誤差逆伝播法による ニューラルネットワーク (BackPropagation Neural Network, BPNN)

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

BPNN とは？

✓ニューラルネットワークおよびその学習法の一つ

✓目的変数の誤差が小さくなるように、各ニューロンの重みを最適化

✓深層学習 (ディープラーニング) も基本的には同じ学習方法で可能

- ディープニューラルネットワーク

✓隠れ層の数が多くなると、入力変数 (説明変数) に近くなるにつれて学習が進まなくなるので注意

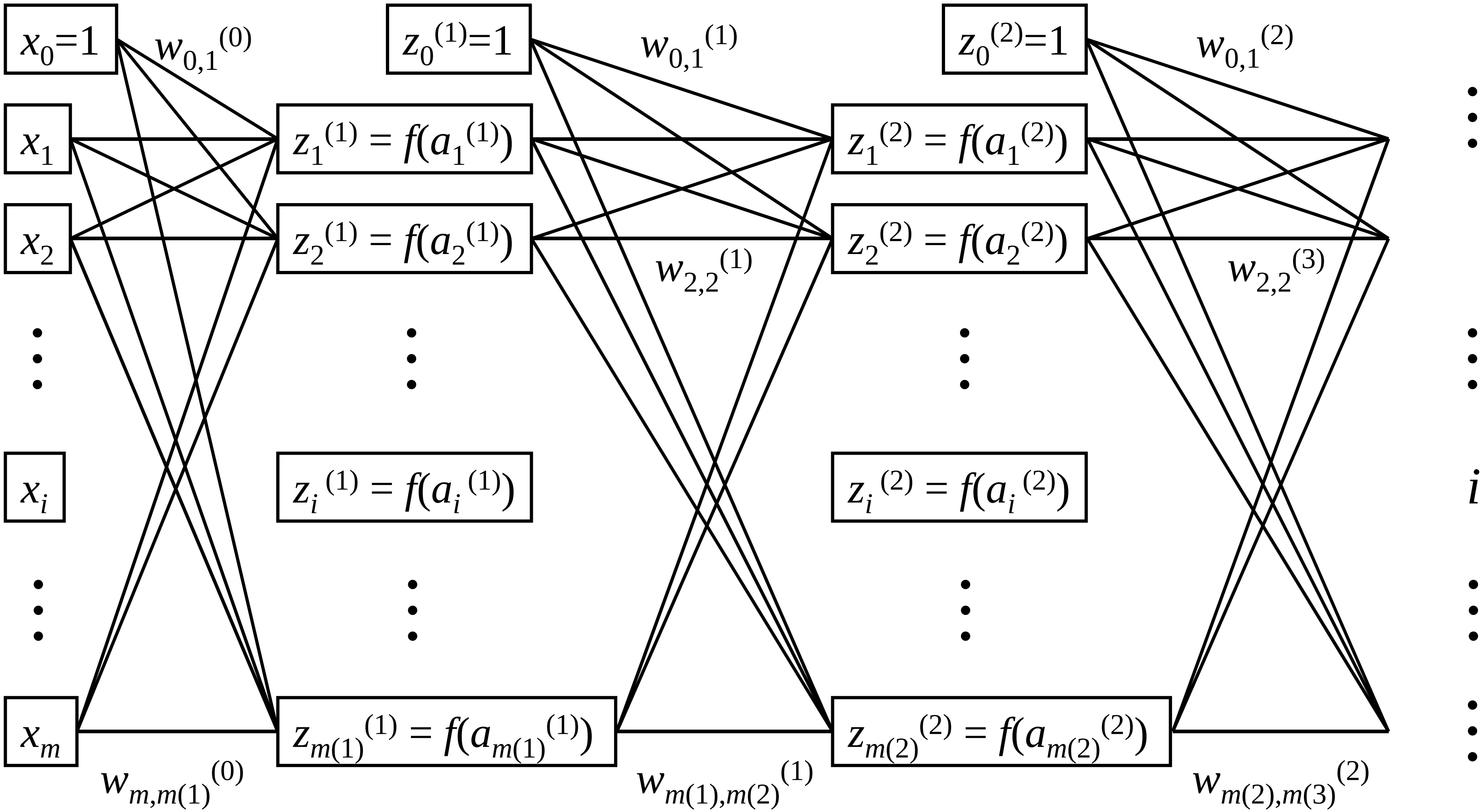
ニューラルネットワークの構造 1/2

入力変数
(説明変数)

