学習理論の練習　1

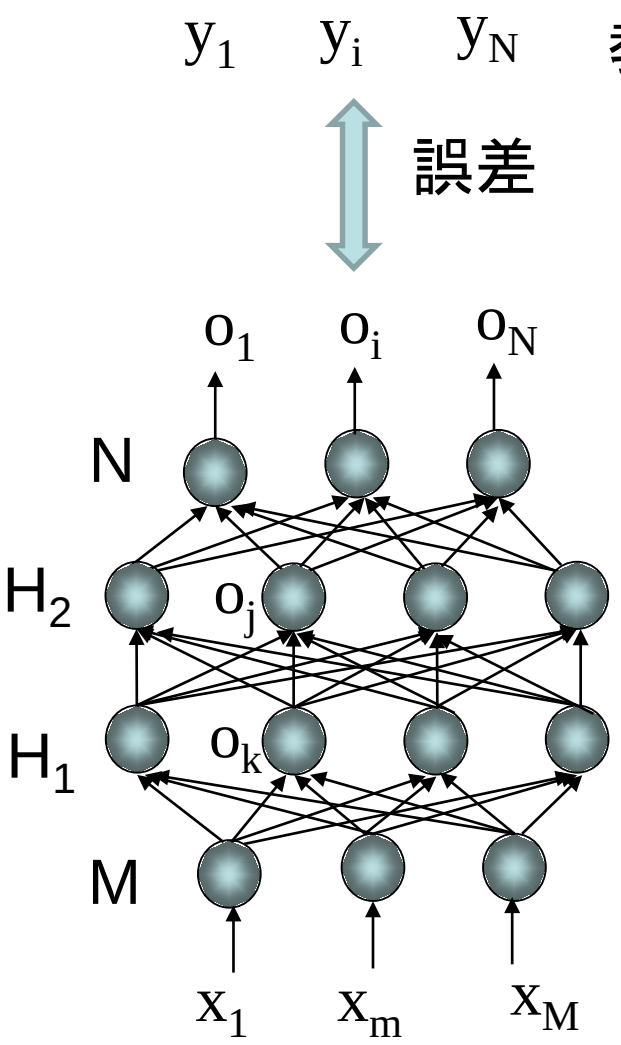
データ

学習用のデータ　この組が n 個ある

x_1	x_m	x_M	入力 x ：M次元
y_1	y_i	y_N	教師データ y : N次元

テスト用のデータ　学習データとは異なる組が n’ 個ある

教師あり学習



