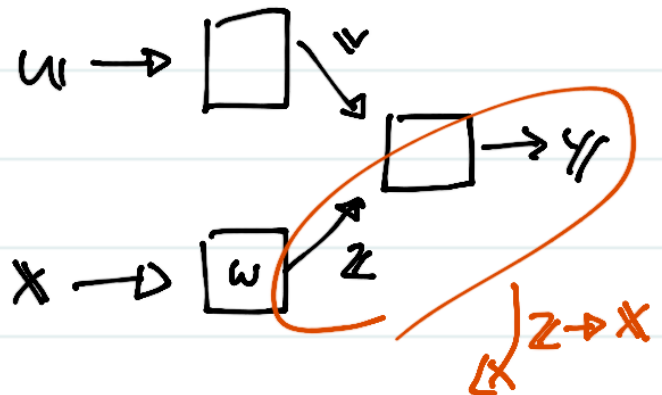


5.3.4 ヤコビ"行列

$$\frac{\partial \mathcal{E}}{\partial w} = \sum_{kj} \frac{\partial \mathcal{E}}{\partial y_k} \frac{\partial y_k}{\partial z_j} \frac{\partial z_j}{\partial w} \quad (5.71)$$

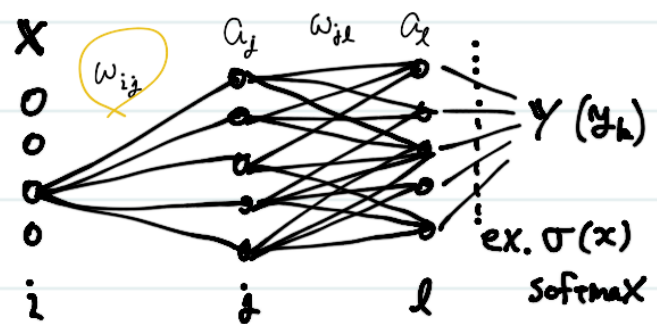
ヤコビ"行列



$$J_{ki} = \frac{\partial y_k}{\partial x_i} = \sum_j \frac{\partial y_k}{\partial a_j} \frac{\partial a_j}{\partial x_i}$$

$$= \sum_j w_{ji} \frac{\partial y_k}{\partial a_j} \quad (5.73)$$

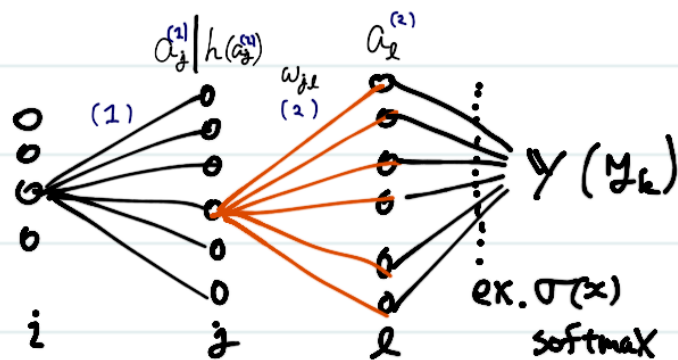
($a_j = \sum_i w_{ji} x_i$)



$$\frac{\partial y_k}{\partial a_j^{(1)}} = \sum_l \frac{\partial y_k}{\partial a_l} \frac{\partial a_l}{\partial a_j^{(1)}}$$

$$= \sum_l \frac{\partial y_k}{\partial a_l^{(2)}} \frac{\partial}{\partial a_j^{(1)}} \left\{ \sum_i w_{il} h(a_i^{(1)}) \right\}$$

$$= h'(a_j^{(1)}) \sum_l w_{lj} \frac{\partial y_k}{\partial a_l^{(2)}} \quad (5.74)$$



ex. 出力がシグモイド $y_k = \sigma(a_k^{(2)})$

$$\frac{\partial y_k}{\partial a_k^{(2)}} = \delta_{kl} \sigma'(a_k)$$

ex. 出力がソフトマックス $y_k = \frac{e^{a_k}}{\sum_i e^{a_i}}$

$$\frac{\partial y_k}{\partial a_k^{(2)}} = \delta_{kl} y_k - y_k y_l$$

$$\frac{\partial}{\partial a_j} \left[\sum_i e^{a_i} \right]^{-1} = - \left[\sum_i e^{a_i} \right]^{-2} \sum_i \delta_{ij} e^{a_i}$$

$$= -z^2 e^{a_j}$$

5.4 ハッセル行列

逆伝播は誤差の2階微分分にもつたえる。

すなわちの重みとバイアスパラメータをひとつのベクトル W としてみる。

$$H_{ij} = \frac{\partial^2 E}{\partial W_{ij}^2}$$

5.4.1 対角近似

ハッセル行列の逆行列が必要になる事が多い

対角近似によつて逆行列を求めやすくしておく

誤差は $E = \sum_n E_n$ と考える (ミニバッチ学習)