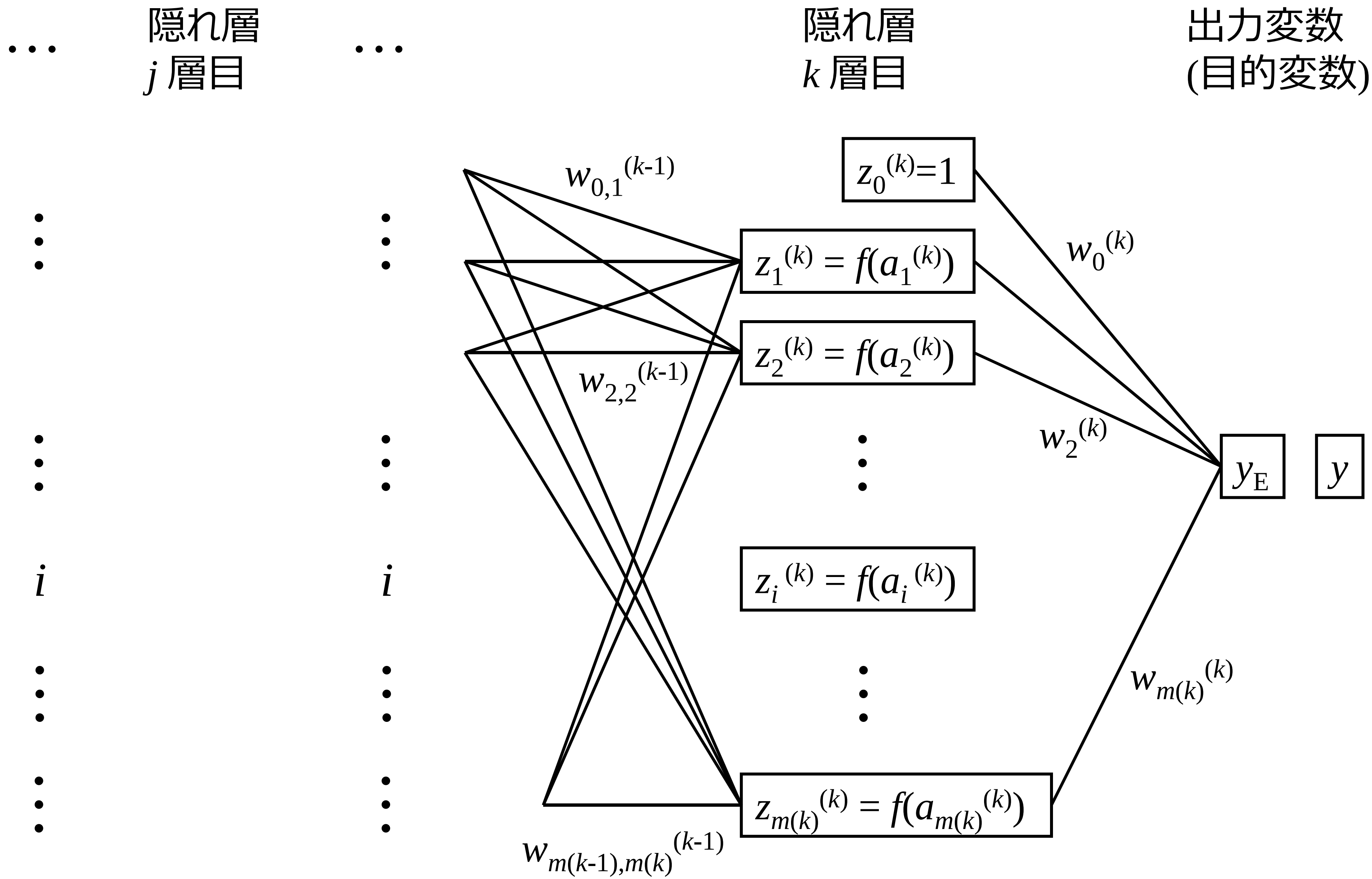
隠れ層
1 層目

隠れ層
2 層目

・・・



ニューラルネットワークの構造 2/2



ニューラルネットワークの構造の補足

- ✓入力変数 (説明変数) と隠れ層 1 層目の間や、隠れ層の前後の間や、隠れ層 k 層目と出力変数 (目的変数) の間の、すべての線 (−) に重み w がある

- ✓それぞれの隠れ層における四角 ($\boxed{z_i^{(j)} = f(a_i^{(j)})}$) をニューロンと呼ぶ

- ✓入力変数 (説明変数) には $x_0 = 1$ が、それぞれの隠れ層にも 0 番目のニューロンとして $z_0^{(j)} = 1$ が、バイアスパラメータとしてある

- ✓隠れ層 j-1 層目の p 番目のニューロンと、隠れ層 j 層目の i 番目のニューロンとの間の重みを、$w_{p,i}^{(j)}$ とする

ニューラルネットワークの構造を式で表す

$$a_i^{(1)} = \sum_{p=1}^{m} w_{p,i}^{(0)} x_p + w_{0,i}^{(0)} = \sum_{p=0}^{m} w_{p,i}^{(0)} x_p \qquad (x_0 = 1)$$

$$a_i^{(j)} = \sum_{p=1}^{m^{(j-1)}} w_{p,i}^{(j-1)} z_p^{(j-1)} + w_{0,i}^{(j-1)} = \sum_{p=0}^{m^{(j-1)}} w_{p,i}^{(j-1)} z_p^{(j-1)} \qquad \left(z_0^{(j-1)} = 1\right)$$

$$z_i^{(j)} = f\left(a_i^{(j)}\right)$$