神経回路網の答えと
正しい答えの誤差は

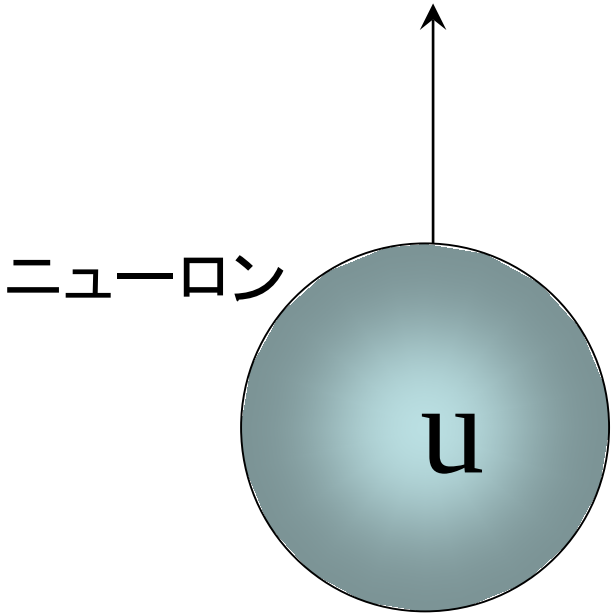
$\Sigma (y_i - o_i(x,w))^2$

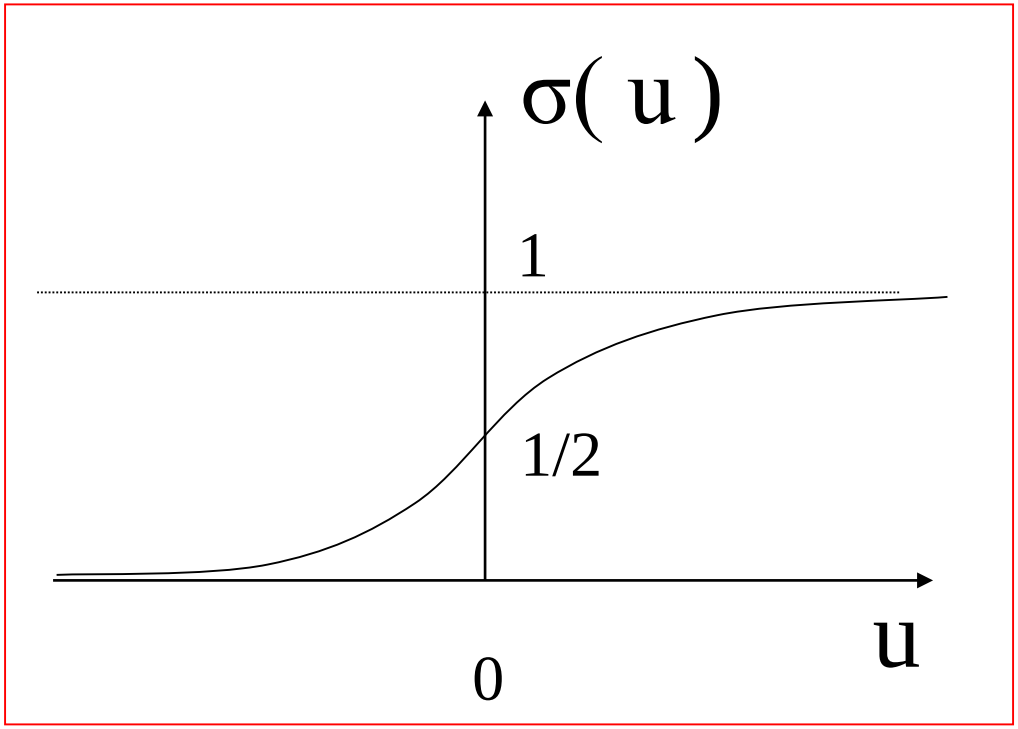
結合
荷重 w

誤差が小さくなるように
w を変えていく

非線型応答:

