lecture 4

Dong-Geol Choi
Hanbat Nat'l Univ.

Pizza delivery time by multi input

staff capacity	weather	delivery distance	delivery time
100	1	100m	20min
200	1	150m	24min
300	2	300m	36min
400	1	400m	47min
500	1	130m	22min
600	2	240m	32min
700	1	350m	47min
800	1	200m	42min
900	2	100m	21min
1000	1	110m	21min

Pizza delivery time by multi input

linear Hypothesis with single input

$$H(x) = Wx$$

linear Hypothesis with multi input

$$H(x_1, x_2, x_3) = w_1 x_1 + w_2 x_2 + w_3 x_3 + b$$

Cost Function

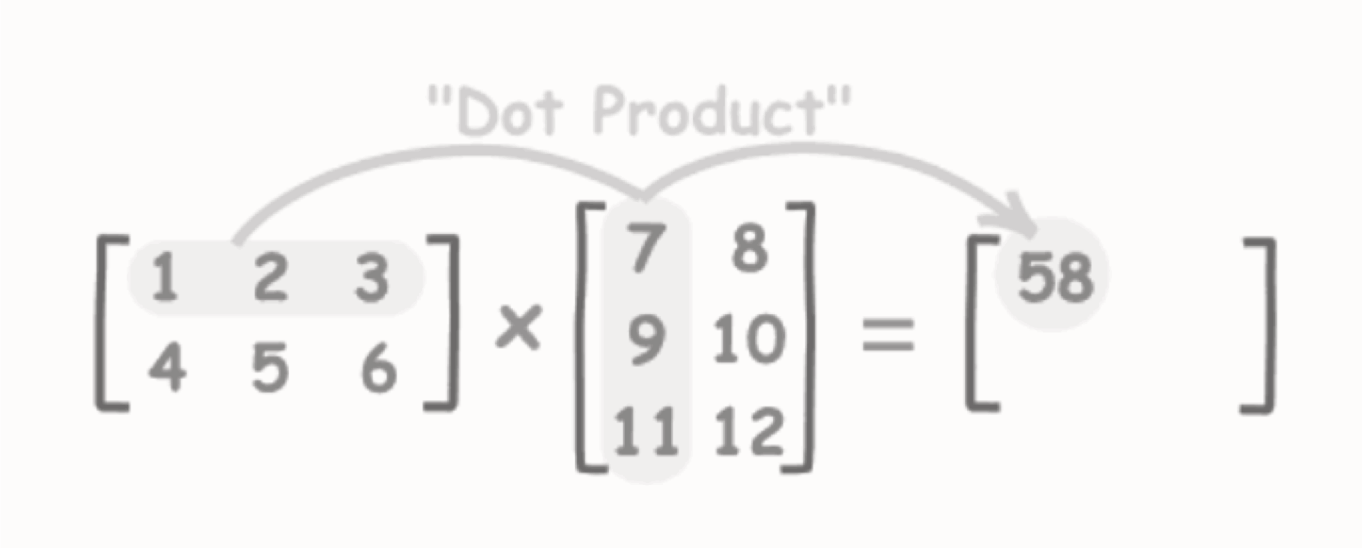
$$H(x_1, x_2, x_3) = w_1 x_1 + w_2 x_2 + w_3 x_3 + b$$

$$cost(W, b) = \frac{1}{m} \sum_{I=1}^{m} (H(x_1^{(i)}, x_2^{(i)}, x_3^{(i)}) - y^{(i)})^2$$

Matrix

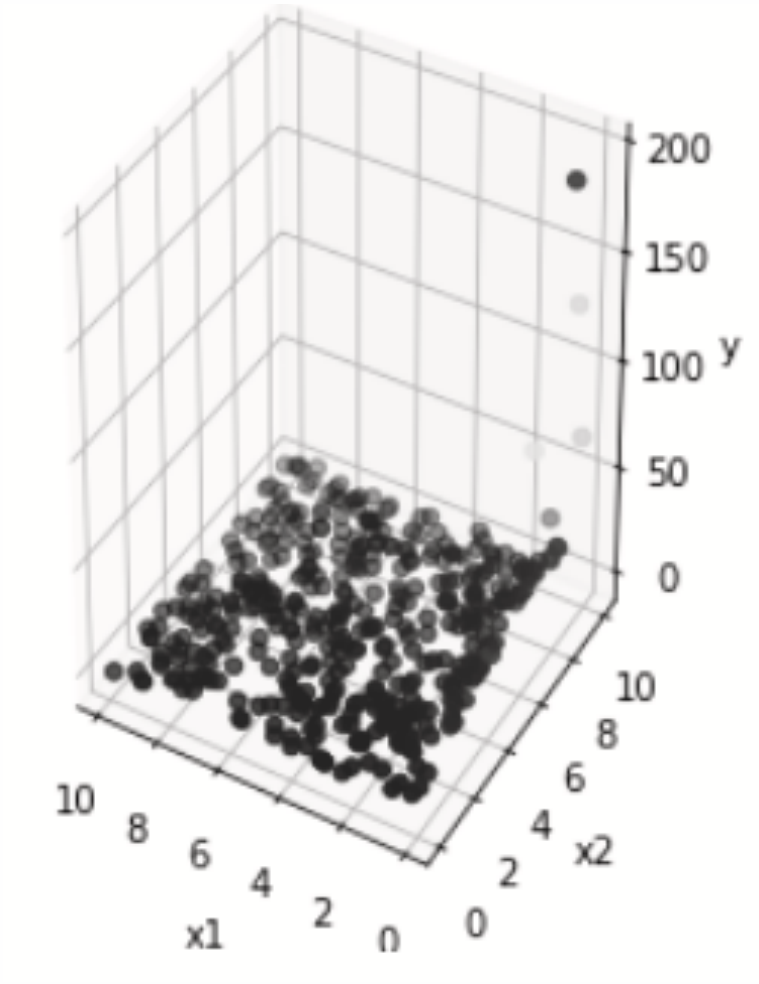
$$w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$$

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} x_1w_1 + x_2w_2 + x_3w_3 \end{pmatrix}$$



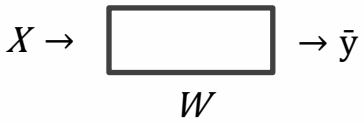
Forward Propagation

staff capacity (x_1)	delivery distance (x_2)	delivery time (Y)
100	100m	20min
200	150m	24min
300	300m	26min
400	400m	24min
500	130m	22min
600	240m	10min
700	350m	22min
800	200m	24min
900	100m	25min
1000	110m	25min



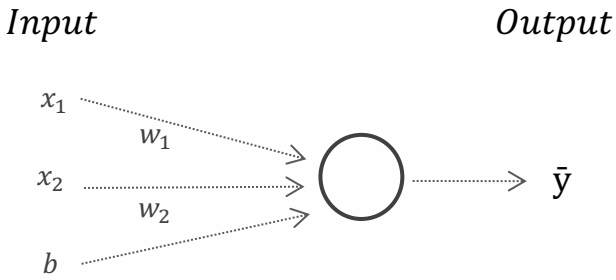
Forward Propagation

staff capacity (x_1)	delivery distance (x_2)	delivery time (Y)
100	100m	20min
200	150m	24min
300	300m	26min
400	400m	24min
500	130m	22min
600	240m	10min
700	350m	22min
800	200m	24min
900	100m	25min
1000	110m	25min



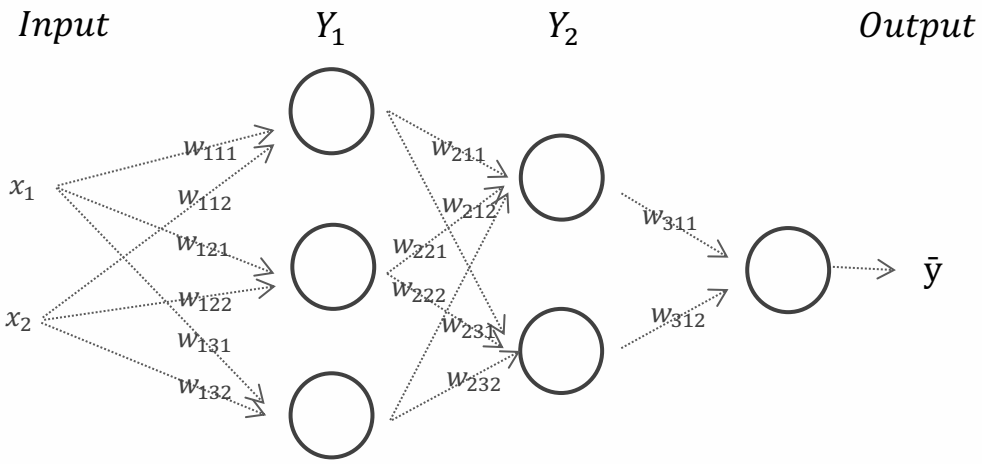
$$Y = W \cdot X + b$$

$$Y = [w_1\ w_2\ ] \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + [b] = [w_1 x_1 + w_2 x_2 + b]$$



