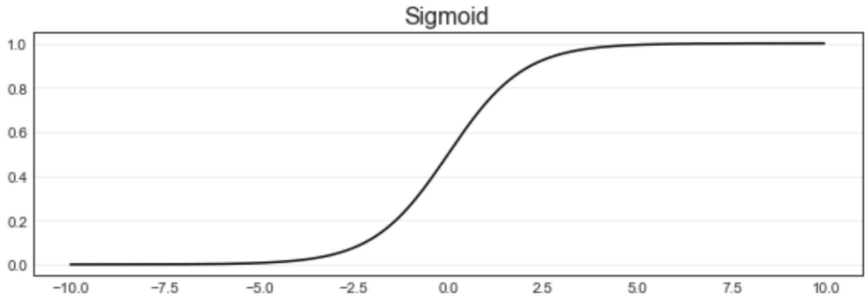
$$\hat{z} = b + x_1 w_1 + x_2 w_2$$

Single layer neural network (binary classification)

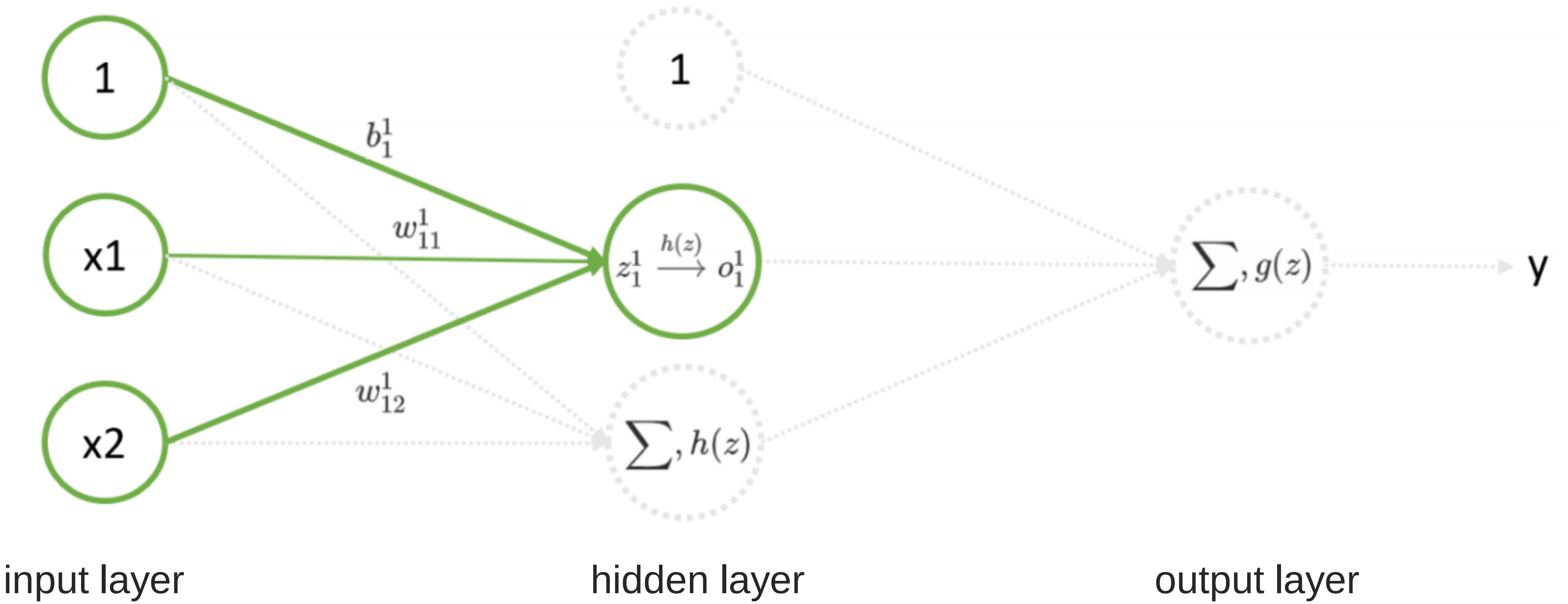


1. Summation: $\hat{z} = b + x_1 w_1 + x_2 w_2$

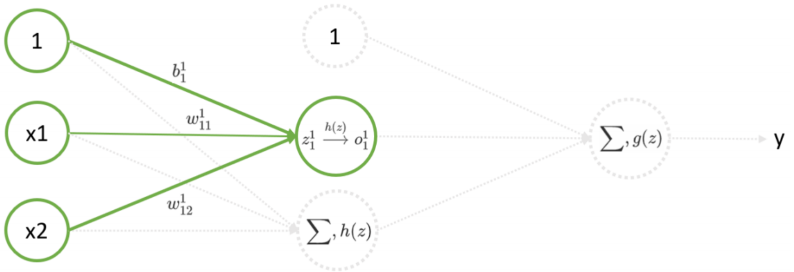
2. Activation: $g(z) = \dfrac{1}{1 + e^{-z}}$



Multi-layer neural network



Multi-layer neural network