$$\sigma(u) = 1/(1+e^{-u})$$: シグモイド関数





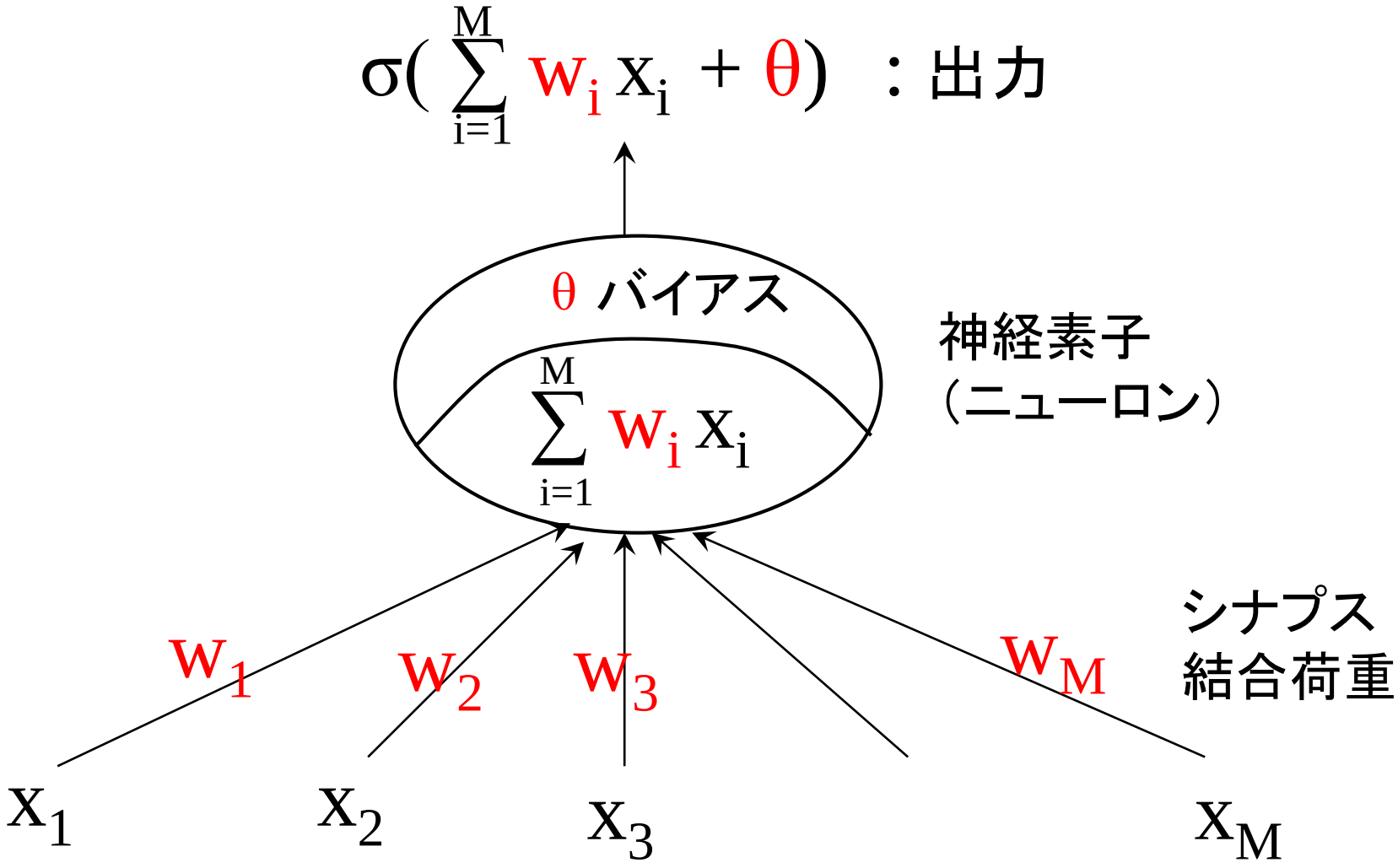
（注意）この形の関数に数学的な意味があるかどうかはわかっていない。最近ではReLu関数なども。