Forward Propagation (multi-layer)

staff capacity (x_1)	delivery distance (x_2)	delivery time (Y)
100	100m	20min
200	150m	24min
300	300m	26min
400	400m	24min
500	130m	22min
600	240m	10min
700	350m	22min
800	200m	24min
900	100m	25min
1000	110m	25min

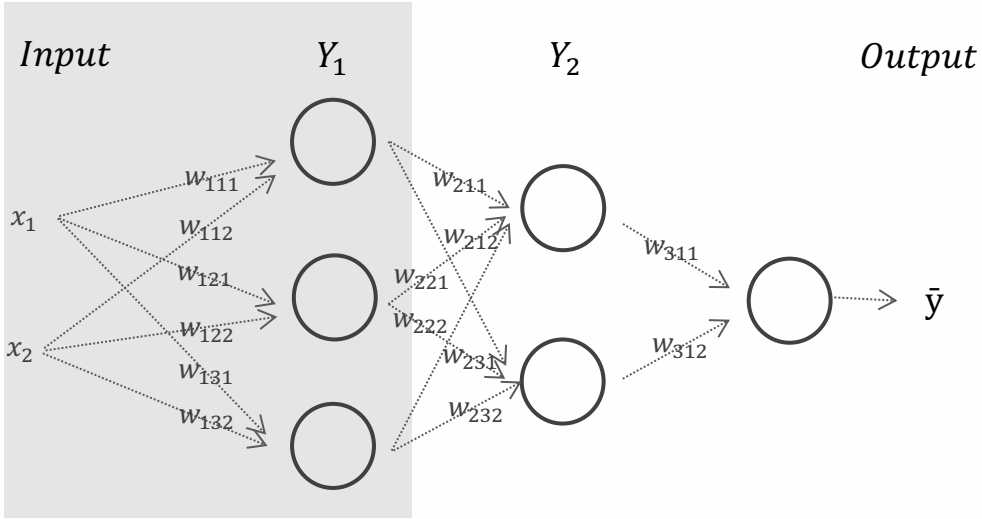


Forward Propagation (multi-layer)

staff capacity (x_1)	delivery distance (x_2)	delivery time (Y)
100	100m	20min
200	150m	24min
300	300m	26min
400	400m	24min
500	130m	22min
600	240m	10min
700	350m	22min
800	200m	24min
900	100m	25min
1000	110m	25min

$$Y_{11} = [w_{111}\ w_{112}\ ] \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b_1 = [w_{111}x_1 + w_{112}x_2 + b_1]$$
$$Y_{12} = [w_{121}\ w_{122}\ ] \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b_1 = [w_{121}x_1 + w_{122}x_2 + b_1]$$
$$Y_{13} = [w_{131}\ w_{132}\ ] \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b_1 = [w_{131}x_1 + w_{132}x_2 + b_1]$$

$$Y_1 = \begin{bmatrix} w_{111}\ w_{112} \\ w_{121}\ w_{122} \\ w_{131}\ w_{132} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b_1 = \begin{bmatrix} w_{111}x_1 + w_{112}x_2 + b_1 \\ w_{121}x_1 + w_{122}x_2 + b_1 \\ w_{131}x_1 + w_{132}x_2 + b_1 \end{bmatrix}$$

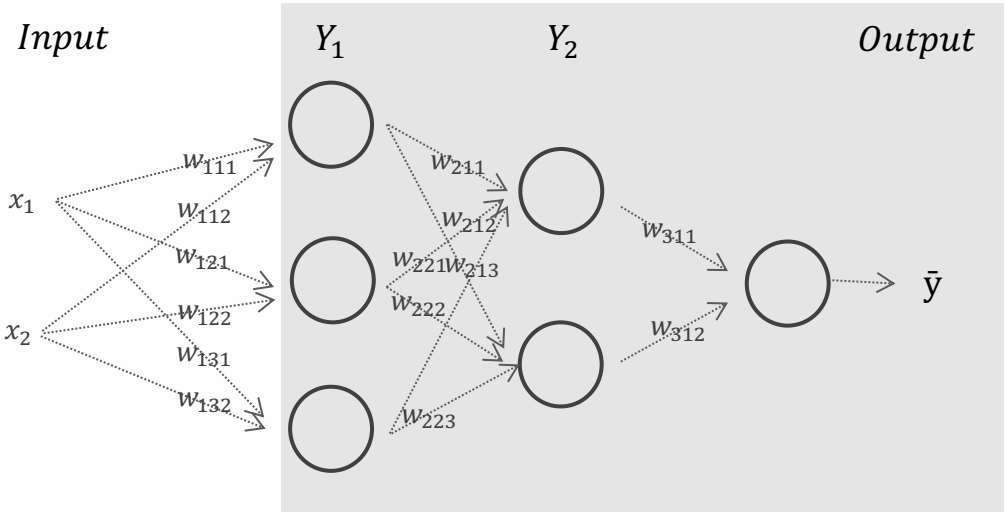


Forward Propagation (multi-layer)

staff capacity (x_1)	delivery distance (x_2)	delivery time (Y)
100	100m	20min
200	150m	24min
300	300m	26min
400	400m	24min
500	130m	22min
600	240m	10min
700	350m	22min
800	200m	24min
900	100m	25min
1000	110m	25min

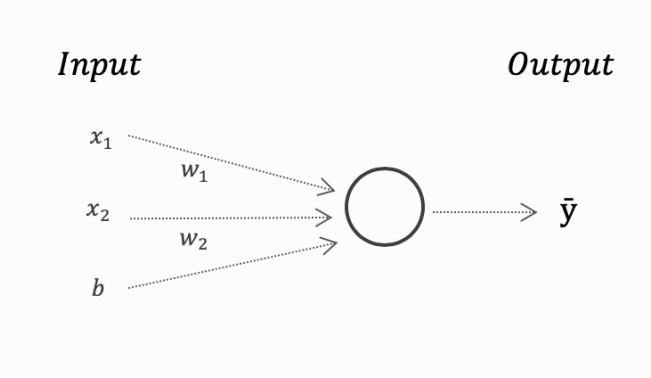
$$Y_2 = \begin{bmatrix} w_{211}\, w_{212}\, w_{213} \\ w_{221}\, w_{222}\, w_{223} \end{bmatrix} \cdot \begin{bmatrix} w_{111}x_1 + w_{112}x_2 + b_1 \\ w_{121}x_1 + w_{122}x_2 + b_1 \\ w_{131}x_1 + w_{132}x_2 + b_1 \end{bmatrix} + b_2$$