$$z_1^1 = 1 * b_1^1 + x_1 w_{11}^1 + x_2 w_{12}^1$$

$$o_1^1 = h(b_1^1 + x_1 w_{11}^1 + x_2 w_{12}^1)$$

layer #

z_1^1 o_1^1

neuron #

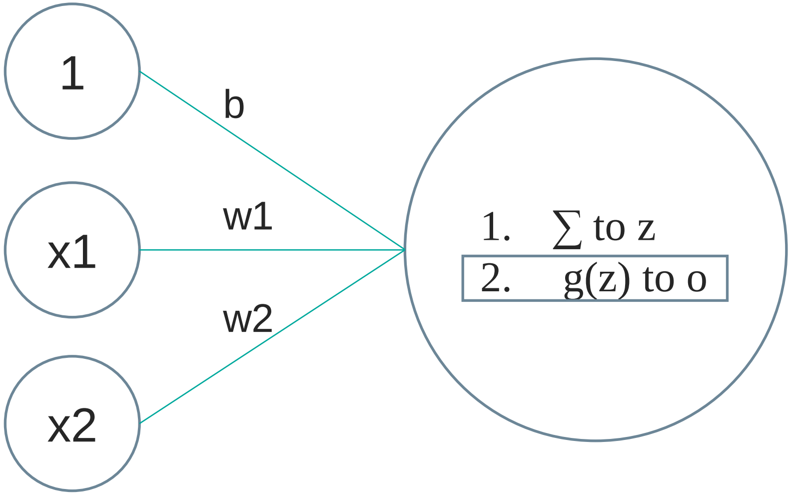
next layer #

w_{12}^1 b_1^1

next layer neuron #

pre layer neuron #

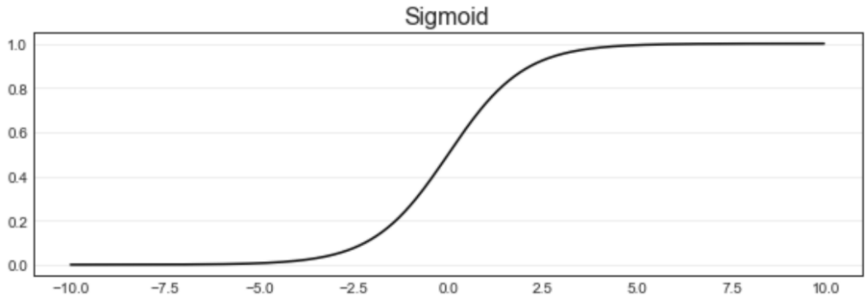
Single layer neural network (binary classification)



