

# ニューラルネットワークについて

前川真一

19980328, 051102,0320, 20151013,20210624, 20210719,20dnc,21,27dnc  
,20220720,20230710,12,2024. 10. 8

## 1 はじめに

一変量の基準変数  $y_i$  と  $q$  変量の説明変数  $x_i$  の組  $(y_i, \mathbf{x}_i), i = 1, 2, \dots, N$ , が与えられた時に、 $y$  を  $x$  で説明する一番簡単なモデルは重回帰分析モデル

$$y_i = \beta_0 + \sum_{j=1}^q x_{ij}\beta_j + e_i \quad (1)$$

である。これは  $y$  を  $x$  の線形結合で予測するモデルである。

これを Figure 1 の様に拡張して、2段階で重回帰分析を行うことを考える。すなわち、 $x$  の線形結合として  $A$  個の中間的な変数  $z_{ik}, k = 1, 2, \dots, A$ , を

$$z_{ik} = \alpha_{0k} + \sum_{j=1}^q x_{ij}\alpha_{jk} \quad (2)$$

として作り、 $y$  を  $z$  に

$$y_i = \gamma_0 + \sum_{k=1}^A z_{ik}\gamma_k + e_{i*} \quad (3)$$

と回帰するモデルである。このモデルは、 $A$  が  $q$  よりも大きい場合、一見よりよい説明力を持つ

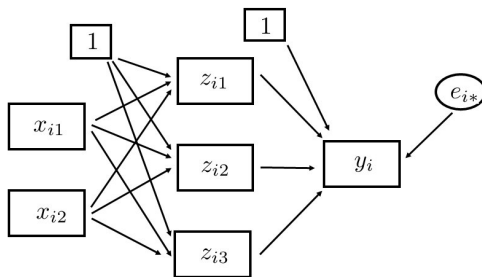


Figure1 重回帰分析の拡張

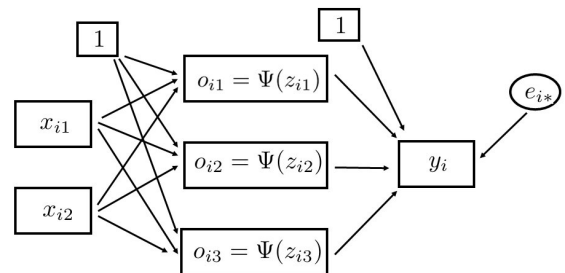


Figure2 ニューラルネットワーク

つように見えるが、実はそうではない。なぜなら、上式の  $z_{ik}$  に Eq.(2) を代入して整理すると

$$y_i = \gamma_0 + \sum_{k=1}^A \left( \alpha_{0k} + \sum_{j=1}^q x_{ij} \alpha_{jk} \right) \gamma_k + e_{i*} \quad (4)$$

$$= \gamma_0 + \sum_{k=1}^A \alpha_{0k} \gamma_k + \sum_{k=1}^A \sum_{j=1}^q x_{ij} \alpha_{jk} \gamma_k + e_{i*} \quad (5)$$

$$= \gamma_0 + \sum_{k=1}^A \alpha_{0k} \gamma_k + \sum_{j=1}^q x_{ij} \sum_{k=1}^A \alpha_{jk} \gamma_k + e_{i*} \quad (6)$$

$$= \beta_{0*} + \sum_{j=1}^q x_{ij} \beta_{j*} + e_{i*} \quad (7)$$

となり、結局、Eq.(1) に示した元の  $q$  変数を用いた重回帰分析と同じ形になるからである。

しかし、Eq.(11) に示すように、 $z_{ik}$  を関数  $\Psi$  を用いて非線形に変換したものを

$$o_{ik} = \Psi(z_{ik}) = \Psi\left(\alpha_{0k} + \sum_{j=1}^q x_{ij} \alpha_{jk}\right) \quad (8)$$

とし、 $y$  を  $o$  に

$$y_i = w_0 + \sum_{k=1}^A o_{ik} w_k + e_{i**} \quad (9)$$

$$= w_0 + \sum_{k=1}^A \Psi(z_{ik}) w_k + e_{i**} \quad (10)$$

$$= w_0 + \sum_{k=1}^A \Psi\left(\alpha_{0k} + \sum_{j=1}^q x_{ij} \alpha_{jk}\right) w_k + e_{i**} \quad (11)$$