$$y_{\mathrm{E}} = \sum_{p=1}^{m^{(k)}} w_p^{(k)} z_p^{(k)} + w_0^{(k)} = \sum_{p=0}^{m^{(k)}} w_p^{(k)} z_p^{(k)} \qquad \left(z_0^{(k)} = 1\right)$$

x_i : i 番目の入力変数 (説明変数)
y : 出力変数 (目的変数)
y_{E} : 推定された出力変数 (目的変数)
f : 活性化関数

活性化関数の例 1/2

$$z_i^{(j)} = f\left(a_i^{(j)}\right)$$

シグモイド関数

$$z_i^{(j)} = \frac{1}{1+\exp\left(-a_i^{(j)}\right)}$$

ソフトサイン

$$z_i^{(j)} = \frac{a_i^{(j)}}{1+\left|a_i^{(j)}\right|}$$

ソフトマックス関数

$$z_i^{(j)} = \frac{\exp\left(a_i^{(j)}\right)}{\sum_{p=1}^{m^{(j)}} \exp\left(a_p^{(j)}\right)}$$

活性化関数の例 1/2

$$z_i^{(j)} = f\left(a_i^{(j)}\right)$$

ReLU
(Rectified Linear Unit)

$$z_i^{(j)} = \max\left(0, a_i^{(j)}\right)$$

動径基底関数

$$z_i^{(j)} = \exp\left(-\gamma\left(a_i^{(j)}\right)^2\right)$$

活性化関数についての補足

✓活性化関数によって、ニューラルネットワークが非線形になる

✓活性化関数は微分可能である必要がある (後述)

✓以前はシグモイド関数が多く用いられたが、近年は隠れ層の数 k を多くすることもあり、ReLU やその改良版が用いられることが多い (後述)