input layer

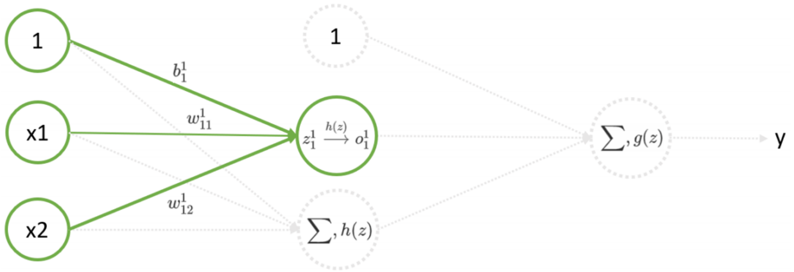
output layer

1. Summation: $\hat{z} = b + x_1 w_1 + x_2 w_2$
2. Activation: $g(z) = \frac{1}{1 + e^{-z}}$



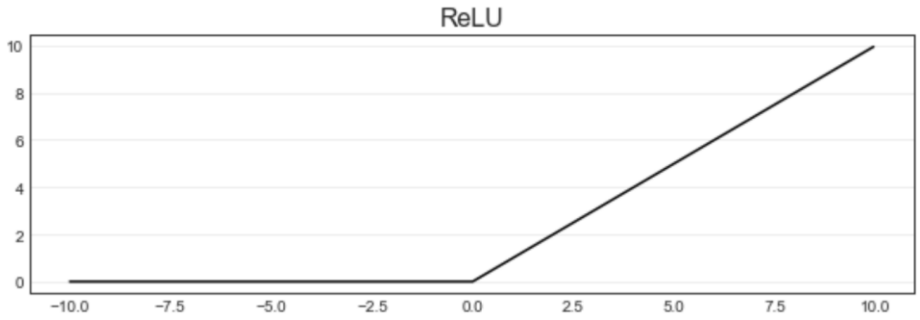
Linear and non-linear

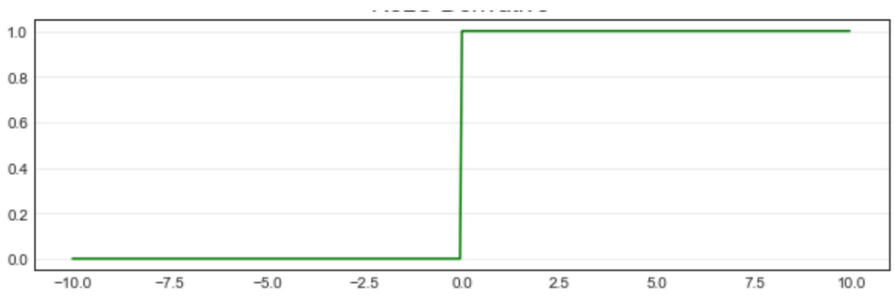
- without h(z), what a layer does is alike affine transformation, (linear in, linear out)
- suitble non-linear h(z) can bring non-linear transformation
- g(z) only relates to the output and is independent with the network
- when we talk activation function, it means h(z)
- g(z) is called out activation function



Activation function (ReLU)

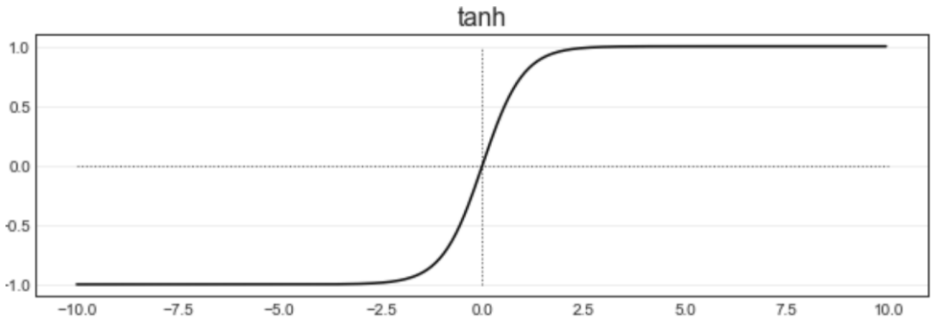
- ReLU (Rectified Linear Unit) $ReLU: h(z) = \begin{cases} z & (z > 0) \\ 0 & (z \leq 0) \end{cases}$

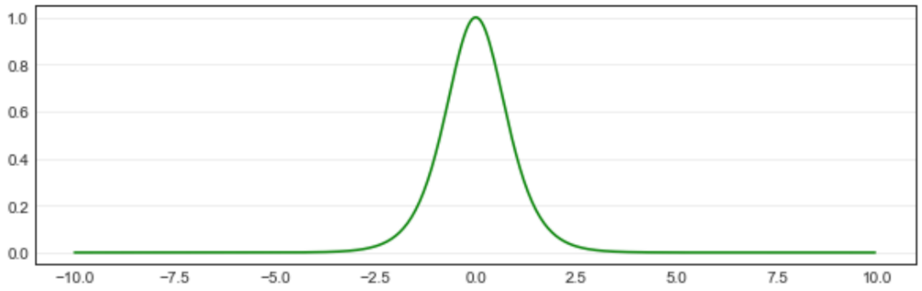




Activation function (tanh)