神経素子

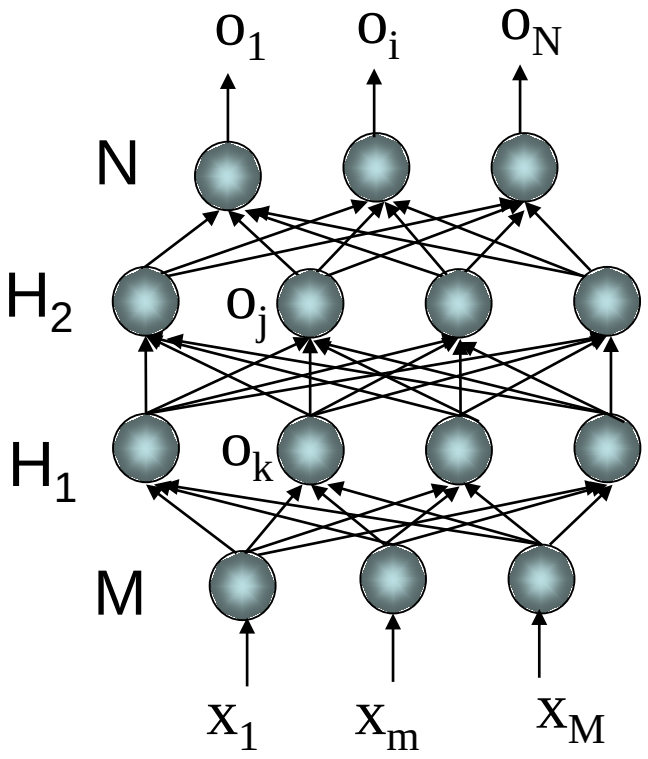
1個の Neuron のモデル

$(x_1,\ldots,x_M)$: 外界からの入力

$(w_1,\ldots w_M, \theta)$: パラメータ



深層学習のネットワーク



中間層2から出力へ

$$o_i = \sigma\left(\sum_{j=1}^{H_2} w_{ij} o_j + \theta_i\right)$$

中間層1から中間層2へ

$$o_j = \sigma\left(\sum_{k=1}^{H_1} w_{jk} o_k + \theta_j\right)$$

入力から中間層1へ

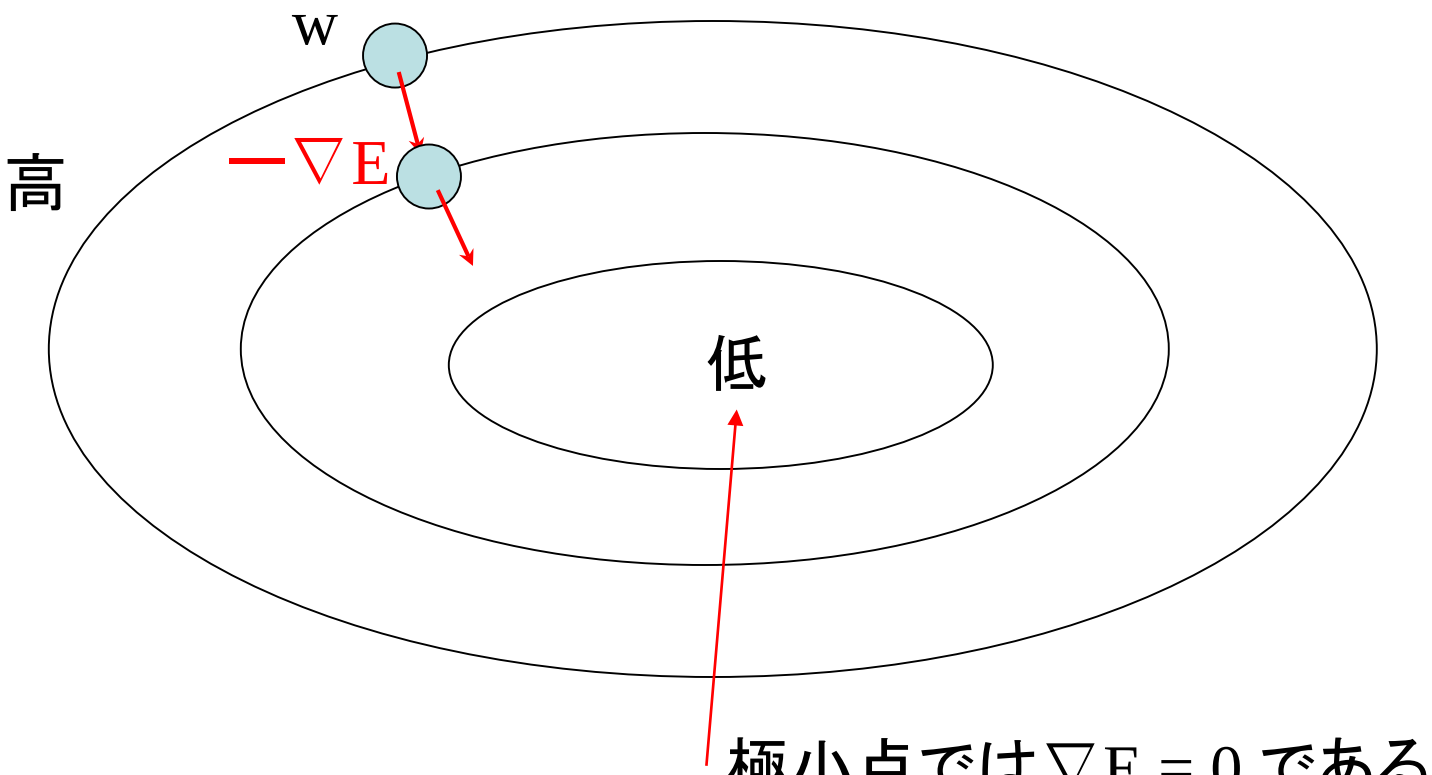
$$o_k = \sigma\left(\sum_{m=1}^{M} w_{km} x_m + \theta_k\right)$$