✓クラス分類のときは、出力層にソフトマックス関数を用いる

✓動径基底関数を用いたとき、RBF (Radial Basis Function) ネットワークの一つになる

ネットワークを構築するとは？

すべての重み ($w_p^{(k)}$ や $w_{p,i}^{(j)}$) を決めるということ

決める方法の一つが、誤差逆伝播法

誤差逆伝播法 サンプルごとの誤差 E

サンプルごとの誤差 E は　$E = \frac{1}{2}\left(y_{\mathrm{E}} - y\right)^2$

サンプルごとに E が小さくなるように、重み w を変化させていけばよい

⇨ w を微小変化させることでどう E が変化するか、を
求めるため、E を w で微分する

・・・ これにより w を変化させるべき方向 (大きくするか小さくするか) が
求まるので、たとえば確率的勾配降下法 [1] などで重みを更新する

[1] https://ja.wikipedia.org/wiki/確率的勾配降下法

誤差逆伝播法 隠れ層 k 層目から y への重み

$$\frac{\partial E}{\partial w_p^{(k)}} = \frac{\partial E}{\partial y_{\mathrm{E}}} \frac{\partial y_{\mathrm{E}}}{\partial w_p^{(k)}}$$

(連鎖則) ・・・ $\frac{\mathrm{d}}{\mathrm{d}x} \log\left(x^2\right) = \frac{1}{x^2} 2x$

でやっていることと同じ

p. 10 より $\frac{\partial E}{\partial y_{\mathrm{E}}} = y_{\mathrm{E}} - y$

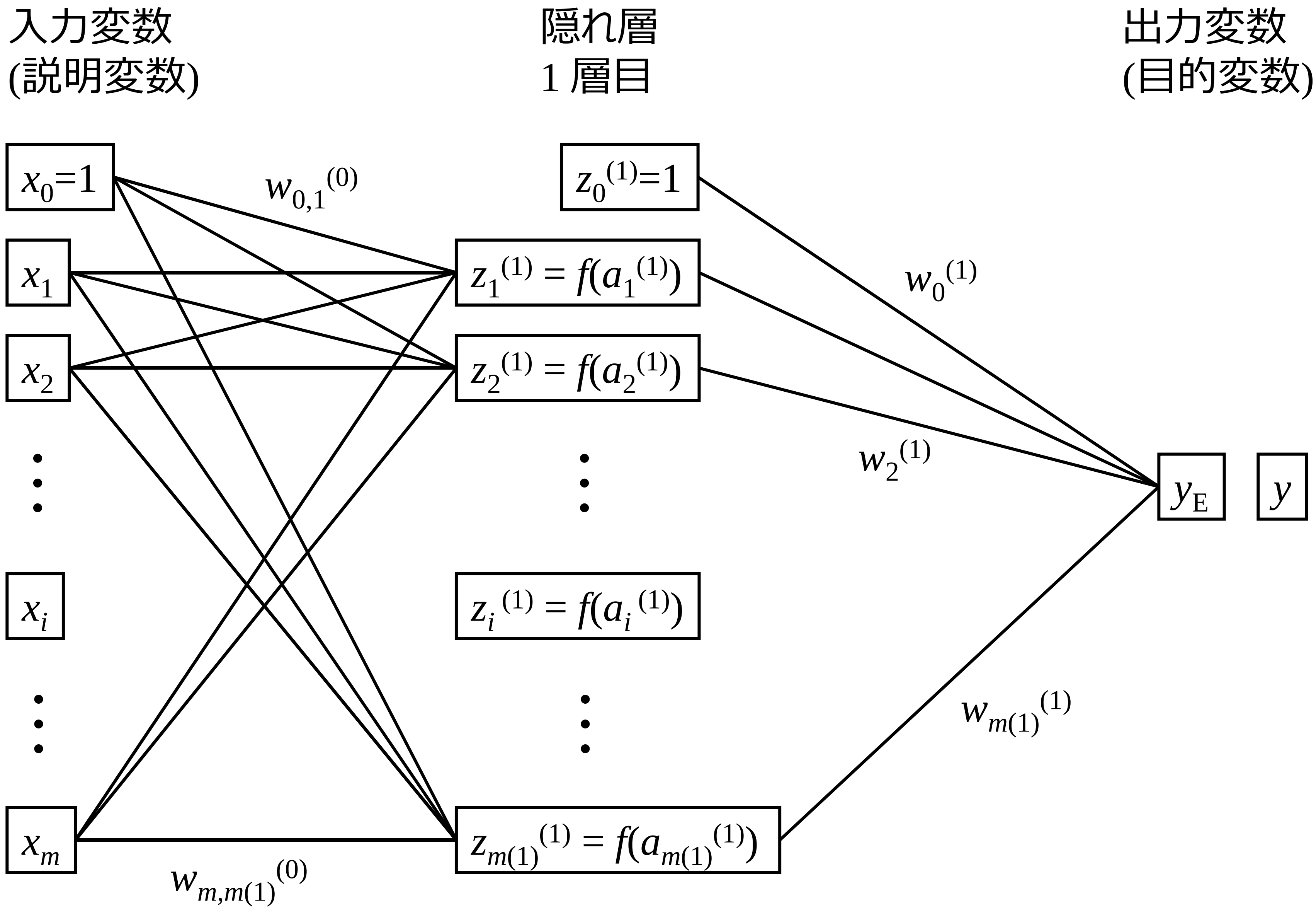
p. 5 の４式目より $\frac{\partial y_{\mathrm{E}}}{\partial w_p^{(k)}} = z_p^{(k)}$