- tanh (hyperbolic tangent) $$tanh : h(z) = \frac{e^{2z} - 1}{e^{2z} + 1}$$





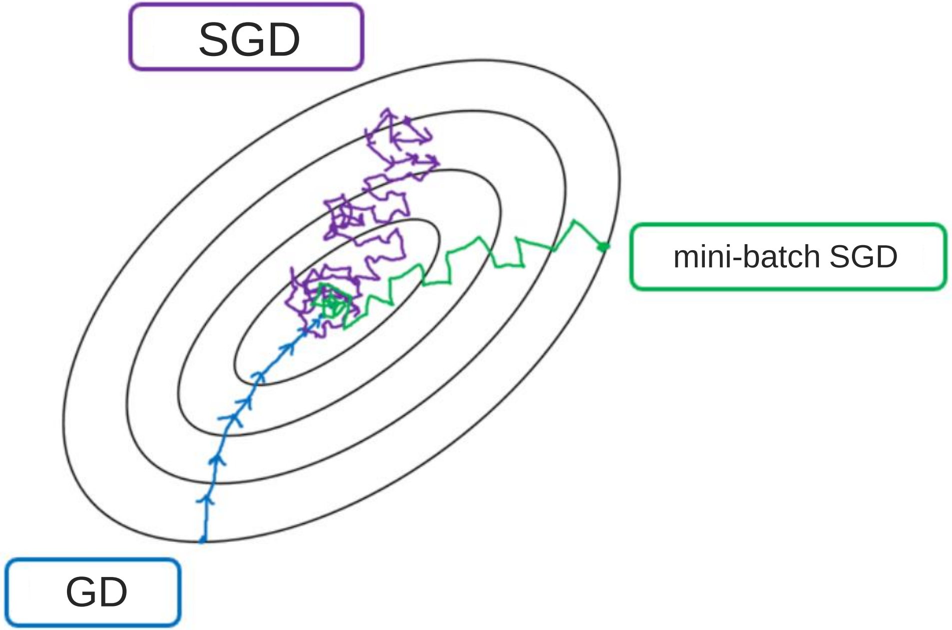
Example #1

- try if one hidden layer works
- 3 to 5 hidden layers
- number of neurons on input > number of neurons on hidden layer > number of neurons on output
- number of neurons on hidden layer = 2/3 * number of neurons on input + number of neurons on output
- number of neurons on hidden layer < 2*number of neurons on input

Learning of ANN parameters

how to determine parameters of “b” and “w” ?

- “lbfgs”,
- “mini-batch S(stochastic)G(gradient)D(descent)” and
- “adam”.



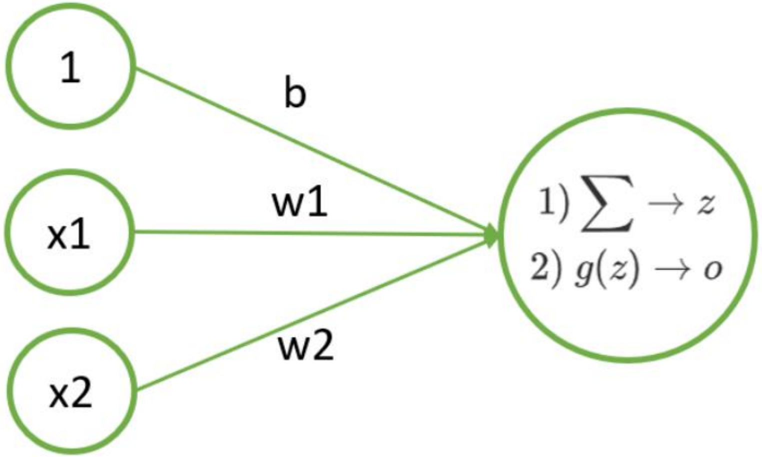
Workflow of learning parameters

Loss functions

- Regression: MSE
- Classification: Cross-Entropy (log loss)

$$L_x(\mathbf{w}, b) = \sum_{i=1}^{N} \frac{1}{2} ||\hat{y}_i - y_i||_2^2 + \frac{\alpha}{2} ||w||_2^2$$

$$\frac{\partial L_x(w_1, w_2, b)}{\partial w}, \frac{\partial L_x(w_1, w_2, b)}{\partial b}$$



