Back Propagation
誤差逆伝播則– Chapter.4

Kato Sotaro
2020/04/28

04. Agenda

1. 勾配計算の難しさ
2. 二層ネットワークでの計算
3. 多層ネットワークへの一般化

04. Agenda

1. 勾配計算の難しさ
2. 二層ネットワークでの計算
3. 多層ネットワークへの一般化

04.1 勾配計算の難しさ

要約：勾配降下法による誤差関数の勾配成分

- （前回の復習）勾配降下法.
  - 誤差関数Eの（局所的な）最小値を求めるための手法.

$$W^{t+1} = W^t - \varepsilon \nabla E \qquad \varepsilon\text{:学習係数, W:重み}$$

  - 実質的には重みによる偏微分

$$\nabla E = \frac{\partial E(w)}{\partial W}$$

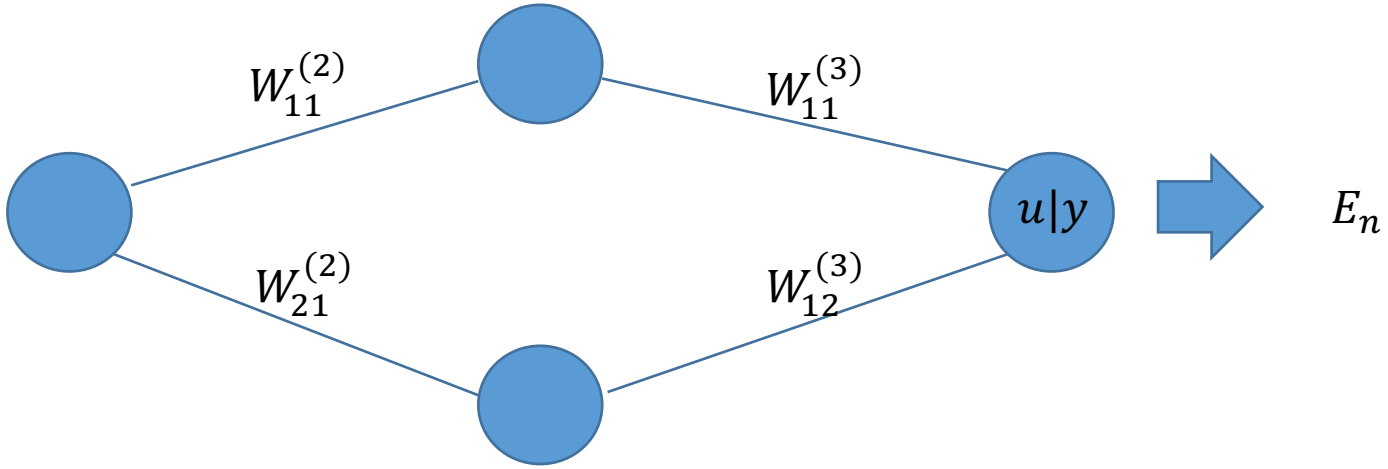
  - ∇Eの各成分

$$\nabla E \rightarrow \frac{\partial E}{\partial W_{ji}}\ and \frac{\partial E}{\partial b} \qquad b\text{:バイアス}$$

04.1 勾配計算の難しさ

要約：$\frac{\partial E}{\partial W_{ji}}$の計算が面倒な理由

- 実際に、勾配∇Eを求めるために二層ネットワークでの誤差関数を偏微分する



- ∇Eを求めるにはE_nを重みで偏微分すればいい
  →$E_n = \frac{1}{2}\sum_{n=1}^{n} |d_n - y(x_n - W)|^2$ (復習)を用いて重みで微分する.