よって、 $\frac{\partial E}{\partial w_p^{(k)}} = \left(y_{\mathrm{E}} - y\right) z_p^{(k)}$

誤差逆伝播法 k = 1 のとき

続いて、隠れ層 j 層目から $j+1$層目への重みを考えたいが、
まずは、最も単純な $k = 1$ つまり隠れ層が 1 層のときを考える

ニューラルネットワークの構造 隠れ層 1 層



誤差逆伝播法 x から隠れ層 1 層目への重み

$$\frac{\partial E}{\partial w_{p,i}^{(0)}} = \frac{\partial E}{\partial a_i^{(1)}} \frac{\partial a_i^{(1)}}{\partial w_{p,i}^{(0)}}$$ (連鎖則)

p. 5 の１式目より $$\frac{\partial a_i^{(1)}}{\partial w_{p,i}^{(0)}} = x_p$$

また、 $$\frac{\partial E}{\partial a_i^{(1)}} = \frac{\partial E}{\partial y_{\mathrm{E}}} \frac{\partial y_{\mathrm{E}}}{\partial a_i^{(1)}} = \left(y_{\mathrm{E}} - y \right) \frac{\partial y_{\mathrm{E}}}{\partial a_i^{(1)}}$$

(連鎖則) (p. 11)

p. 5 の４式目より $$\frac{\partial y_{\mathrm{E}}}{\partial a_i^{(1)}} = w_i^{(1)} \frac{\partial z_i^{(1)}}{\partial a_i^{(1)}}$$

よって、 $$\frac{\partial E}{\partial w_{p,i}^{(0)}} = x_p \left(y_{\mathrm{E}} - y \right) w_i^{(1)} \frac{\partial z_i^{(1)}}{\partial a_i^{(1)}}$$

誤差逆伝播法 活性化関数の微分

$\dfrac{\partial z_i^{(1)}}{\partial a_i^{(1)}}$ は、p. 6, 7 の活性化関数を $a_i^{(j)}$ で微分して、その導関数に