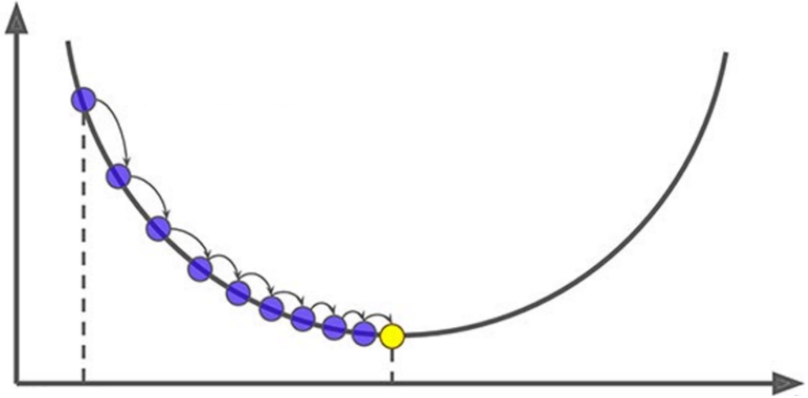
Workflow of learning parameters

$$w_1^{(1)} = w_1^{(0)} - \frac{\boxed{\eta^{(0)}}}{\boxed{N_{\mathrm{B}}}} * \sum_{\boxed{i\in \mathrm{B}_0}} \frac{\partial L_{\boxed{x_i}}(w_1, w_2, b)}{\partial w_1^{(0)}}$$

$$w_2^{(1)} = w_2^{(0)} - \frac{\eta^{(0)}}{N_{\mathrm{B}}} * \sum_{i\in \mathrm{B}_0} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial w_2^{(0)}}$$

$$b^{(1)} = b^{(0)} - \frac{\eta^{(0)}}{N_{\mathrm{B}}} * \sum_{i\in \mathrm{B}_0} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial b^{(0)}}$$

$$\eta^{(1)} = \frac{\eta^{(0)}}{1^{power\_t}}$$



Workflow of learning parameters

$$w_1^{(2)} = w_1^{(1)} - \frac{\eta^{(1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_1} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial w_1^{(1)}}$$

$$w_2^{(2)} = w_2^{(1)} - \frac{\eta^{(1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_1} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial w_2^{(1)}}$$

$$b^{(2)} = b^{(1)} - \frac{\eta^{(1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_1} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial b^{(1)}}$$

$$\eta^{(2)} = \frac{\eta^{(0)}}{2^{power\_t}}$$

$$w_1^{(t)} = w_1^{(t-1)} - \frac{\eta^{(t-1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_{t-1}} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial w_1^{(t-1)}}$$

$$w_2^{(t)} = w_2^{(t-1)} - \frac{\eta^{(t-1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_{t-1}} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial w_2^{(t-1)}}$$

$$b^{(t)} = b^{(t-1)} - \frac{\eta^{(t-1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_{t-1}} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial b^{(t-1)}}$$

$$\eta^{(t)} = \frac{\eta^{(0)}}{t^{power\_t}}$$

Workflow of learning parameters

$$w_1^{(2)} = w_1^{(1)} - \frac{\eta^{(1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_1} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial w_1^{(1)}}$$

$$w_2^{(2)} = w_2^{(1)} - \frac{\eta^{(1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_1} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial w_2^{(1)}}$$

$$b^{(2)} = b^{(1)} - \frac{\eta^{(1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_1} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial b^{(1)}}$$

$$\eta^{(2)} = \frac{\eta^{(0)}}{2^{power\_t}}$$

$$w_1^{(t)} = w_1^{(t-1)} - \frac{\eta^{(t-1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_{t-1}} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial w_1^{(t-1)}}$$

$$w_2^{(t)} = w_2^{(t-1)} - \frac{\eta^{(t-1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_{t-1}} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial w_2^{(t-1)}}$$

$$b^{(t)} = b^{(t-1)} - \frac{\eta^{(t-1)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_{t-1}} \frac{\partial L_{x_i}(w_1, w_2, b)}{\partial b^{(t-1)}}$$