パラメータ w_{ij}, w_{jk}, w_{km} , θ_k , θ_j , θ_i を学習により最適化する

復習

$$\nabla E(w) = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \cdots\cdots, \frac{\partial E}{\partial w_d} \right)$$

$\nabla E = \mathrm{grad}\, E$ とも書く



最急降下法

学習データの二乗誤差の和を $E(w)$ とする

常微分方程式

$$\frac{dw}{dt} = -\nabla E(w)$$

離散化　$t=0,1,2,3,\ldots$　　$\eta>0$ は定数

$$w(t+1) - w(t) = -\eta \nabla E(w(t))$$

最急降下法　→　確率降下法、ミニバッチなど

確率降下法：
学習用のデータをランダムに一個ずつとってきて
そのデータについての際急降下を行う。
　$\eta>0$ を少しずつ小さくしていく。

ミニバッチ：
一個ずつではなく複数個ずつにする。

最急降下法を改善する研究はたくさんある。Adam, adgrad,…
(目標1)　学習が早くなる
(目標2)　未知のデータについてより当たる

誤差逆伝播法

(3) 中間層2から出力へ

$$o_i = \sigma\left(\sum_{j=1}^{H_2} w_{ij} o_j + \theta_i\right)$$

(2) 中間層1から2へ

$$o_j = \sigma\left(\sum_{k=1}^{H_1} w_{jk} o_k + \theta_j\right)$$

(1) 入力から中間層1へ

$$o_k = \sigma\left(\sum_{m=1}^{M} w_{km} x_m + \theta_k\right)$$

(4) 中間層2から出力へ

$$\delta_i = (o_i - y_i)\, o_i\, (1 - o_i)$$

$$\Delta w_{ij} = -\eta(t)\, \delta_i\, o_j$$

$$\Delta \theta_i = -\eta(t)\, \delta_i$$

(5) 中間層1から2へ

$$\delta_j = \Sigma_i\, \delta_i\, w_{ij} o_j (1 - o_j)$$

$$\Delta w_{jk} = -\eta(t)\, \delta_j\, o_k$$

$$\Delta \theta_j = -\eta(t)\, \delta_j$$

(6) 入力から中間層1へ

$$\delta_k = \Sigma_j\, \delta_j\, w_{jk} o_k (1 - o_k)$$

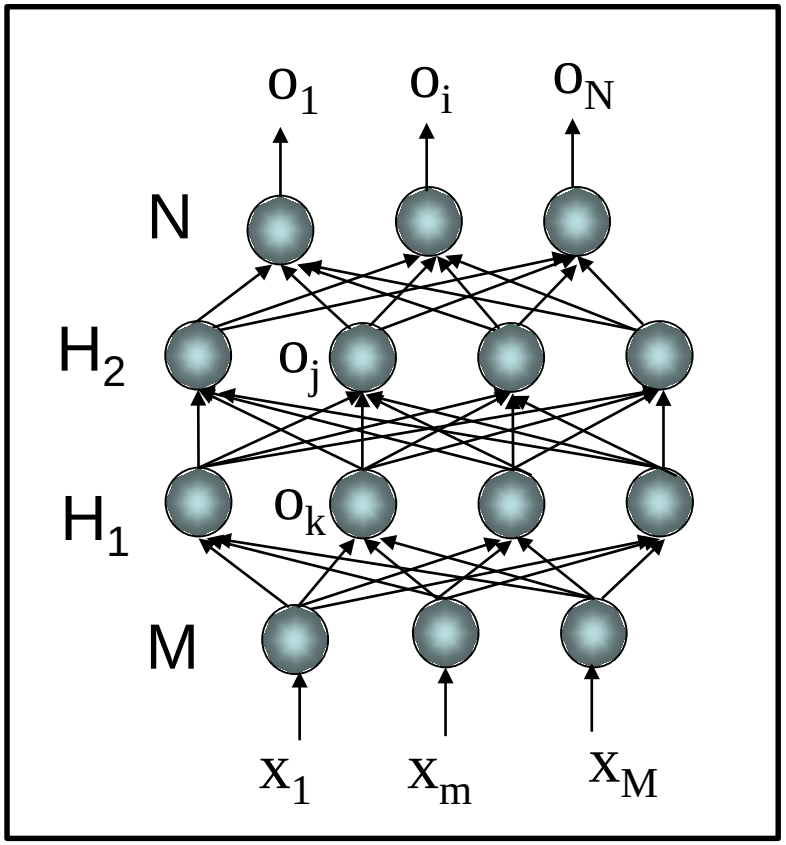