このときハッセル行列の対角成分は

$$\frac{\partial^2 E_n}{\partial W_{ji}^2} = \frac{\partial}{\partial W_{ji}} \left(\frac{\partial E_n}{\partial a_j} z_i \right) = \frac{\partial^2 E_n}{\partial a_j^2} z_i^2 \quad (5.79)$$

$$a_j^{(l)} = \sum_i w_{ij}^{(l)} z_i^{(l-1)} \quad (5.48)$$

$$\frac{\partial E_n}{\partial W_{ji}} = \frac{\partial E_n}{\partial a_j} \frac{\partial a_j}{\partial w_{ij}} = \frac{\partial E_n}{\partial a_j} z_i$$

逆伝播方程式

$$z_j^{(l)} = h(a_j^{(l)}) \dots \quad (5.49)$$

$$\frac{\partial^2 E_n}{\partial a_j^{(l)2}} = \frac{\partial}{\partial a_j^{(l)}} \left(\frac{\partial E_n}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial a_j^{(l)}} \right) = \frac{\partial}{\partial a_j^{(l)}} \left[h'(a_j^{(l)}) \frac{\partial E_n}{\partial a_k^{(l+1)}} \frac{\partial a_k^{(l+1)}}{\partial z_j^{(l)}} \right]$$

$$= \frac{\partial}{\partial a_j^{(l)}} \left[h'(a_j^{(l)}) \sum_k w_{kj}^{(l+1)} \frac{\partial E_n}{\partial a_k^{(l+1)}} \right]$$

$$= h'(a_j^{(l)})^2 \sum_k \sum_{k'} w_{kj}^{(l+1)} w_{k'j}^{(l+1)} \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \quad (5.80)$$

$$+ h''(a_j^{(l)}) \sum_k w_{kj}^{(l+1)} \frac{\partial E_n}{\partial a_k}$$

(5.80) の第2項の 1727A-4 を無視 (非対角をみない)

$$\frac{\partial^2 E_n}{\partial a_j^2} = k(a_j)^2 \sum_k \omega_{kj}^2 \frac{\partial^2 E_n}{\partial a_k^2} + h'(a_j) \sum_k \omega_{kj} \frac{\partial E_n}{\partial a_k} \dots (5.81)$$

しかしハッセ行列は極端に非対角である。←よって近似は注意が必要

5.4.2 外積による近似

NN 回帰問題への応用

= 乗和誤差関数

$$E = \frac{1}{2} \sum_{n=1}^N (y_n - t_n)^2 \quad (5.82)$$

ハッセ行列

$$H = \nabla(\nabla E)^T = \nabla \left[\sum_{n=1}^N \nabla y_n (y_n - t_n) \right]^T$$

$$= \sum_{n=1}^N \nabla y_n (\nabla y_n)^T + \sum_{n=1}^N (y_n - t_n) \nabla \nabla y_n \quad (5.83)$$


平均ゼロ
の確率変数



↓
∇y_n と相関がなければ $\sum_{n=1}^N z_n^2$ の第2項は0になるはず。

$$H \approx \sum_{n=1}^N \mathbf{b}_n \mathbf{b}_n^T \quad (5.84) \quad \text{Levenberg-Marquardt 近似}$$

$$\mathbf{b}_n \equiv \nabla y_n = \nabla a_n$$

出力層は恒等写像 \rightarrow 

$$f(a_n) = y_n$$

適切に訓練された

ネットワークに対しての近似!!