$$\eta^{(t)} = \frac{\eta^{(0)}}{t^{power\_t}}$$

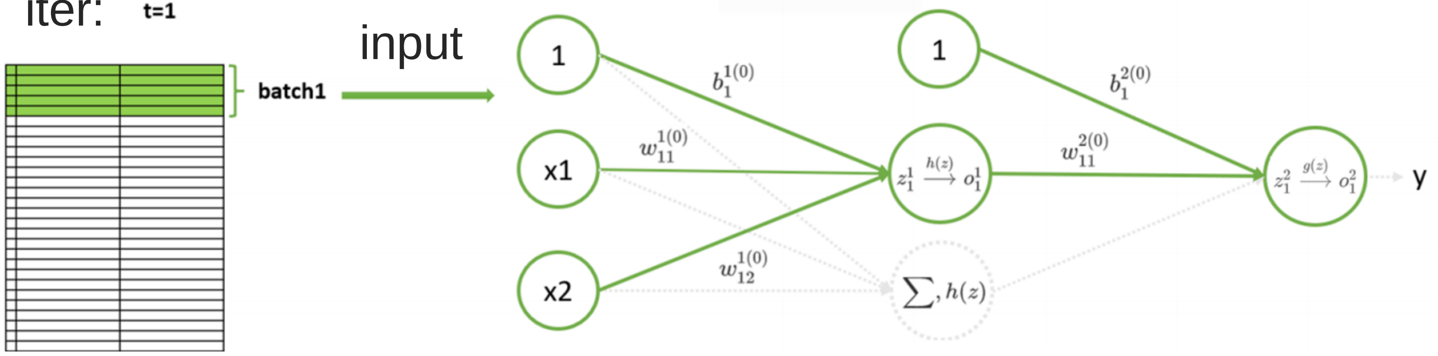
Example #2

Forward broadcast

$$\mathbf{b}^{(0)} = [\, b_1^{1(0)}, b_2^{1(0)}, b_1^{2(0)} \,]$$

$$\mathbf{w}^{(0)} = [\, w_{11}^{1(0)}, w_{12}^{1(0)}, w_{21}^{1(0)}, w_{22}^{1(0)}, w_{11}^{2(0)}, w_{12}^{2(0)} \,]$$

iter: t=1



Forward broadcast

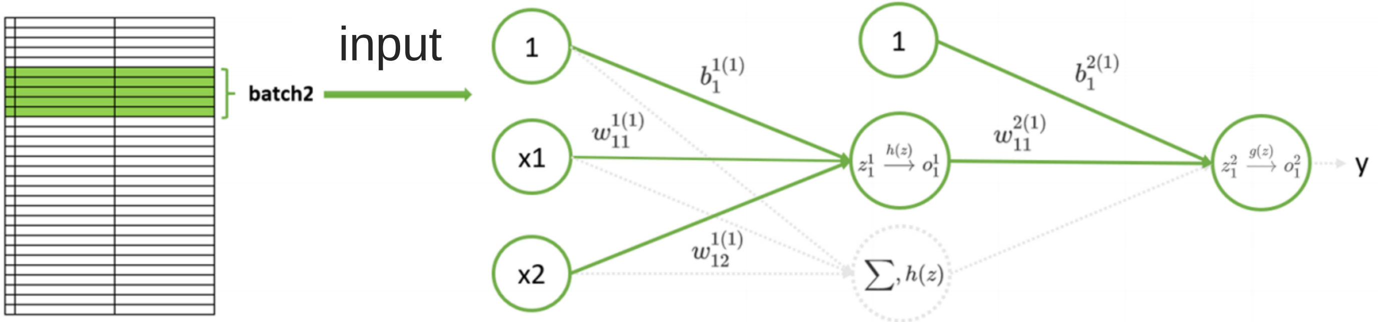
new batch

after first iter

$$\mathbf{b}^{(1)} = [\, b_1^{1(1)}, b_2^{1(1)}, b_1^{2(1)} \,]$$

$$\mathbf{w}^{(1)} = [\, w_{11}^{1(1)}, w_{12}^{1(1)}, w_{21}^{1(1)}, w_{22}^{1(1)}, w_{11}^{2(1)}, w_{12}^{2(1)} \,]$$