$$Y_3 = w_{311}\, w_{312} \cdot Y_2 + b_3$$

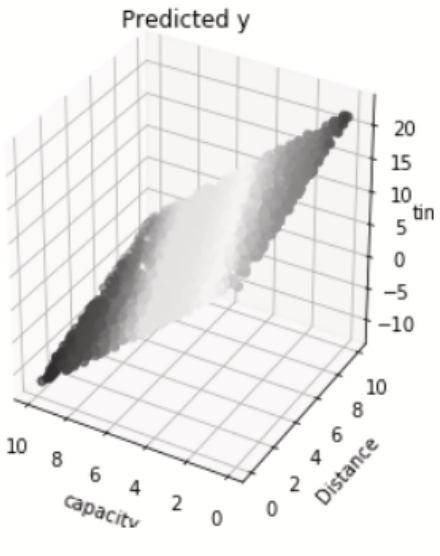


Forward Propagation

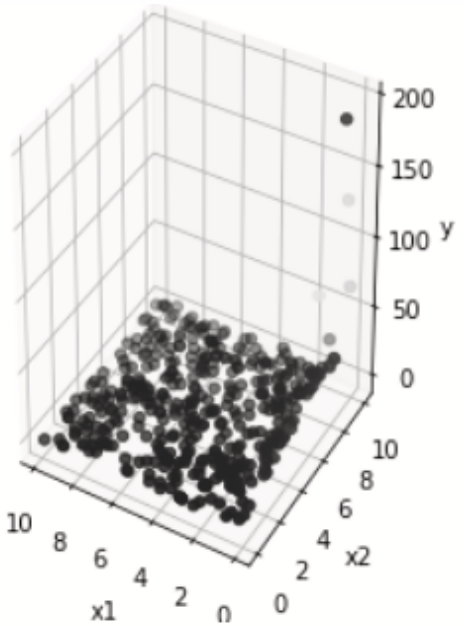
model



model predicted

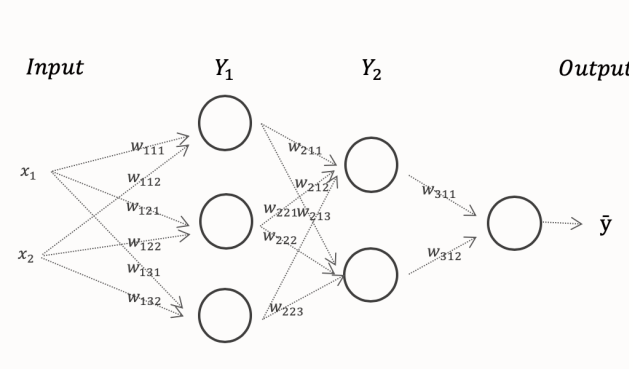


real data

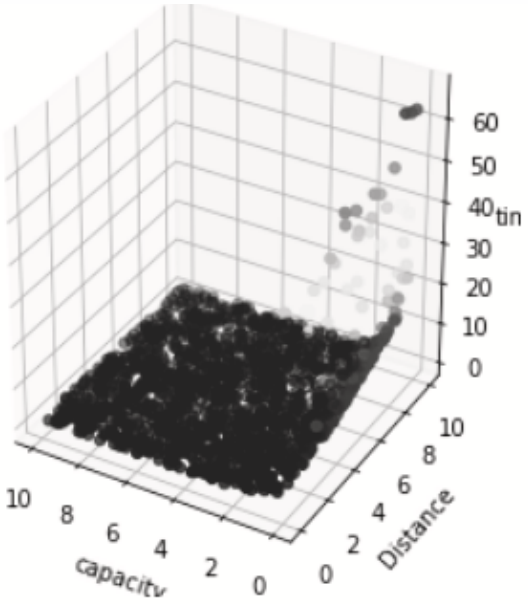


Forward Propagation (multi-layer)

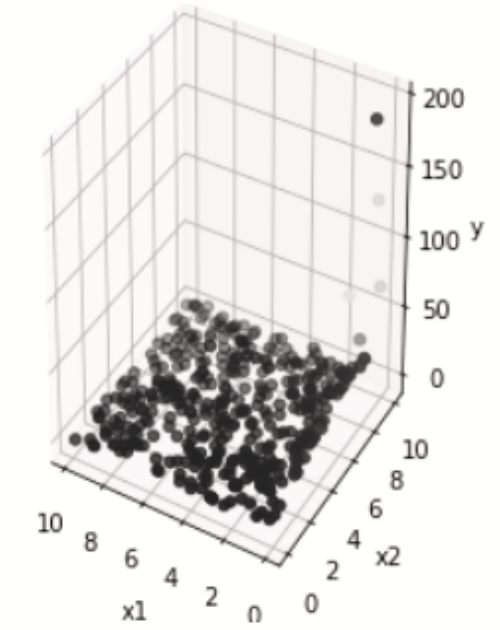
model



model predicted

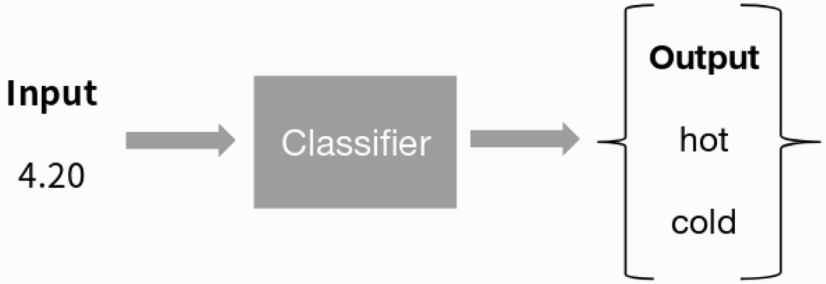


real data

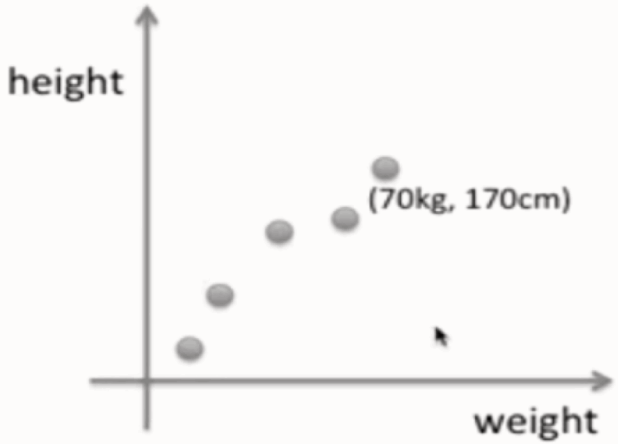


Logistic Classification

Classification vs Regression



classify input into categorical output

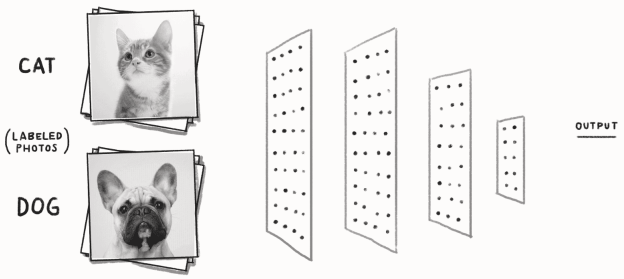


Regression outputs continuous results.

Classification vs Regression

What is logistic Classification?

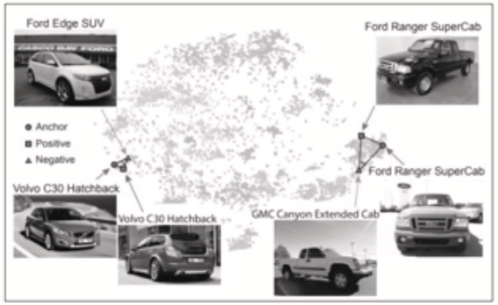
Binary Classification algorithm.
{spam, not spam}, {man or woman}



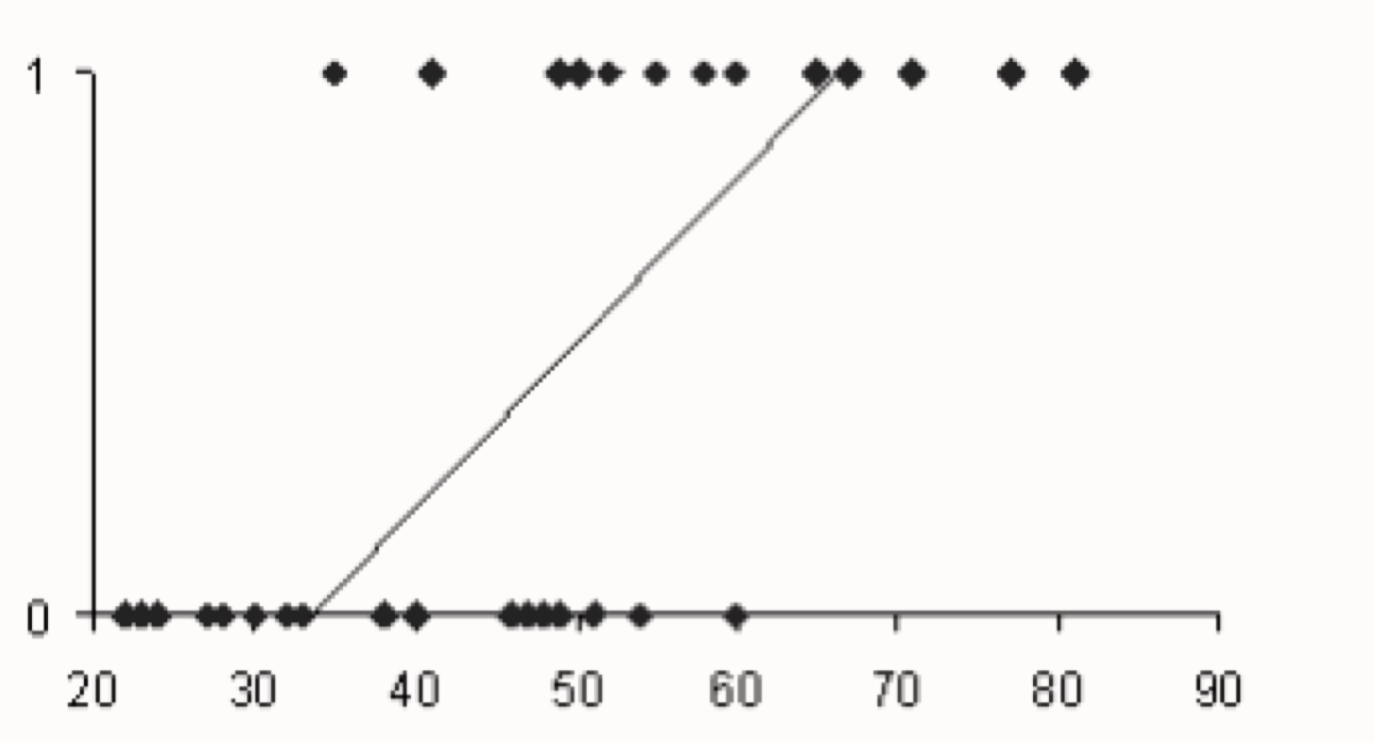
Difference with linear regression?