- 例えば$W_{11}^{(2)}$でEを偏微分すると

$$\nabla E = \frac{\partial E}{\partial W_{11}^{(2)}} = (y(x) - d)^T \boxed{\frac{\partial y}{\partial W_{11}^{(2)}}}$$

← この計算が面倒

04.1 勾配計算の難しさ

要約：$\frac{\partial E}{\partial W_{ji}}$の計算が面倒な理由

- yの数式

$$
\begin{aligned}
y(x) &= f\left(u^{(3)}\right) \\
&= f\left(W^{(3)}Z^{(2)} + b^3\right) \\
&= f\left(W^{(3)}f\left(W^{(2)}Z^{(1)} + b^2\right) + b^3\right)
\end{aligned}
$$

入力値

→ $W_{11}^{(2)}$で微分するときに、3回も微分しなければならない(微分の連鎖規則).

- 微分の連鎖規則
  - yの微分は合成関数の微分の考えと似ている.
  - 例えば以下の微分を考える.

$$y = \log(\sin^2 x)$$

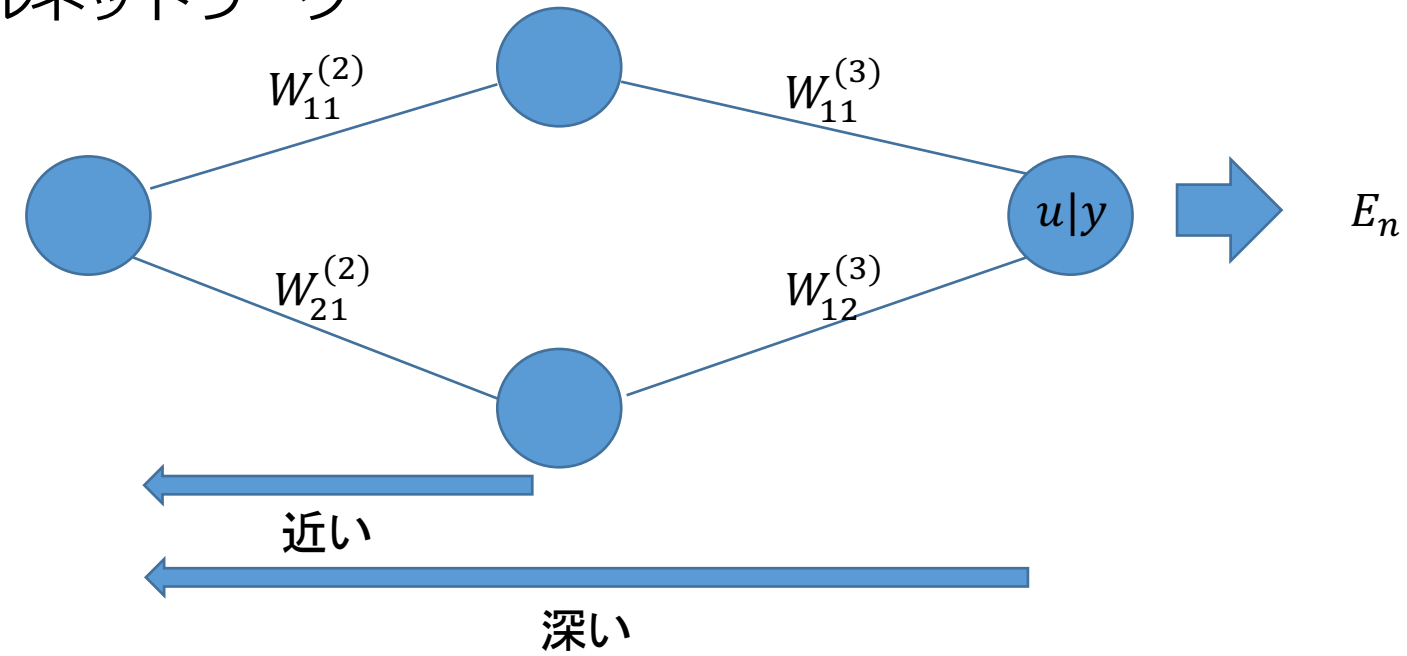
$$\frac{dy}{dx} = \frac{(\sin^2 x)'}{\sin^2 x}$$

→xの微分ができるまで微分する.