iter: t=2



Forward broadcast

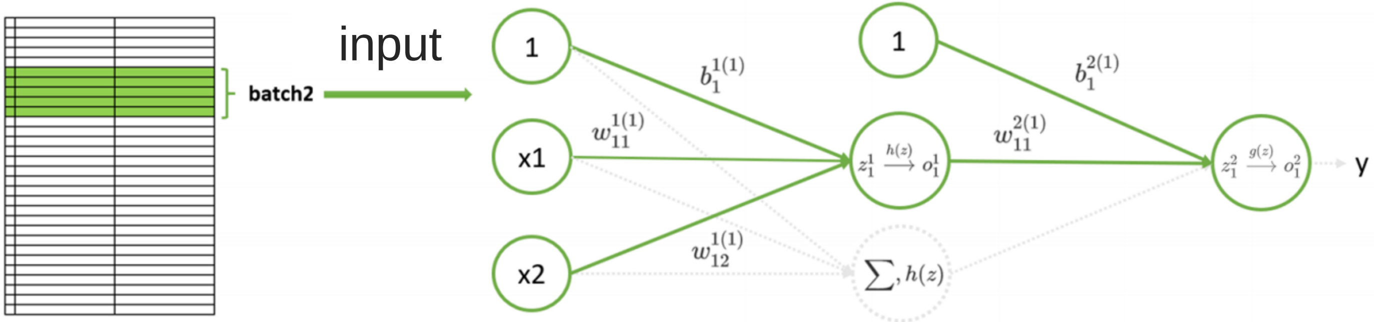
after first iter

new batch

$$\mathbf{b}^{(1)} = [\, b_1^{1(1)}, b_2^{1(1)}, b_1^{2(1)} \,]$$

$$\mathbf{w}^{(1)} = [\, w_{11}^{1(1)}, w_{12}^{1(1)}, w_{21}^{1(1)}, w_{22}^{1(1)}, w_{11}^{2(1)}, w_{12}^{2(1)} \,]$$

iter: **t=2**



Forward broadcast

Forward broadcast

- from input to output layers
- prediction
- generate and save intermediate variables

after sec iter

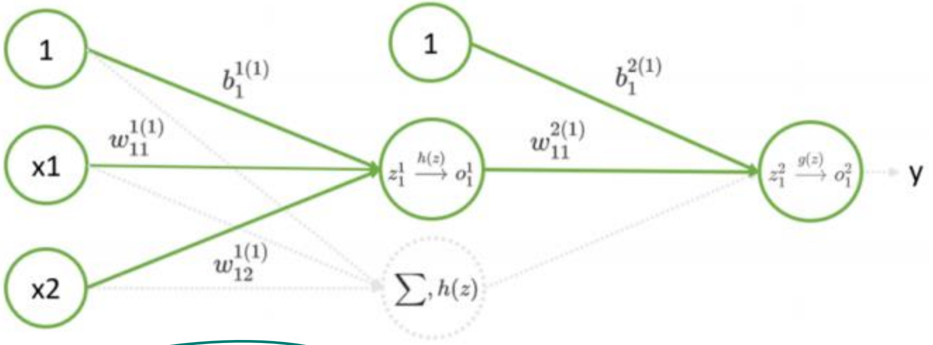
......

after t-th iter

iter: t

$$\mathbf{b^{(t)}} = [\, b_1^{1(t)}, b_2^{1(t)}, b_1^{2(t)} \,]$$

$$\mathbf{w^{(t)}} = [\, w_{11}^{1(t)}, w_{12}^{1(t)}, w_{21}^{1(t)}, w_{22}^{1(t)}, w_{11}^{2(t)}, w_{12}^{2(t)} \,]$$



$$o_1^2 = g(z_1^2)$$

$$\begin{aligned} z_1^2 &= f_{11}^2(o_1^1, o_2^1) \\ &= b_2^{1(0)} + o_1^1 w_{11}^{2(0)} + o_2^1 w_{12}^{2(0)} \end{aligned}$$

$$o_1^1 = h(z_1^1), \;\; o_2^1 = h(z_2^1)$$

$$\begin{aligned} z_1^1 &= f_{11}^1(x_1, x_2) \\ &= b_1^{1(0)} + x_1 w_{11}^{1(0)} + x_2 w_{12}^{1(0)} \\ z_2^1 &= f_{21}^1(x_1, x_2) \\ &= b_2^{1(0)} + x_1 w_{21}^{1(0)} + x_2 w_{22}^{1(0)} \end{aligned}$$

$$\begin{aligned} o_1^2 &= g(z_1^2) \\ &= g(f_{11}^2(o_1^1, o_2^1)) \\ &= g(f_{11}^2(h(z_1^1), h(z_2^1))) \\ &= g\left\{ f_{11}^2 \left[h\left\langle f_{11}^1(x_1, x_2)\right\rangle, h\left\langle f_{21}^1(x_1, x_2)\right\rangle \right] \right\} \end{aligned}$$

$$-y_i ln(o_1^2) = -y_i ln(g\left\{ f_{11}^2 \left[h\left\langle f_{11}^1(x_1, x_2)\right\rangle, h\left\langle f_{21}^1(x_1, x_2)\right\rangle \right] \right\})$$