Linear regression's output is continuous,
Logistic classification output is categorized

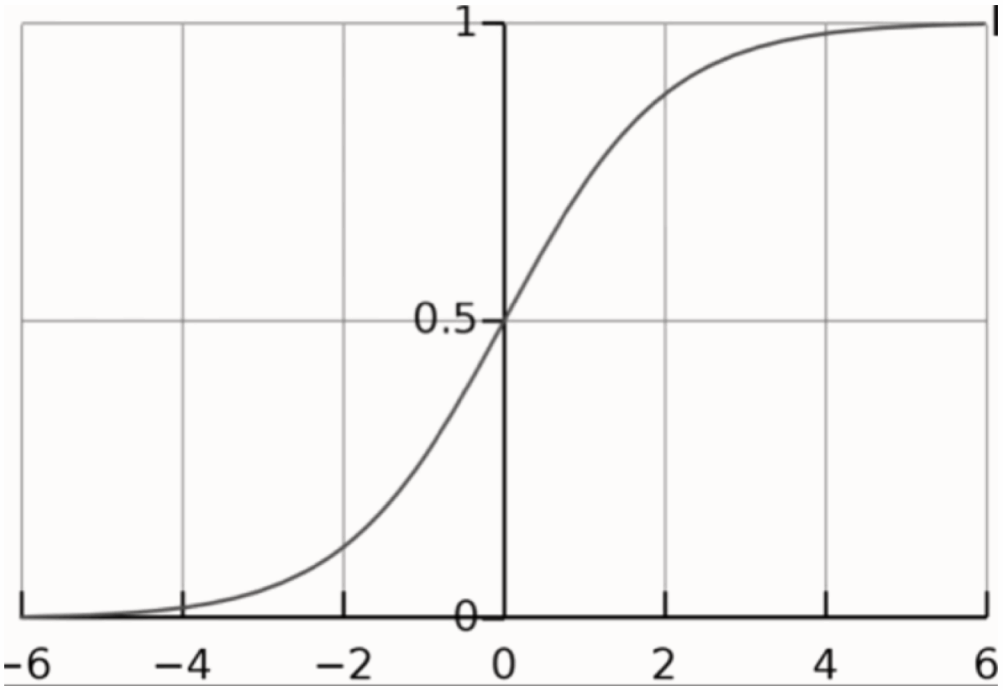
....



Why no linear regression for classification?



Why no linear regression for classification?