$$\Delta w_{km} = -\eta(t)\, \delta_k x_m$$

$$\Delta \theta_k = -\eta(t)\, \delta_k$$

ミニバッチを実装してみる

ミニバッチでは
k 個のデータについて
並列処理ができる。

n : データの数
$x : M \times n$, $y : N \times n$



k : ミニバッチデータの数

$x : M \times k$

$y : N \times k$

$o^1, o^2, o^3 : H_1 \times k, H_2 \times k, N \times k$

$\delta^1, \delta^2, \delta^3 : H_1 \times k, H_2 \times k, N \times k$

$\theta^1, \theta^2, \theta^3 : H_1 \times 1, H_2 \times 1, N \times 1$

$w^1, w^2, w^3 : H_1 \times M, H_2 \times H_1, N \times H_2$

行列で書く

積は行列としての積
(A)’ は A の転置
[x . y] は要素毎の積
σ(x) は x の要素毎の関数
I_k は 1×kで要素がすべて1

(3) 中間層2から出力へ

$$o^3=\sigma(w^3o^2+\theta^3 I_k)$$

(2) 中間層1から2へ

$$o^2=\sigma(w^2o^1+\theta^2 I_k)$$

(1) 入力から中間層1へ

$$o^1=\sigma(w^1x+\theta^1 I_k)$$

(4) 中間層2から出力へ

$$\delta^3= (o^3-y).\, o^3.\,(1- o^3)$$

$$\Delta w^3 = -\eta(t)\, \delta^3\, (o^2)'$$

$$\Delta \theta^3 = -\eta(t)\, \delta^3\, (I_k)'$$

(5) 中間層1から2へ

$$\delta^2 = ((w^3)'\, \delta^3).o^2.(1-o^2)$$

$$\Delta w^2 = -\eta(t)\, \delta^2\, (o^1)'$$

$$\Delta \theta^2 = -\eta(t)\, \delta^2\, (I_k)'$$

(6) 入力から中間層1へ

$$\delta^1 = ((w^2)'\, \delta^2).o^1.(1-o^1)$$

$$\Delta w^1 = -\eta(t)\, \delta^1\, (x)'$$

$$\Delta \theta^1 = -\eta(t)\, \delta^1\, (I_k)'$$

学習誤差とテスト誤差

ネットワーク出力　$o = (o_i)$,　教師出力　$y = (y_i)$.