という形で回帰する場合には、説明力が増すことが知られている。なぜなら、Eq.(11) は、推定すべきモデルのパラメタ  $\alpha_{jk}, k = 0, 1, \dots, A, j = 1, 2, \dots, q, w_k, k = 0, 1, \dots, A$  を持つ  $\boldsymbol{x}$  の非線形の関数であるからである。これを多変量の  $\boldsymbol{y}$  に拡張したものを3層ニューラルネットワークと呼ぶ。

また、Eq.(8) で定義される  $A$  個の  $o$  の  $B$  個の線形結合を作り、それを非線形に変換したものに  $y$  を回帰する、といったように多段階に拡張したものを多層ニューラルネットワークと呼ぶ。なお、以下の記述では、 $\boldsymbol{x}$  の線形結合として表される  $z$  の事を  $i$  と書き、段階数を表す添字を右上に付けている。また、 $\alpha_0$  や  $w_0$  の様な切片（定数項）は  $\theta$  という文字を使用している。

## 2 教師信号ありの順伝搬型ニューラルネットワーク

$Q$  個の説明変数  $\boldsymbol{x}$  を第 1 層に、 $m$  個の基準変数  $\boldsymbol{y}$  を予測する最終層を第  $n$  層とするネットワークを考える。各層のニューロンの数は  $n_k, k = 1, 2, \dots, n$  とする。（ただし、後述する常に 1 を出力するニューロンは数えない。）従って、 $n_1 = Q, n_n = m$  である。また、個体  $a$  の第  $k$  層の第  $j$  番目のニューロンへの入力を  $i_{aj}^k$ 、同じく出力を  $o_{aj}^k$  と書く。なお、本節では、上付きの添え字  $k, k-1, k+1, n$  等はすべて添え字であり冪乗を示すものではないことに注意された

い。各個体を表す添え字は  $a, a = 1, 2, \dots, N$  である。各ニューロンはそれぞれ上位のニューロンに向かって同一の出力を発する。また、各ニューロンは下位のニューロンの発する出力を重み付けして加えた形でそのニューロンへの入力形成する。すなわち、個体  $a$  の第  $k-1$  層の第  $i$  番目のニューロンの出力  $o_{ai}^{k-1}$  から個体  $a$  の第  $k$  層の第  $j$  番目のニューロンへの入力  $i_{aj}^k$  を作る際の重みを  $w_{ij}^k$  とすれば、

$$i_{aj}^k = \sum_{i=1}^{n_{k-1}} o_{ai}^{k-1} w_{ij}^k + \theta_j^k \quad (12)$$

である。ただし、 $\theta_k$  はバイアスと呼ばれる定数項（閾値）である。また、第  $k$  層の入力はアクティベーション関数  $\psi$  により、

$$o_{aj}^k = \psi(i_{aj}^k) = \frac{1}{1 + \exp(-i_{aj}^k)} \quad (13)$$

という変換を受けた形で出力されるが、アクティベーション関数としては、例えばロジスティック関数の様なシグモイド関数（S字型をした0と1の間の値を取る関数）が用いられることが多い。ただし、最終層  $o_{aj}^n$  は  $y_{aj}$  を近似するためのものであるため、アクティベーション関数による変換を施さない場合もあるが、その場合は、 $o_{aj}^n = i_{aj}^n$  である。また、判別分析のように、最終層の出力として背反な  $m$  個のカテゴリへの所属確率的なものが望ましい場合には、

$$o_{aj}^n = \frac{i_{aj}^n}{\sum_{j'=1}^m i_{aj'}^n} \quad (14)$$

を用いることもあるがこれを softmax と呼ぶ。

この過程を、第  $k$  層の前後に関して図示したものが Figure 3 である。図中、第  $k-1$  層のニューロンの数は3、第  $k$  層のニューロンの数は2、第  $k+1$  層のニューロンの数は2であるが、定数項  $\theta$  は、常に1を出力するニューロンへの重みとして記述してある。また、個体を表す添字  $a$  は省略してある。 $n=3$  とすれば、この図は  $k-1$  層が  $\mathbf{x}$  を  $k=2$  が中間層、 $k+1$  層が以下に示す  $\hat{\mathbf{y}}$  となる3層ニューラルネットワークを表している。（図中の (input) と (output) に相当。）