sigmoid

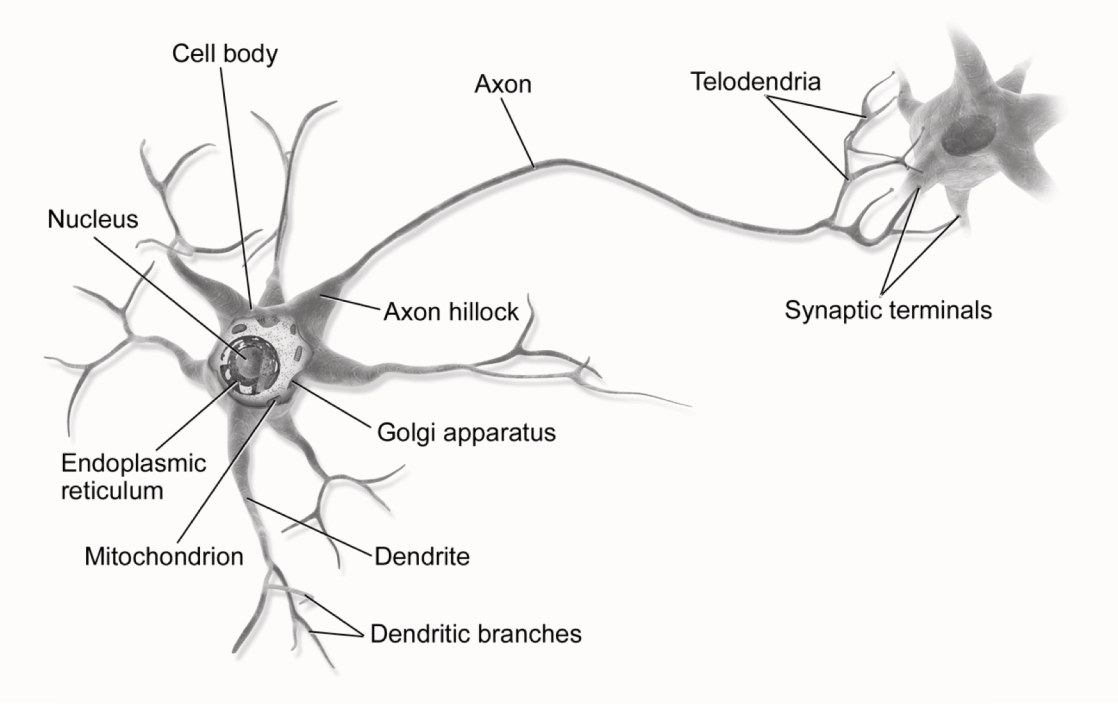
input x

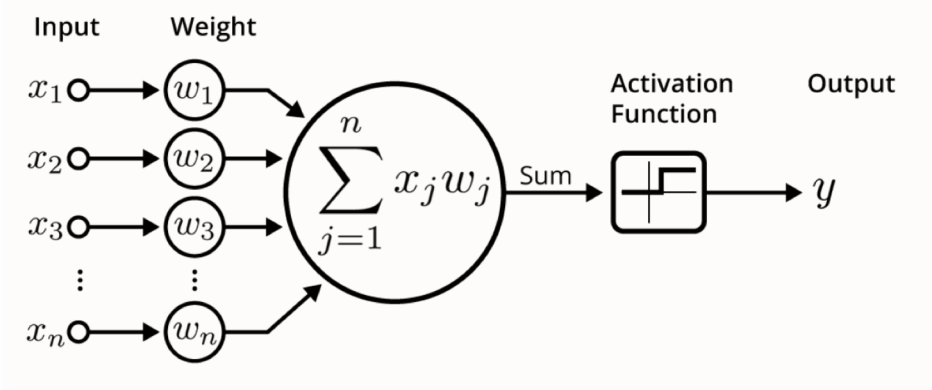
$z = Wx + b$

$h(x) = \frac{1}{1+e^{-z}}$

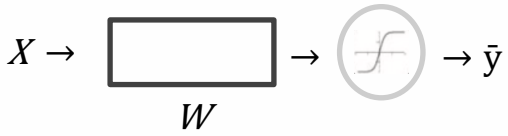
Output $= \frac{1}{1+e^{-(Wx+b)}}$

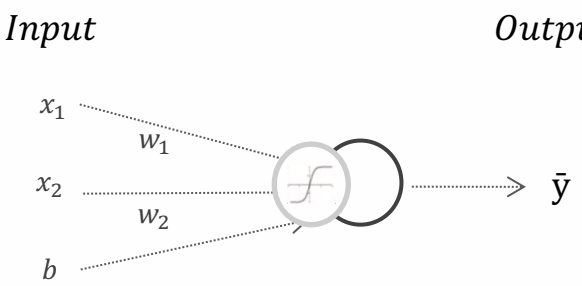
Activation Function

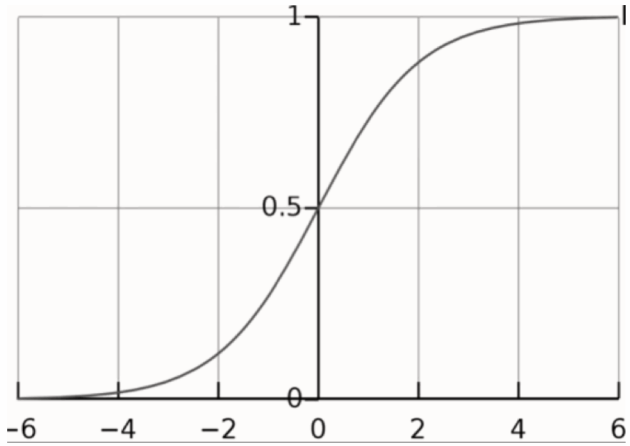




Activation Function

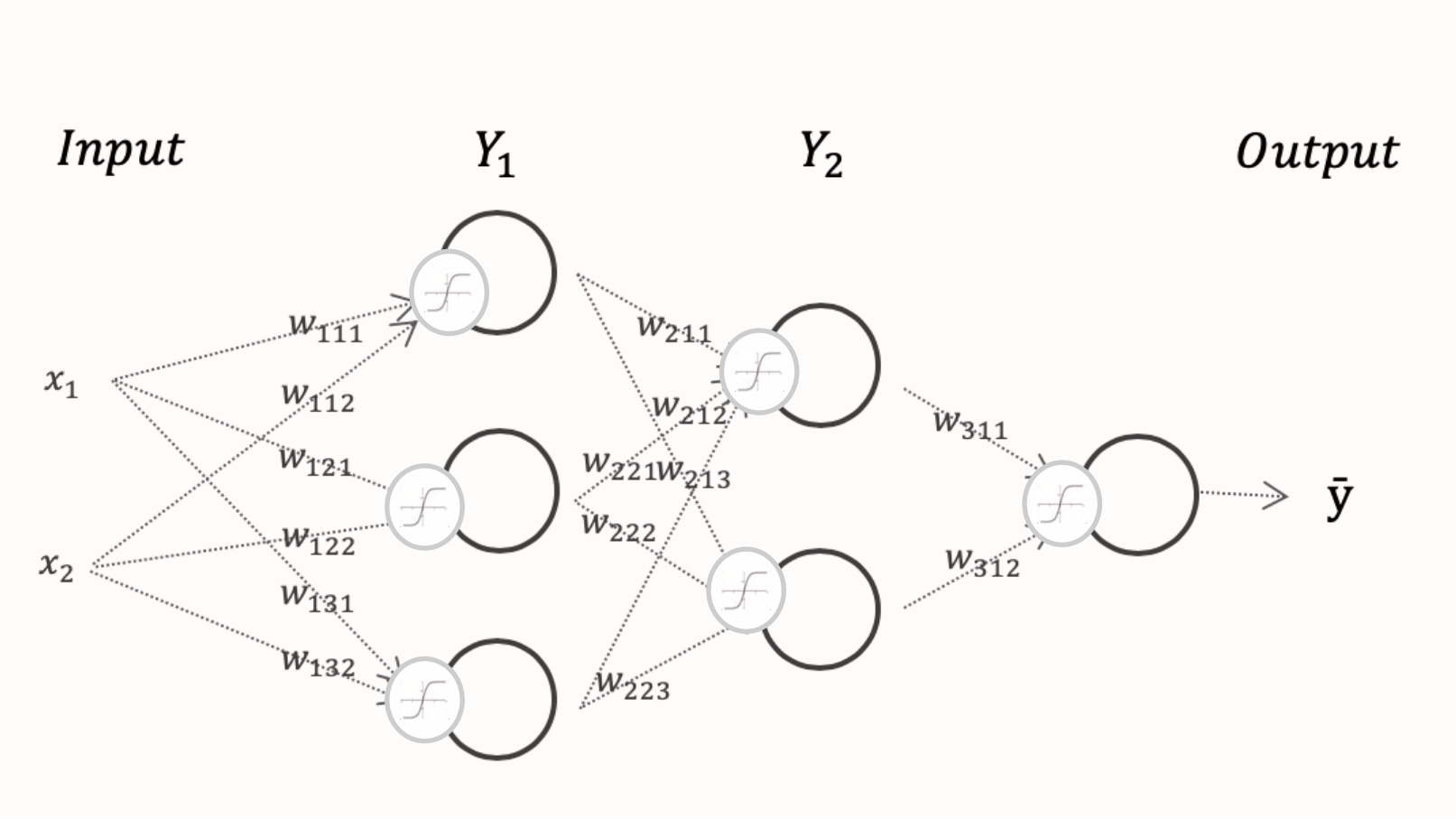






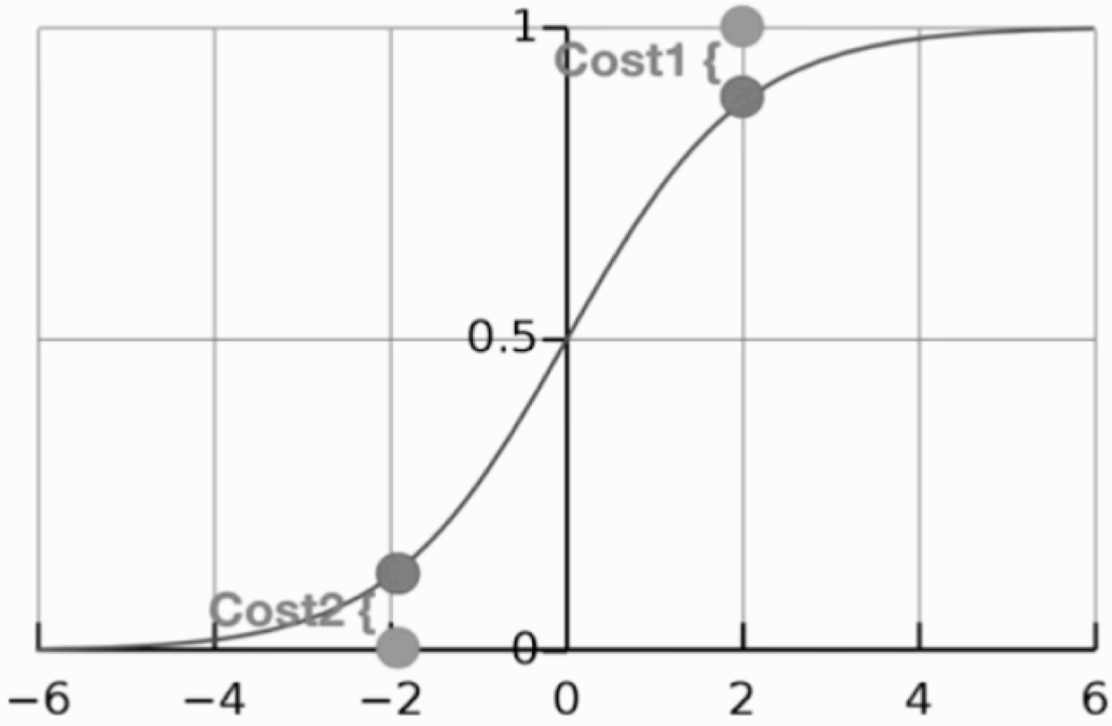
sigmoid

Activation Function



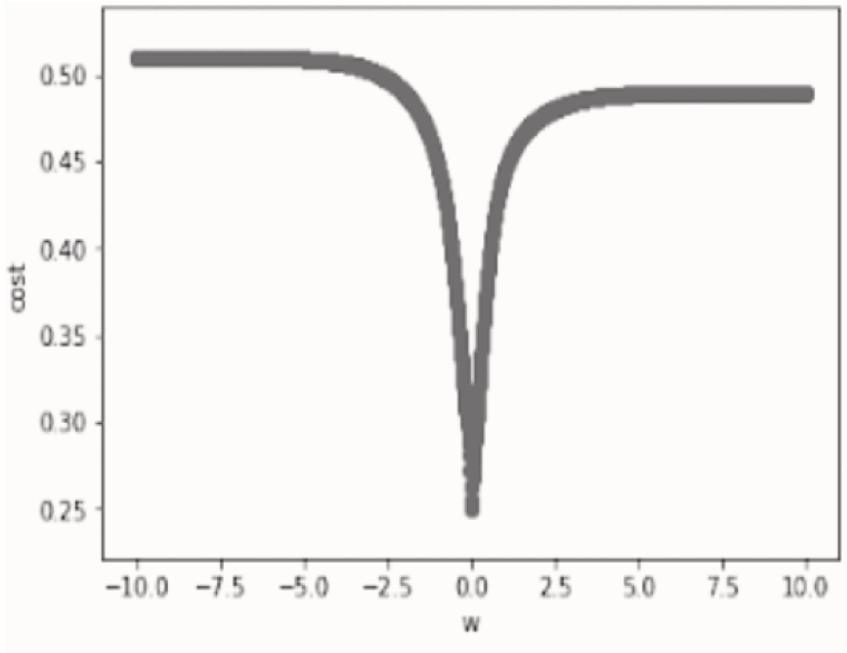
How to minimize cost function?

$$\textbf{Minimize } \frac{1}{m}\sum_{i=1}^{m} cost(h_\theta(x^i), y^i)$$



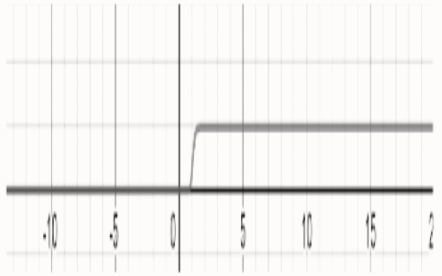
How to minimize cost function?