$a_i^{(1)}$ を代入したもの

⇨ 活性化関数は $a_i^{(j)}$ で微分可能である必要がある

誤差逆伝播法 k > 1 のとき 1/3

隠れ層が 2 層以上のとき (ディープニューラルネットワーク)、
隠れ層 j-1 層目から j 層目への重みを考える

$$\frac{\partial E}{\partial w_{p,i}^{(j-1)}} = \frac{\partial E}{\partial a_i^{(j)}} \frac{\partial a_i^{(j)}}{\partial w_{p,i}^{(j-1)}} \qquad \text{(連鎖則)}$$

p. 5 の 2 式目より $$\frac{\partial a_i^{(j)}}{\partial w_{p,i}^{(j-1)}} = z_p^{(j-1)}$$

隠れ層 j 層目の i 番目のニューロンにつながっている、
隠れ層 j+1 層目の $m^{(j+1)}$ 個のニューロンを考えると、

$$\frac{\partial E}{\partial a_i^{(j)}} = \sum_{q=1}^{m^{(j+1)}} \frac{\partial E}{\partial a_q^{(j+1)}} \frac{\partial a_q^{(j+1)}}{\partial a_i^{(j)}}$$

誤差逆伝播法 k > 1 のとき 2/3

$$\frac{\partial E}{\partial a_i^{(j)}} = \sum_{q=1}^{m^{(j+1)}} \frac{\partial E}{\partial a_q^{(j+1)}} \frac{\partial a_q^{(j+1)}}{\partial a_i^{(j)}}$$

p. 5 の 2 式目より　$a_q^{(j+1)} = \sum_{p=0}^{m^{(j-1)}} w_{p,q}^{(j)} z_p^{(j)}$　から、　$\frac{\partial a_q^{(j+1)}}{\partial a_i^{(j)}} = w_{i,q}^{(j)} \frac{\partial z_i^{(j)}}{\partial a_i^{(j)}}$

$\frac{\partial z_i^{(j)}}{\partial a_i^{(j)}}$　は、p. 6, 7 の活性化関数を $a_i^{(j)}$ で微分して、その導関数に

$a_i^{(1)}$ を代入したもの

誤差逆伝播法 k > 1 のとき 3/3

p. 16, 17 をまとめると、 $$\frac{\partial E}{\partial w_{p,i}^{(j-1)}} = z_p^{(j-1)} \sum_{q=1}^{m^{(j+1)}} \frac{\partial E}{\partial a_q^{(j+1)}} w_{i,q}^{(j)} \frac{\partial z_i^{(j)}}{\partial a_i^{(j)}}$$

$\frac{\partial E}{\partial a_q^{(j+1)}}$ について、p. 17 と同様にして、隠れ層 j+1 層目の q 番目の

ニューロンにつながっている、隠れ層 j+2 層目の $m^{(j+2)}$ 個のニューロンを考えることができる

これを繰り返すと、j+2 層目、j+3 層目、・・・となり、最後は y になる

つまり、 $$\frac{\partial E}{\partial y_{\mathrm{E}}} = y_{\mathrm{E}} - y$$

誤差逆伝播法 名前の由来

以上のように、y の誤差 $\mathrm{y_E} - \mathrm{y}$ が、隠れ層 k 層目、k-1 層目、・・・と逆に伝播して、重みの変化に寄与していることから、誤差逆伝播法と呼ぶ

誤差逆伝播法 注意点

y の誤差が伝播するとき、隠れ層 k 層目、k-1 層目、・・・ と入力変数 (説明変数) に近くなるにつれて、値が小さくなってしまうことに注意

p. 18 に活性化関数の微分係数 $\dfrac{\partial z_i^{(j)}}{\partial a_i^{(j)}}$ があるが、たとえば

シグモイド関数の微分係数の最大値は 0.25 であり、隠れ層の層が深くなるにつれて、最大でも 0.25^j と指数関数的に小さくなってしまう

⇒ 重み w が変化しなくなってしまう

そこで、特に隠れ層の数 k を多くするときには、微分係数が 1 になる ReLU が使われる