04.1 勾配計算の難しさ

要約：$\frac{\partial E}{\partial W_{ji}}$の計算が面倒な理由

- 2層ニューラルネットワーク



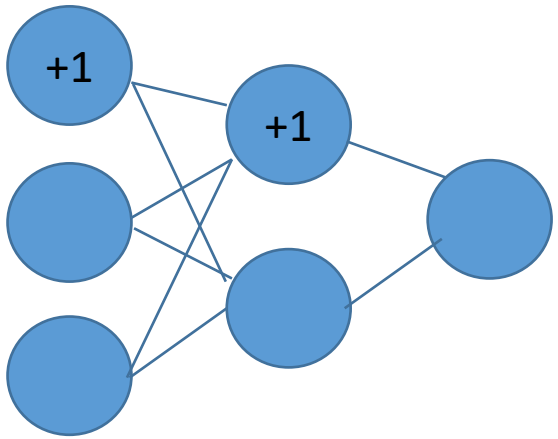
- 深い層で、かつ入力に近い重みでの微分ほど計算量が多くなる.
- 勾配降下法の∇Eの面倒くさい計算を誤差逆伝播則が解決する.

04.1 勾配計算の難しさ

要約 :表記の簡略化　①:出力uの表記　②:誤差関数Eの表記

- 表記の簡略化① 出力uの表記

$$u^{(l)} = \sum_{i=1}^{n} W_{ji}^{(l)} z^{(l-1)} + b_j$$
$$= \sum_{i=0}^{n} W_{ji}^{(l)} z^{(l-1)}$$



z_1を1にすることで
簡単に記載(右図参照)

- 0層での出力$z^{(l)}$を常に1にすることで出力の式からバイアスを消去

04.1 勾配計算の難しさ

要約 :表記の簡略化　①:出力uの表記　②:誤差関数Eの表記

- 表記の簡略化② 誤差関数Eの表記
  - バッチ学習：

$$E(W) = \sum_{n=1}^{N} E_n(W)$$

  - ミニバッチ学習：

$$E_t(W) = \frac{1}{N_t} \sum_{n \in Dt} E_n(w)$$

誤差関数の微分とは

→誤差関数の総和の微分　言い換えると　誤差関数の微分の総和

$$\frac{\partial E\ (W)}{\partial W} = \frac{\partial E_1(W)}{\partial W} + \frac{\partial E_2(W)}{\partial W} + \frac{\partial E_3(W)}{\partial W} + \cdots + \frac{\partial E_n(W)}{\partial W}$$

１つのサンプルnに関して微分できればいい.

$$\frac{\partial E_n(W)}{\partial W} \rightarrow \frac{\partial E\ (W)}{\partial W}$$

04. Agenda

1 勾配計算の難しさ

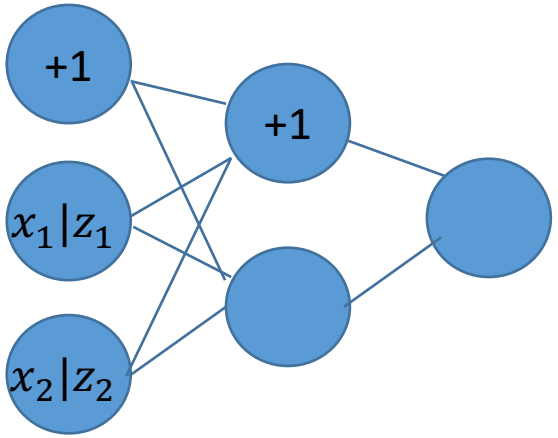
2 二層ネットワークでの計算

3 多層ネットワークへの一般化

04.2 二層ネットワークでの計算

要約 :勾配を計算する上で$\frac{\partial E_n}{\partial w_{ji}^{(2)}}$と$\frac{\partial E_n}{\partial w_{ji}^{(3)}}$の計算に必要な前提知識を紹介.
二層ネットワークの出力の前提条件を提示し中間層の出力zの式とyの出力の式を表した.