1個の出力の二乗誤差　$E(w) = \frac{1}{2} \sum_{i=1}^{N} (o_i - y_i)^2$

(N = 出力の次元)

学習誤差 = 学習に使ったデータについての
二乗誤差の平均

テスト誤差 = テストに使ったデータについての
二乗誤差の平均

誤差逆伝播法の導出 (1)

ネットワーク出力 $o = (o_i)$, 教師出力 $y = (y_i)$.

(N = 出力の次元)

二乗誤差 $E(w) = \frac{1}{2} \sum_{i=1}^{N} (o_i - y_i)^2$

(1) 中間層2から出力への学習

$$\frac{\partial E}{\partial w_{ij}} = (o_i - y_i) \frac{\partial o_i}{\partial w_{ij}}$$

$$= \boxed{(o_i - y_i)\, o_i (1 - o_i)}\, o_j$$

$$= \delta_i$$

中間層2から出力へ

$$o_i = \sigma\left(\sum_{j=1}^{H_2} w_{ij} o_j + \theta_i\right)$$

中間層1から中間層2へ

$$o_j = \sigma\left(\sum_{k=1}^{H_1} w_{jk} o_k + \theta_j\right)$$

入力から中間層1へ

$$o_k = \sigma\left(\sum_{m=1}^{M} w_{km} x_m + \theta_k\right)$$

誤差逆伝播法の導出 (2)

$$E(w) = \frac{1}{2}\sum_{i=1}^{N}(o_i - y_i)^2$$

(2) 中間層1から中間層2へ

$$\frac{\partial E}{\partial w_{jk}} = \sum_{i=1}^{N}(o_i - y_i)\frac{\partial o_i}{\partial o_j}\frac{\partial o_j}{\partial w_{jk}}$$

$$\begin{cases} \dfrac{\partial o_i}{\partial o_j} = o_i(1-o_i)w_{ij} \\ \dfrac{\partial o_j}{\partial w_{jk}} = o_j(1-o_j)x_k \end{cases}$$

$$= \sum_{i=1}^{N}(o_i - y_i)\, o_i(1-o_i)w_{ij}\, o_j(1-o_j)o_k$$

$$= \delta_j = \Sigma_i \delta_i\, w_{ij} o_j(1-o_j)$$

中間層2から出力へ

$$o_i = \sigma\left(\sum_{j=1}^{H_2} w_{ij}o_j + \theta_i\right)$$

中間層1から中間層2へ

$$o_j = \sigma\left(\sum_{k=1}^{H_1} w_{jk}o_k + \theta_j\right)$$

入力から中間層1へ

$$o_k = \sigma\left(\sum_{m=1}^{M} w_{km}x_m + \theta_k\right)$$

誤差逆伝播法の導出 (3)

$$E(w) = \frac{1}{2}\sum_{i=1}^{N}(o_i - y_i)^2$$

(3) 入力から中間層1へ

$$\frac{\partial E}{\partial w_{km}} = \sum_{i=1}^{N}\sum_{j=1}^{H_2}(o_i - y_i)\frac{\partial o_i}{\partial o_j}\frac{\partial o_j}{\partial o_k}\frac{\partial o_k}{\partial w_{km}}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{H_2}(o_i - y_i)\, o_i(1-o_i)w_{ij}\, o_j(1-o_j)w_{jk}o_k(1-o_k)x_m$$

$$= \delta_k = \Sigma_j \delta_j\, w_{jk}o_k(1-o_k)$$

中間層2から出力へ

$$o_i = \sigma\left(\sum_{j=1}^{H_2} w_{ij}o_j + \theta_i\right)$$

中間層1から中間層2へ

$$o_j = \sigma\left(\sum_{k=1}^{H_1} w_{jk}o_k + \theta_j\right)$$

入力から中間層1へ

$$o_k = \sigma\left(\sum_{m=1}^{M} w_{km}x_m + \theta_k\right)$$