1階微分をよいのと $O(w)$ とする

5.4.3 ハッセ行列の逆行列

外積による近似を用いてハッセ行列の逆行列を効率的に得る。

$$H_N = \sum_{n=1}^N b_n b_n^T \quad (5.86)$$

$$b_n = \nabla_w a_n$$

ハッセ行列を逐次的に求める。L個のデータ点ですべてハッセ行列が得られているとする。

$L = (L-1) + 1$ したときのハッセ行列は

$$H_{L+1} = H_L + b_{L+1} b_{L+1}^T$$

Woodburyの公式 (C.7) から

$$(M + uv^T)^{-1} = M^{-1} - \frac{(M^{-1}u)(v^T M^{-1})}{1 + v^T M^{-1}u}$$

を用いてハッセ行列の逆行列を求める。

$$(H_{L+1})^{-1} = H_L^{-1} - \frac{(H_L^{-1} b_{L+1})(b_{L+1}^T H_L^{-1})}{1 + b_{L+1}^T H_L^{-1} b_{L+1}}$$

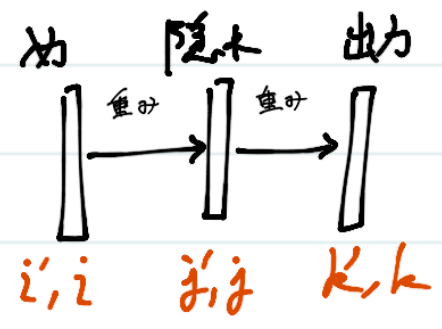
この式から逐次的に逆行列を求めることができる。

実際には αI (α は小さい) という初期値からスタートすることで $H + \alpha I$ の逆行列を求めたいことになる。

5.4.5 ハッセ行列の厳密な公平性

今まで近似手法をみてきたけど実は逆伝播の手法で良い感じに厳密に評価することができる。

ここでは2層の重みをネットワークを考える。



$$\text{誤差 } \delta_k = \frac{\partial E_n}{\partial a_k}, \quad M_{kk} \equiv \frac{\partial^2 E_n}{\partial a_k \partial a_k} \quad E = \sum_{n=1}^N E_n$$

ハッセ行列を3つのブロックに分ける。

1. 両方の重みが第2層

$$\begin{aligned} \frac{\partial^2 E_n}{\partial \omega_{kj}^{(1)} \partial \omega_{kj}^{(2)}} &= \frac{\partial}{\partial \omega_{kj}^{(2)}} \left[\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial \omega_{kj}^{(2)}} \right] & a_k &= \sum_j \omega_{kj} z_j \\ &= \frac{\partial a_k}{\partial \omega_{kj}^{(2)}} \frac{\partial}{\partial a_k} \left[\frac{\partial E_n}{\partial a_k} z_j \right] \\ &= z_j z_j M_{kk} \quad (5.93) \end{aligned}$$

2. 両方第1層

$$\begin{aligned} \frac{\partial}{\partial \omega_{ji}^{(1)}} &= \frac{\partial a_j}{\partial \omega_{ji}^{(1)}} \frac{\partial}{\partial a_j} = x_i \frac{\partial z_j}{\partial a_j} \frac{\partial}{\partial z_j} & a_k &= \sum \omega z \\ &= x_i h'(a_j) \sum_k \frac{\partial a_k}{\partial z_j} \frac{\partial}{\partial a_k} = x_i h'(a_j) \sum_k \omega_{kj}^{(1)} \frac{\partial}{\partial a_k} \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 E_n}{\partial \omega_{ji}^{(1)} \partial \omega_{j'j'}^{(1)}} &= x_i x_{i'} h''(a_j) I_{jj'} \sum_k \omega_{kj}^{(1)} \delta_k \\ &\quad + x_i x_{i'} h'(a_j) h'(a_{j'}) \sum_k \sum_{k'} \omega_{kj}^{(2)} \omega_{k'j'}^{(2)} M_{kk} \quad (5.94) \end{aligned}$$

3. 重みは1つの層に1つずつある

$$\frac{\partial^2 E_n}{\partial \omega_{ji}^{(1)} \partial \omega_{kj}^{(2)}} = x_i h'(a_j) \left\{ \delta_k I_{jj'} + z_{j'} \sum_{k'} \omega_{k'j}^{(2)} M_{kk'} \right\} \quad (5.95)$$

$$\begin{aligned} \frac{\partial}{\partial \omega_{kj}^{(2)}} \left[x_i h'(a_j) \sum_{k'} \omega_{k'j}^{(2)} \frac{\partial E_n}{\partial a_k} \right] &= x_i h'(a_j) \sum_{k'} I_{kk'} I_{jj'} \delta_{k'} + x_i h'(a_j) \sum_{k'} \omega_{k'j}^{(2)} M_{kk'} \\ &= x_i h'(a_j) \left\{ \delta_k I_{jj} + z_j \sum_{k'} \omega_{k'j}^{(2)} M_{kk'} \right\} \end{aligned}$$

5.4.6

ハザード行列はベクトル V の積で全評価したい事が多い

$$V^T H$$

を全評価する。これは N 個の要素しか持たないのだから $O(N)$ で計算できないか?