- 二層ネットワーク



・考える上での条件
1.出力層の活性化関数を恒等写像
$y = u^{(l)} = \sum_{i=1}^{n} W_{ji}^{(l)} z^{(l-1)} + b_j$
2.中間層ユニットは任意の活性化関数
3.入力層の出力を$z_i^{(1)} = x_i$

・入力が中間層に向けて次のように伝播.

$$z_j^{(2)} = f\left(u_j^{(2)}\right) = f\left(\sum_i w_{ji}^{(2)} z_i^{(1)}\right)$$

・出力層の活性化関数は恒等写像なので出力層の式は次のようになる.

$$y_j(x) = z_j^{(3)} = u_j^{(3)} = \sum_i w_{ji}^{(3)} z_i^{(2)}$$

重みの微分$\frac{\partial E_n}{\partial w_{ji}^{(2)}}$と$\frac{\partial E_n}{\partial w_{ji}^{(3)}}$はどのように計算されるのかを見ていく

04.2 二層ネットワークでの計算

要約：$\frac{\partial E_n}{\partial w_{ji}^{(3)}}$の計算

- $\frac{\partial E_n}{\partial w_{ji}^{(3)}}$の計算は$\frac{\partial E_n}{\partial w_{ji}^{(2)}}$に比べて簡単にできる.

  - 誤差関数$E_n = \frac{1}{2}|y(x) - d_n|^2$を用いて$\frac{\partial E_n}{\partial w_{ji}^{(3)}}$を計算.

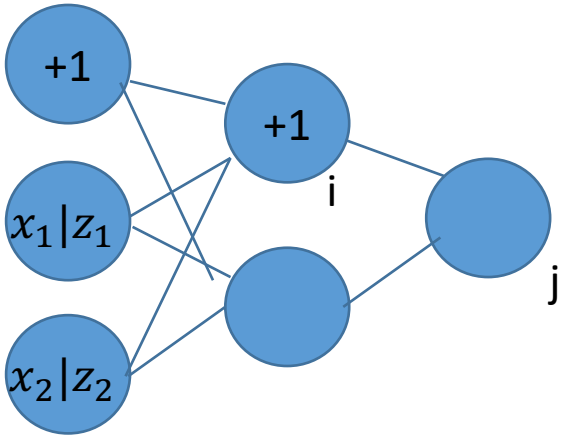
$$\frac{\partial E_n}{\partial w_{ji}^{(3)}} = (y(x) - d)^T \frac{\partial y}{\partial w_{ji}^{(3)}}$$

  - $y_j(x) = \sum_i w_{ji}^{(3)} z_i^{(2)}$よりj成分のみが$z_i^{(2)}$でそれ以外は0.

$$\frac{\partial y}{\partial w_{ji}^{(3)}} = [0 \dots 0 \ z_i^{(2)} \ 0 \ \dots 0]^T$$

  - よって

$$\frac{\partial E_n}{\partial w_{ji}^{(3)}} = (y(x) - d)^T \ z_i^{(2)}$$



04.2 二層ネットワークでの計算

要約 :中間層の重みで偏微分した微分$\frac{\partial E_n}{\partial w_{ji}^{(2)}}$の計算.

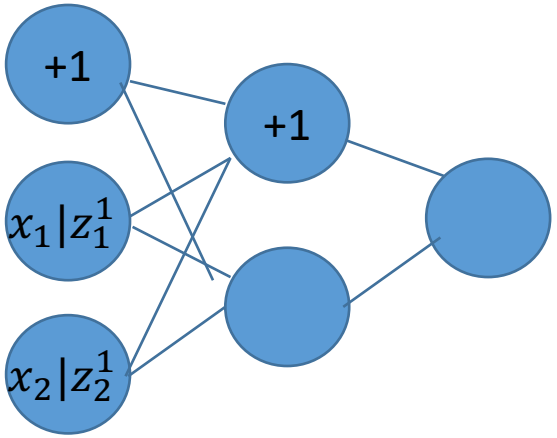
- 微分$\frac{\partial E_n}{\partial w_{ji}^{(2)}}$の計算.
  - $w_{ji}^{(2)}$は$u_j^{(2)}$の中に存在するので連鎖規則を使うと.