これを全ての  $i_{aj}^k, o_{aj}^k, a = 1, 2, \dots, N$  に関してまとめて書くと、 $\mathbf{O}^k$  を  $N \times n_k$  の行列、 $\mathbf{W}^k$  を  $n_{k-1} \times n_k$  の行列、 $\boldsymbol{\theta}^k$  を  $1 \times n_k$  のベクトルとして、

$$\mathbf{O}^k = \Psi(\mathbf{I}^k) \quad (15)$$

$$= \Psi(\mathbf{O}^{k-1} \mathbf{W}^k + \mathbf{1}\boldsymbol{\theta}^k) \quad (16)$$

$$= \Psi(\Psi(\mathbf{I}^{k-1}) \mathbf{W}^k + \mathbf{1}\boldsymbol{\theta}^k) \quad (17)$$

$$= \Psi(\Psi(\mathbf{O}^{k-2} \mathbf{W}^{k-1} + \mathbf{1}\boldsymbol{\theta}^{k-1}) \mathbf{W}^k + \mathbf{1}\boldsymbol{\theta}^k) \quad (18)$$

$$\vdots \quad \vdots \quad (19)$$

となるが、これは

$$\mathbf{O}^k = \Psi(\mathbf{O}^{k-1} \mathbf{W}^k + \mathbf{1}\boldsymbol{\theta}^k) = \Gamma_k(\mathbf{O}^{k-1}) \quad (20)$$

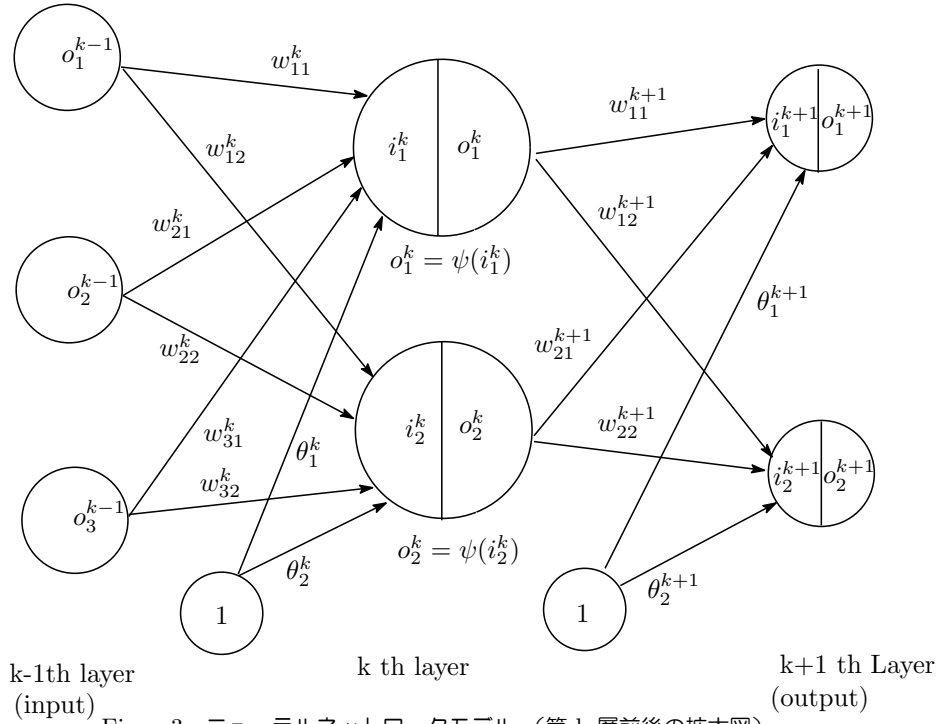


Figure3 ニューラルネットワークモデル (第 k 層前後の拡大図)

と定義をすれば、最終層の出力が

$$\mathbf{O}^n = \Gamma_n(\mathbf{O}^{n-1}) \quad (21)$$

$$= \Gamma_n(\Gamma_{n-1}(\mathbf{O}^{n-2})) \quad (22)$$

$$\vdots \quad (23)$$

$$= \Gamma_n \circ \Gamma_{n-1} \circ \cdots \circ \Gamma_2(\mathbf{O}^1) \quad (24)$$

$$= \Gamma(\mathbf{O}^1) = \Gamma(\mathbf{X}) \quad (25)$$

と書けることを示している。ただし、

$$\Gamma(\cdot) = \Gamma_n \circ \Gamma_{n-1} \circ \cdots \circ \Gamma_2(\cdot) \quad (26)$$

である。従って、ニューラルネットワークのモデルは、基準変数  $\mathbf{Y}$  を説明変数  $\mathbf{X}$  の非線形関数  $\hat{\mathbf{Y}} = \mathbf{O}^n = \Gamma(\mathbf{X})$  で近似するという意味で

$$\mathbf{Y} = \Gamma(\mathbf{X}) + \boldsymbol{\varepsilon} \quad \text{or} \quad \hat{\mathbf{Y}} = \Gamma(\mathbf{X}) \quad (27)$$