Mean Square Error? No

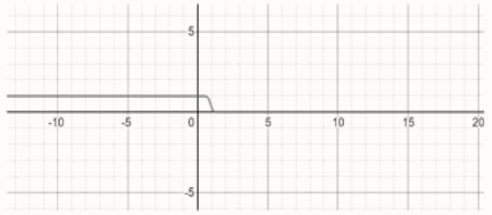


$$\text{cost} = \left(\frac{1}{1+e^{-(Wx[i]+b)}} - y[i]\right)^2$$

Y= 0

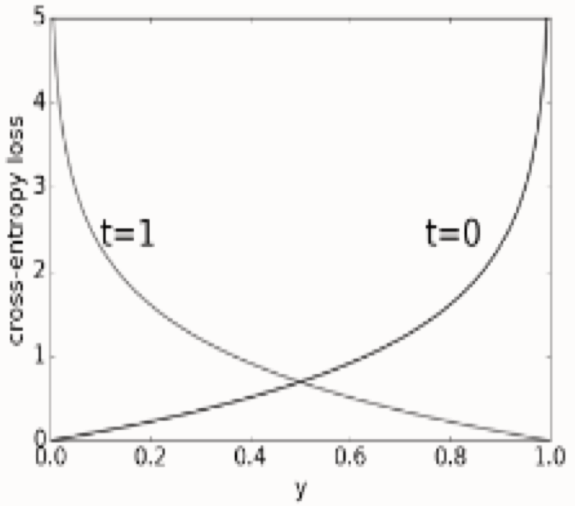


Y= 1



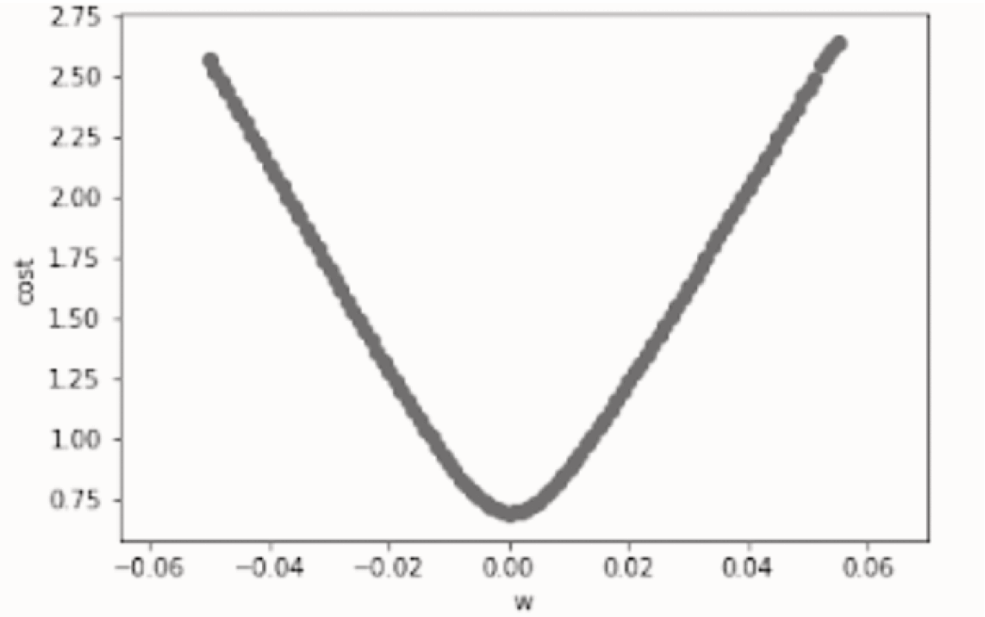
How to minimize cost function?

$$\mathcal{L}_{\mathrm{CE}}(y,t) = \begin{cases} -\log y & \text{if } t = 1 \\ -\log 1 - y & \text{if } t = 0 \end{cases}$$
$$= -t \log y - (1-t) \log 1 - y$$



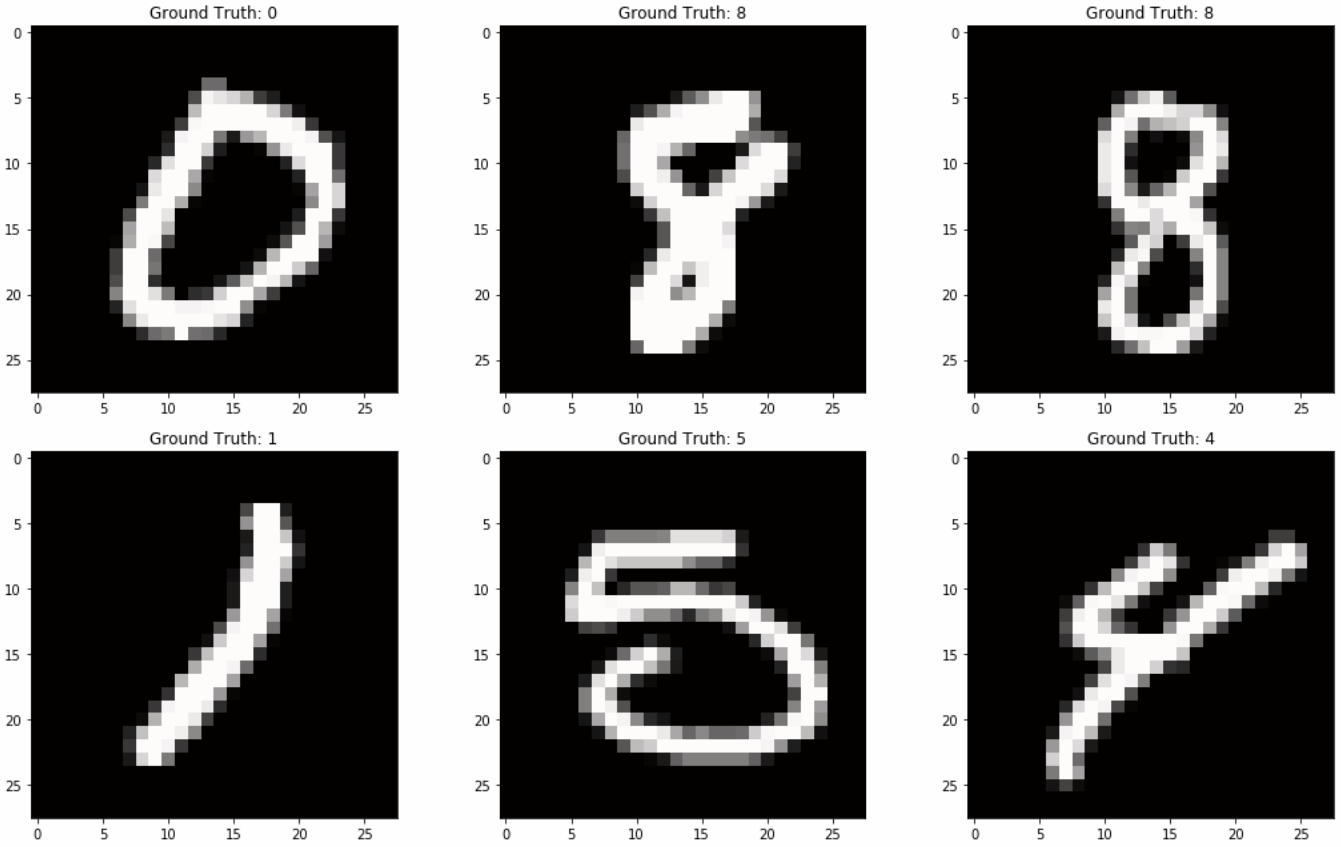
How to minimize cost function?