$$\frac{\partial E_n}{\partial w_{ji}^{(2)}} = \frac{\partial E_n}{\partial u_j^{(2)}} \frac{\partial u_j^{(2)}}{\partial w_{ji}^{(2)}} \quad \cdots(4.5)$$

  - $\frac{\partial u_j^{(2)}}{\partial w_{ji}^{(2)}}$は$u_j^{(2)} = \sum_i w_{ji}^{(2)} z_i^{(1)}$を用いると.

$$\frac{\partial u_j^{(2)}}{\partial w_{ji}^{(2)}} = z_i^{(1)}$$

  - $\frac{\partial E_n}{\partial u_j^{(2)}}$の計算は次のページへ



04.2 二層ネットワークでの計算

要約 :中間層の重みについての微分$\frac{\partial E_n}{\partial w_{ji}^{(2)}}$の計算

- 微分$\frac{\partial E_n}{\partial u_j^{(2)}}$の計算.
  - 中間層の入力$u_j^{(2)}$はそのユニットと繋がりのある出力層のユニットに伝わるので総入力$u_k^{(3)}$に影響する.

$$\frac{\partial E_n}{\partial u_j^{(2)}} = \sum_k \frac{\partial E_n}{\partial u_k^{(3)}} \frac{\partial u_k^{(3)}}{\partial u_j^{(2)}}$$

  - 右辺の$\frac{\partial E_n}{\partial u_k^{(3)}}$は $E_n = \frac{1}{2}\sum_k (y_k(x) - d_k)^2 = \frac{1}{2}\sum_k (u_k^{(3)} - d_k)^2$から

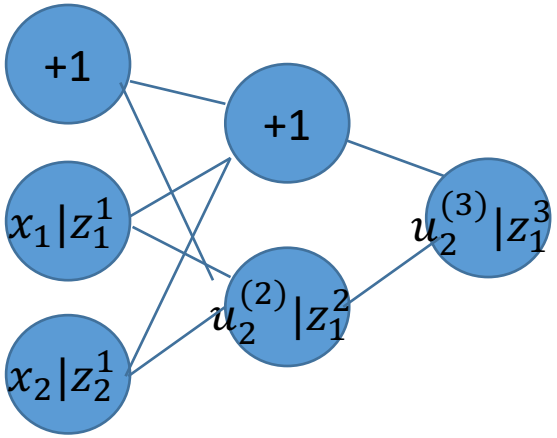
$$\frac{\partial E_n}{\partial u_k^{(3)}} = u_k^{(3)} - d_k$$

  - 右辺の$\frac{\partial u_k^{(3)}}{\partial u_j^{(2)}}$は $u_k^{(3)} = \sum_j (w_{kj}^{(3)} f\,(u_j^{(2)}))$であることから

$$\frac{\partial u_k^{(3)}}{\partial u_j^{(2)}} = w_{kj}^{(3)} f'(u_j^{(2)})$$

これらの式を式*(4.5)*に代入すると

$$\frac{\partial E_n}{\partial w_{ji}^{(2)}} = \left(f'(u_j^{(2)}) \sum_k w_{kj}^{(3)} (u_k^{(3)} - d_k)\right) z_i^{(1)}$$



04. Agenda

1 勾配計算の難しさ

2 二層ネットワークでの計算

3 多層ネットワークへの一般化

04.3 多層ネットワークへの一般化

要約：二層ネットワークの中間層に関する重みの微分を一般化して計算する.