→ ならばハザード行列を直接評価する
ステップ $O(N)$ より効率的

$$V^T H = V^T \nabla (\nabla E) \quad \text{重み } W \text{ の効果} \quad \nabla W$$

ここでよく使われる記法を紹介する。 $V^T \nabla \cdot = R[\cdot]$ と書くことにする。

ex. $R[W] = V^T \nabla \underbrace{W}_1 = V$

方針 $V^T H$ を計算したい

$$V^T H = V^T \nabla (\nabla E) = R[\nabla E]$$

なので ∇E を求める公式に $R[\cdot]$ を作用させまくる。

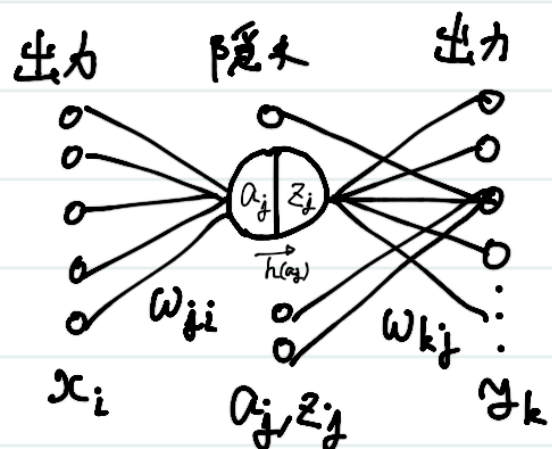
今までと同様に 2 層のネットワークを考えた。

順伝播方程式

$$a_j = \sum_i w_{ji} x_i$$

$$z_j = h(a_j)$$

$$y_k = \sum_j w_{kj} z_j$$



これに $R\{\cdot\}$ を作用させる. → は ω_{ji} に対応する W の要素

$$\begin{array}{ccc} \omega_{ji} = W & & \\ \Downarrow & & \Downarrow \\ z_{ji} = V & & \end{array}$$

$$R\{a_j\} = \sum_i z_{ji} x_i \quad (5.101)$$

$$R\{z_j\} = h(a_j) R\{a_j\} \quad (5.102)$$

$$R\{y_k\} = \sum_j \omega_{kj} R\{z_j\} + \sum_j z_{kj} z_j \quad (5.103)$$

$c = z, R\{a_j\}, R\{z_j\}, R\{y_k\}$ は新しい変数としてみなす.

二乗和誤差を導いているので逆伝播の公式は

$$\delta_k = y_k - t_k \quad (5.104)$$

$$\delta_j = h'(a_j) \sum_k \omega_{kj} \delta_k \quad (5.105)$$

$R\{\cdot\}$ を作用させる

$$R\{\delta_k\} = R\{y_k\}$$

$$R\{\delta_j\} = h''(a_j) R\{a_j\} \sum_k \omega_{kj} \delta_k$$

$$+ h'(a_j) \sum_k z_{kj} \delta_k + h'(a_j) \sum_k \omega_{kj} R\{\delta_k\} \quad (5.107)$$

誤差の1階微分は

$$\frac{\partial E}{\partial \omega_{kj}} = \delta_k z_j, \quad \frac{\partial E}{\partial \omega_{ji}} = \delta_j x_i$$

$R\{\cdot\}$ を作用させる.

$$R\left\{\frac{\partial E}{\partial \omega_{kj}}\right\} = R\{\delta_k\} z_j + \delta_k R\{z_j\} \quad \left. \vphantom{R\left\{\frac{\partial E}{\partial \omega_{kj}}\right\}} \right\} \rightarrow R\left\{\frac{\partial E}{\partial \omega}\right\} \rightarrow W^T H$$

$$R\left\{\frac{\partial E}{\partial \omega_{ji}}\right\} = x_i R\{\delta_j\}$$

を $R\{a\}, R\{z\}, R\{\delta\}$
という変数を導入して書くことも
できる.