Cross Entropy? Yes

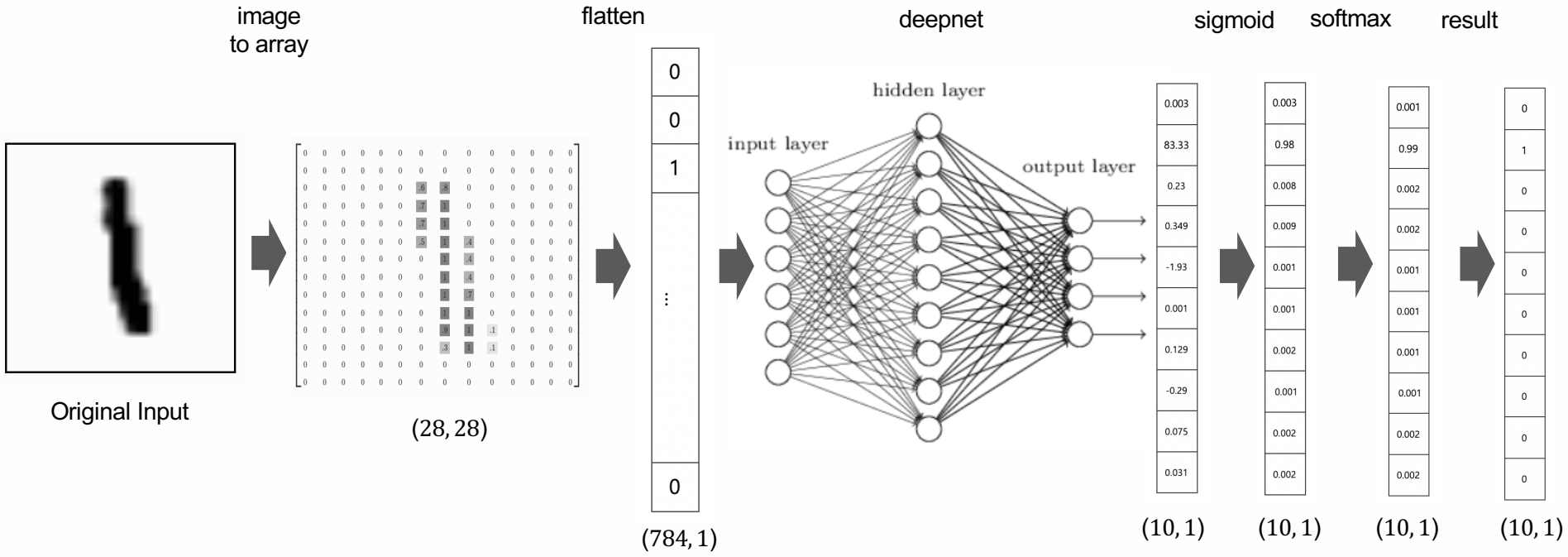


$$L = -\sum_i y_i log\tilde{y}_i + (1 - y_i)\log\,(1 - \tilde{y}_i)$$

MNIST Forward Propagation

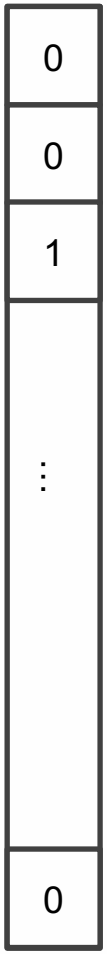


MNIST Forward Propagation Overview



MNIST Forward Propagation

Input



$(784, 1)$

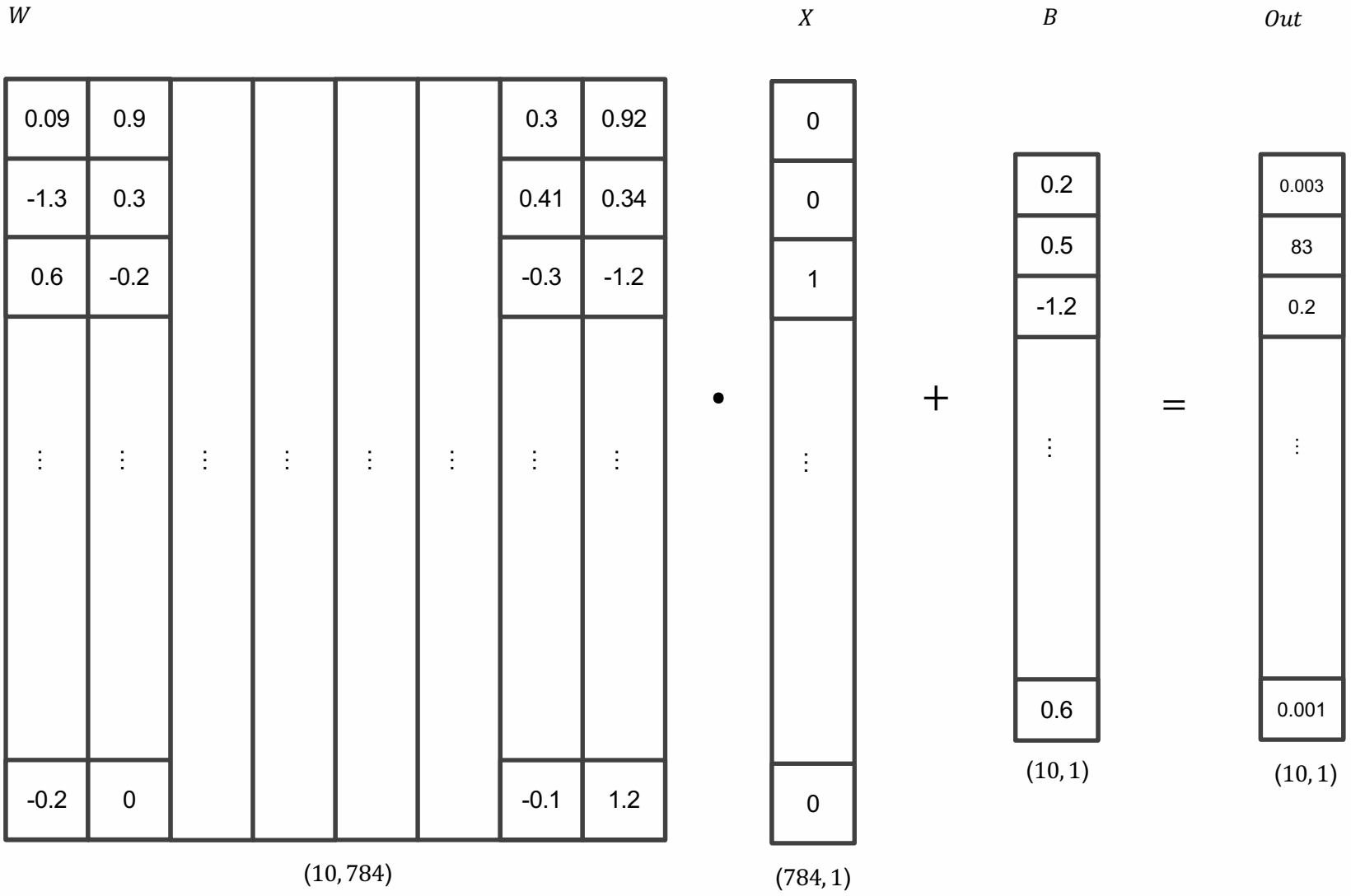
$$WX + B$$

X: $(784, 1)$ → W: $(10, 784)$

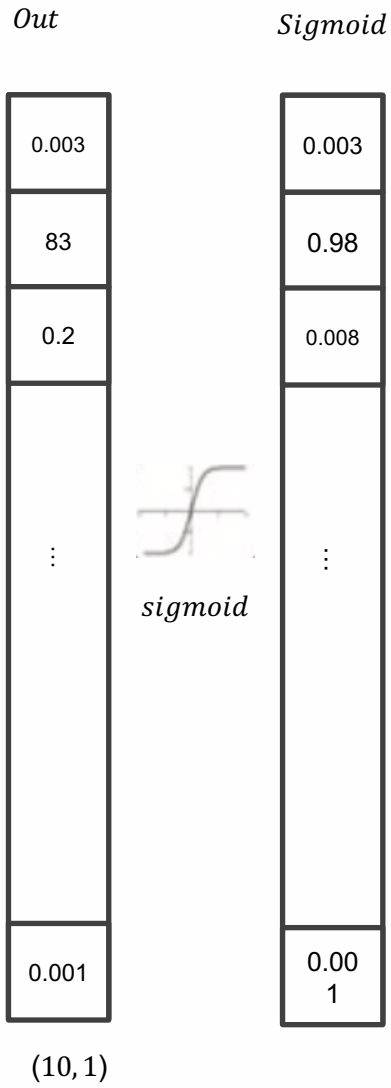
WX: $(10, 784) \cdot (784, 1)$ → $(10, 1)$