- 二層ネットワークの中間層の重みの式を第l層の式をしてみると(一般化)

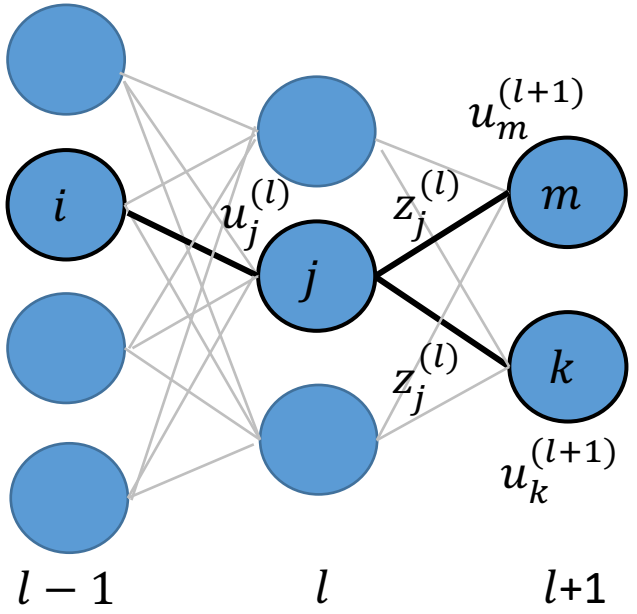
$$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \frac{\partial E_n}{\partial u_j^{(l)}} \frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}}$$

- 右図から$u_j^{(l)}$の変動がE_nに与える影響は出力$z_j^{(l)}$を通じて総入力$u_k^{(l+1)}$を変化させるとわかるので.

$$\frac{\partial E_n}{\partial u_j^{(l)}} = \sum_k \frac{\partial E_n}{\partial u_k^{(l+1)}} \frac{\partial u_k^{(l+1)}}{\partial u_j^{(l)}}$$

- $\frac{\partial E_n}{\partial u_j^{(l)}}$を$\delta_j^{(l)}$:デルタ(delta)とおくと

$$\delta_j^{(l)} = \sum_k \delta_j^{(l+1)} \frac{\partial u_k^{(l+1)}}{\partial u_j^{(l)}}$$



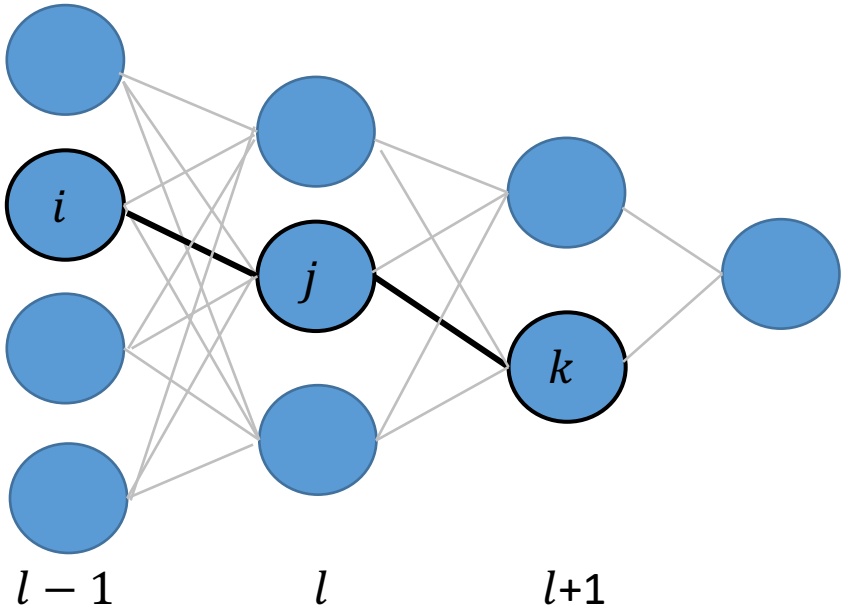
- $u_k^{(l+1)} = \sum_j w_{kj}^{(l+1)} z_j^{(l)} = \sum_j w_{kj}^{(l+1)} f(u_j^{(l)})$より、$\frac{\partial u_k^{(l+1)}}{\partial u_j^{(l)}} = w_{kj}^{(l+1)} f'(u_j^{(l)})$を用いて

$$\delta_j^{(l)} = \sum_k \delta_j^{(l+1)} (w_{kj}^{(l+1)} f'\left(u_j^{(l)}\right)) \quad \cdots(4.12)$$ →この式が意味するのは(次ページ)