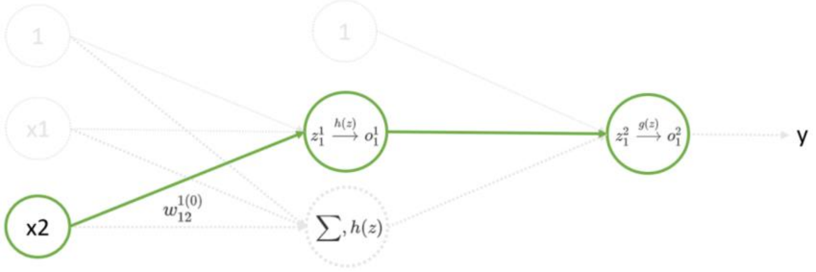
$$L_{x_i}(\mathbf{w}, \mathbf{b}) = -y_i ln(o_1^2) - (1 - y_i) ln(1 - o_1^2)) + \frac{\alpha ||\mathbf{w}||_2^2}{N}$$

Back-propagation

$$L_x(\mathbf{w},\mathbf{b}) = l_x(\mathbf{w},\mathbf{b}) + s(\mathbf{w})$$

$$\begin{aligned}
\frac{\partial l}{\partial w} &= \frac{\partial l(o)}{\partial o_1^2} * \frac{\partial o_1^2}{\partial w} \\
&= \frac{\partial l(o)}{\partial o_1^2} * \frac{\partial(-y_i ln(g\left\{\, f_{11}^2\left[\, h\left\langle f_{11}^1(x_1, x_2)\right\rangle, h\left\langle f_{21}^1(x_1, x_2)\right\rangle \,\right]\right\})}{\partial w} \\
&= \frac{\partial l(o)}{\underline{\partial o_1^2}} * \frac{\partial o_1^2}{\underline{\partial z_1^2}} * \frac{\partial z_1^2}{\partial w} \\
&= \frac{\partial l(o)}{\partial g(z)} * \frac{\partial g(z)}{\partial f_{11}^2(w)} * \frac{\partial f_{11}^2(w)}{\partial w}, \qquad w = w_{11}^{2(0)}
\end{aligned}$$

Back-propagation



$$\frac{\partial l}{\partial w} = \frac{\partial l(o)}{\partial o_1^2} * \frac{\partial o_1^2}{\partial w}$$

$$= \frac{\partial l(o)}{\partial o_1^2} * \frac{\partial(-y_i ln(g\{\, f_{11}^2[\, h\langle f_{11}^1(x_1, x_2)\rangle , h\langle f_{21}^1(x_1, x_2)\rangle \,]\,\})}{\partial w}$$

$$= \frac{\partial l(o)}{\partial o_1^2} * \frac{\partial o_1^2}{\partial z_1^2} * \frac{\partial z_1^2}{\partial o_1^1} * \frac{\partial o_1^1}{\partial z_1^1} * \frac{\partial z_1^1}{\partial w}$$

$$= \boxed{\frac{\partial l(o)}{\partial g(z)} * \frac{\partial g(z)}{\partial f_{11}^2(h)}} * \frac{\partial f_{11}^2(h)}{\partial h(f_{11}^1)} * \frac{\partial h(f_{11}^1)}{\partial f_{11}^1(w)} * \frac{\partial f_{11}^1(w)}{\partial w}, \quad w = w_{12}^{1(0)}$$

$$w_{21}^{1(1)} = w_{21}^{1(0)} - \frac{\eta^{(0)}}{N_{\mathrm{B}}} * \sum_{i \in \mathrm{B}_0} \boxed{\frac{\partial L_{x_i}(\mathbf{w}, \mathbf{b})}{\partial w_1^{(0)}}}$$

CHALMERS

UNIVERSITY OF TECHNOLOGY