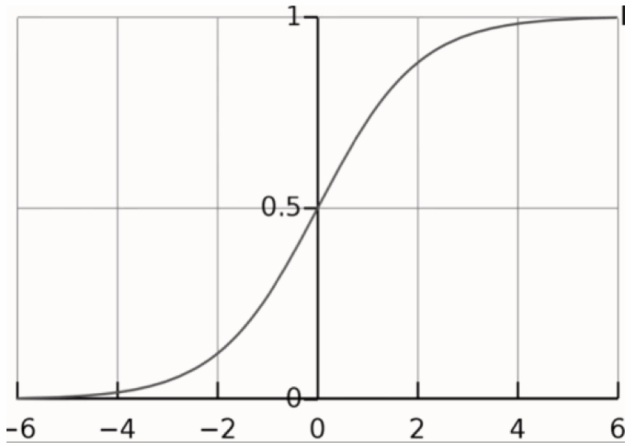
B: $(10, 1)$

MNIST Forward Propagation



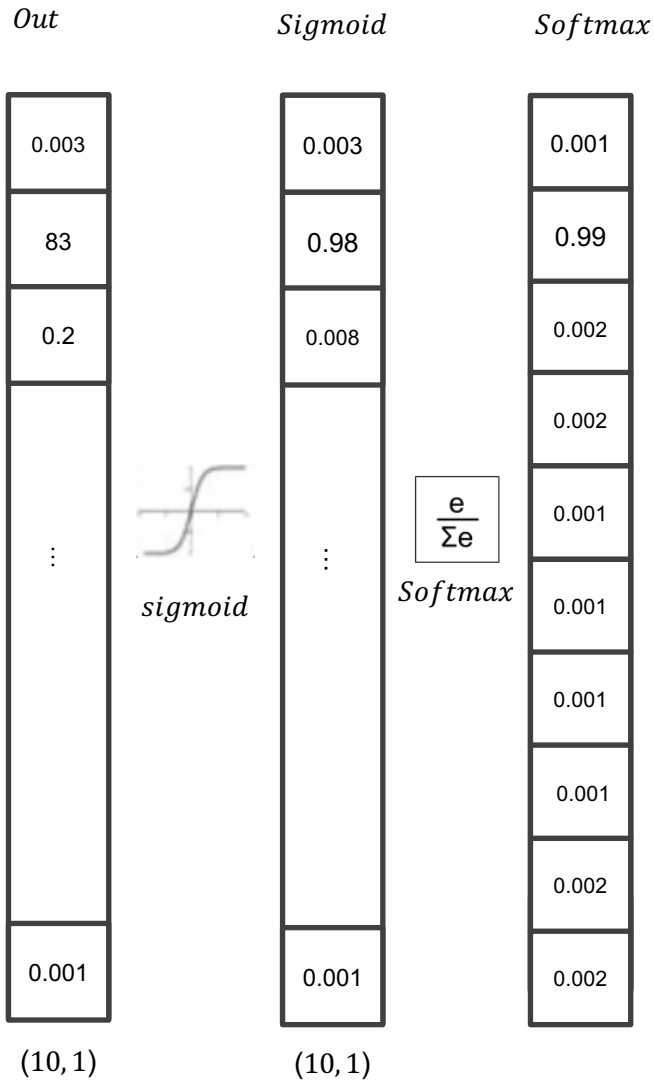
MNIST Forward Propagation: Sigmoid





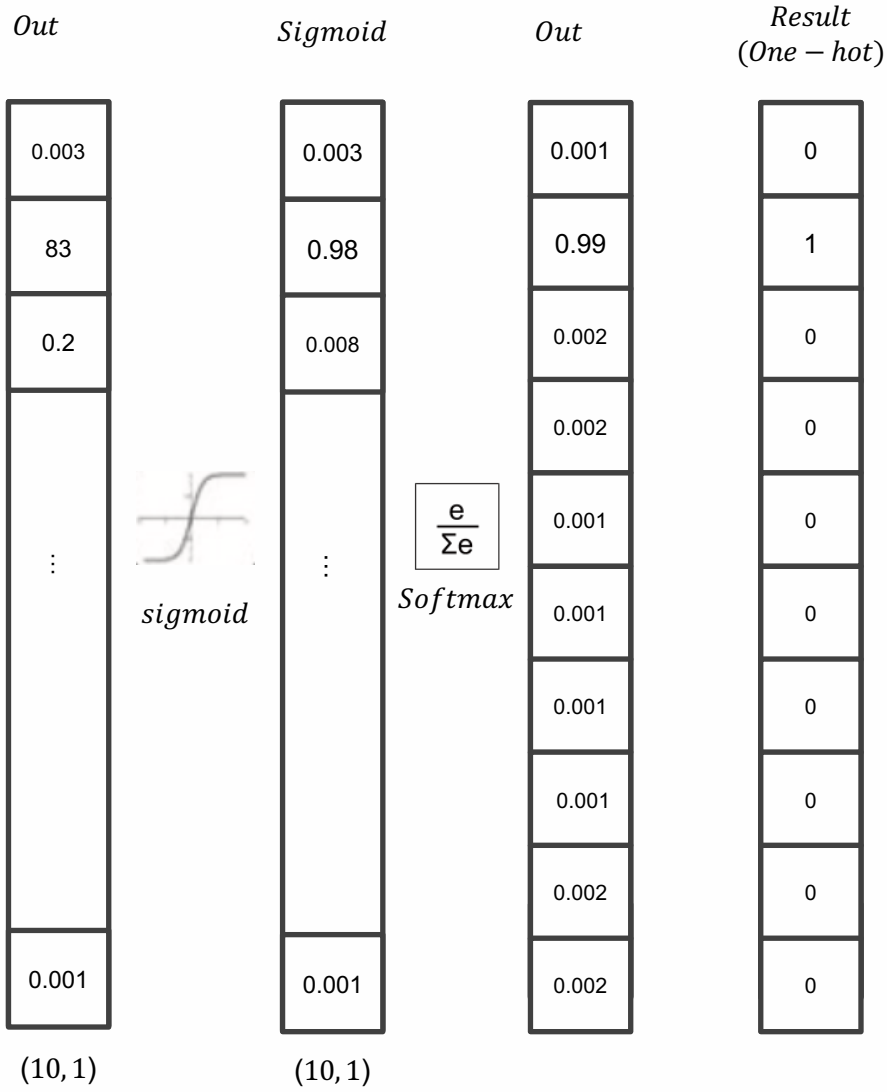
$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}.$$

MNIST Forward Propagation: Softmax



$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}}$$

MNIST Forward Propagation: One-hot



Summary

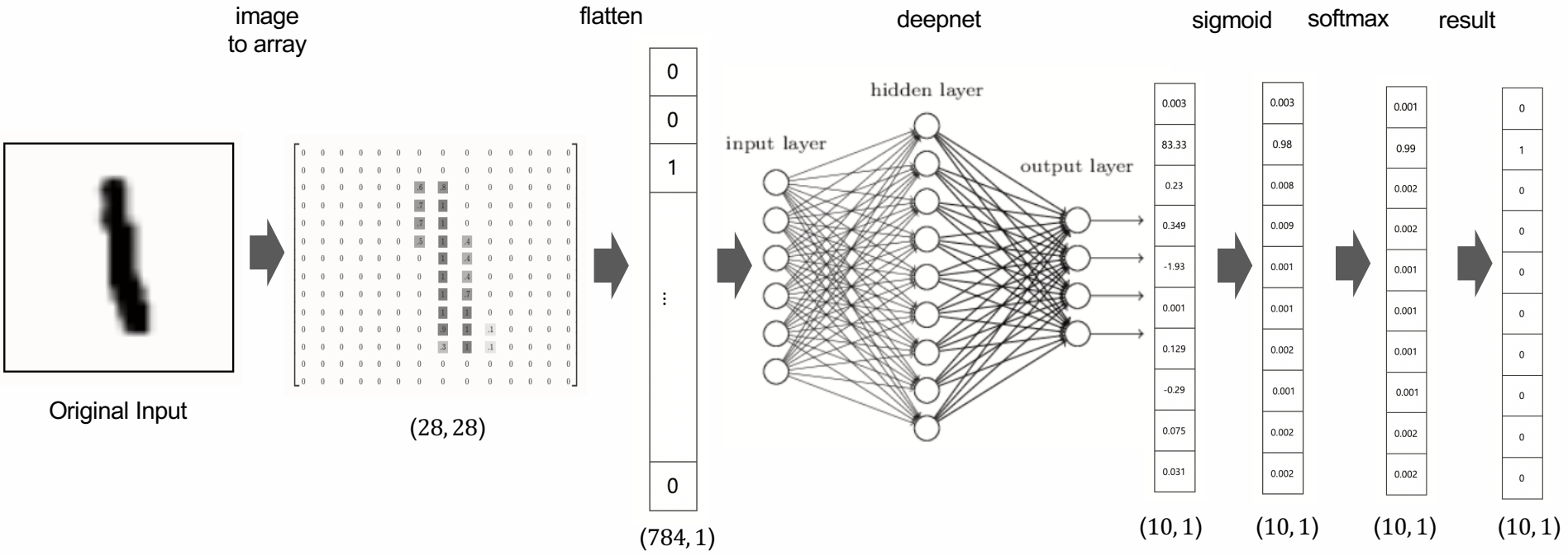


image
to array
flatten
deepnet
sigmoid
softmax
result
Original Input
(28, 28)
0
0
1
0
(784, 1)
input layer
hidden layer
output layer
0.003
83.33
0.23
0.349
-1.93
0.001
0.129
-0.29
0.075
0.031
(10, 1)
0.003
0.98
0.008
0.009
0.001
0.001
0.002
0.001
0.002
0.002
(10, 1)
0.001
0.99
0.002
0.002
0.001
0.001
0.001
0.001
0.002
0.002
(10, 1)
0
1
0
0
0
0
0
0
0
0
(10, 1)

Summary

$$P(y=j \mid \Theta^{(i)}) = \frac{e^{\Theta^{(i)}}}{\sum_{j=0}^{k} e^{\Theta_k^{(i)}}}$$

Softmax function

where $\Theta = w_0 x_0 + w_1 x_1 + \ldots + w_k x_k = \sum_{i=1}^{k} w_i x_i = w^T x$