04.3 多層ネットワークへの一般化

要約：$\delta_j^{(l)} = \sum_k \delta_j^{(l+1)} (w_{kj}^{(l+1)} f'\left(u_j^{(l)}\right))$ は、 $\delta_j^{(l)}$ が$\delta_j^{(l+1)}$から計算できることを意味する.l層のユニットのデルタは
$l+1$層のユニットのデルタから求められ、 $\delta_j^{(l+2)}$ $\delta_j^{(l+3)}$…と出力層から計算できる.（誤差逆伝播法の由来）

- $\delta_j^{(l)} = \sum_k \delta_j^{(l+1)} (w_{kj}^{(l+1)} f'\left(u_j^{(l)}\right))$が意味するのは $\delta_j^{(l)}$ が$\delta_j^{(l+1)}$から計算できること.



- $\delta_j^{(l)}$ が$\delta_j^{(l+1)}$から計算できるということは、最初に出力層の各ユニットのデルタが求まっていれば任意の層のデルタを求めることができる.

  →デルタは出力層から入力層の向きに伝播していくことから誤差逆伝播法と言われる.

04.3 多層ネットワークへの一般化

要約：$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \frac{\partial E_n}{\partial u_j^{(l)}} \frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}}$の第二項$\frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}}$はユニット$i$からの出力$z_i^{(l-1)}$のただの積に変形できる.よって$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} z_i^{(l-1)}$とまとめられる.

- $\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \frac{\partial E_n}{\partial u_j^{(l)}} \frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}}$の第二項$\frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}}$は、 $u_i^{(l)} = \sum_i w_{ji}^{(l)} z_i^{(l-1)}$を用いることで

$$\frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}} = z_i^{(l-1)}$$

- したがって、これと$\delta_j^{(l)} = \frac{\partial E_n}{\partial u_j^{(l)}}$を用いて$\frac{\partial E_n}{\partial w_{ji}^{(l)}}$を表すと

$$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} z_i^{(l-1)} \qquad \ldots (4.13)$$

- 以上から誤差関数の勾配は次の手順(次ページ)で求められる.

04.3 多層ネットワークへの一般化

要約 :ある訓練サンプル(x,d)が与えられたとき、サンプルに関する誤差関数の勾配は次の手順で示される.

入力: 訓練サンプルx_nおよび目標出力d_nのペアの1つ.
出力: 誤差関数$E_n(w)$の各層lのパラメータについての微分$\frac{\partial E_n}{\partial w_{ji}^{(l)}}(l = 2, \dots, L)$.

1. $z^{(1)} = x_n$とし、各層$l(l = 2, \dots, L)$のユニット入出力$u^{(l)}$および$z^{(l)}$を順に計算する.(順伝播)
2. 出力層でのデルタ$\delta_j^{(L)}$を求める.
3. 中間層l $(l = L-1, L-2, \dots, 2)$での$\delta_j^{(l)}$をこの順に式(4.12)にしたがって計算する.
4. 各層l $(l = 2, \dots, L)$のパラメータ$w_{ji}^{(l)}$に関する微分を式(4.13)にしたがって計算する.

- $\delta_j^{(L)}$とは、出力層のデルタであり初期値なので$\delta_j^{(l)} = \frac{\partial E_n}{\partial u_j^{(L)}}$と簡単に求めることができる.
- これの具体的な計算は04.4で行う.
- ミニバッチの扱いを含むより詳細な実装も04.4で行う.

Reference

要約：

1. 岡谷貴之,深層学習, 4th,講談社, 2015