と書くことができる。

なお、以下の記述では、各層の第  $n_k$  番目のニューロンに加えて、常に 1 を出力するニューロンがもう一つ存在し、その重みを  $\boldsymbol{\theta}^k$  としている。すなわち、 $\boldsymbol{\theta}^k$  は  $\mathbf{W}^k$  行列の最初の行、すなわち、第 0 行として取り扱われるが、これは  $\mathbf{O}^k$  行列を  $N \times (1 + n_k)$  の行列と定義し、その第 0 列がすべて 1 であることに相当する。

### 3 最小2乗法によるパラメタの推定に必要な1次微分

データ  $Y$  とモデルの値  $\hat{Y}$  の差の2乗和

$$\text{RSS} = \sum_{s=1}^Q \sum_{a=1}^N (y_{as} - \hat{y}_{as})^2 \quad (28)$$

を最小にするようにモデルのパラメタ

$$w_{ij}^k, \quad k = 2, 3, \dots, n, \quad i = 0, 1, 2, \dots, n_{k-1}, \quad j = 1, 2, \dots, n_k$$

を求めるためには以下に示す  $\hat{y}_{as}$  を各パラメタで偏微分したもの、すなわちヤコビアン行列が必要となる。ただし、 $w_{0j}^k$  は前節まで  $\theta_j^k$  と記していたものであり、

$$\hat{y}_{as} = o_{as}^n \quad (29)$$

である。なお、以下の記述では、 $o$  と  $i$  に付く個体を表す添字  $a$  は省略してある。

$$\frac{\partial o_s^n}{\partial w_{ij}^k} = \frac{\partial o_s^n}{\partial i_j^k} \times \frac{\partial i_j^k}{\partial w_{ij}^k} \quad (30)$$

$$\frac{\partial o_s^n}{\partial i_j^k} = \frac{\partial o_s^n}{\partial o_j^k} \times \frac{\partial o_j^k}{\partial i_j^k} \quad (31)$$

$$\frac{\partial o_s^n}{\partial o_j^k} = \sum_{q=1}^{n_{k+1}} \frac{\partial o_s^n}{\partial i_q^{k+1}} \times \frac{\partial i_q^{k+1}}{\partial o_j^k} \quad (32)$$

$$\frac{\partial o_j^k}{\partial i_j^k} = o_j^k(1 - o_j^k), \quad \text{if } k < n \quad (33)$$

$$\frac{\partial o_j^n}{\partial i_j^n} = \begin{cases} o_j^n(1 - o_j^n) & \text{最終層も } \Psi \text{ で変換する場合} \\ 1 & \text{最終層は変換しない場合} \end{cases} \quad (34)$$

$$\frac{\partial i_q^{k+1}}{\partial o_j^k} = w_{jq}^{k+1} \quad (35)$$

$$\frac{\partial i_j^k}{\partial w_{ij}^k} = \begin{cases} o_i^{k-1} & \text{if } i \geq 1 \quad (w) \\ 1 & \text{if } i = 0 \quad (\theta) \end{cases} \quad (36)$$

なお、Eq.(32) に於いて  $i_0^{k+1}$  は定数 1 であり  $o_j^k$  に依存しないため、総和記号の和は  $q = 1$  からとなる。従って、 $\delta_j^{k,s} = \frac{\partial o_s^n}{\partial i_j^k}$  と定義すると、この値は一つ上の第  $k + 1$  層から最終層である第  $n$  層までの値を使って、 $2 \leq k < n$  に関して

$$\delta_j^{k,s} = \frac{\partial o_s^n}{\partial i_j^k} \quad (37)$$

$$= \sum_{q=1}^{n_{k+1}} \frac{\partial o_s^n}{\partial i_q^{k+1}} w_{jq}^{k+1} \times \frac{\partial o_j^k}{\partial i_j^k} = \sum_{q=1}^{n_{k+1}} \delta_q^{k+1,s} w_{jq}^{k+1} \times \frac{\partial o_j^k}{\partial i_j^k} \quad (38)$$

$$= \sum_{q=1}^{n_{k+1}} \delta_q^{k+1,s} w_{jq}^{k+1} \times o_j^k(1 - o_j^k) \quad (39)$$

と再帰的に書くことが出来るが、これを誤差逆伝搬 error back propagation と呼ぶ。なお、最終層である第  $n$  層は

$$\delta_j^{n,s} = \frac{\partial o_s^n}{\partial i_j^n} = \begin{cases} \begin{cases} o_s^n(1-o_s^n) & \text{最終層も } \Psi \text{ で変換する場合} \\ 1 & \text{最終層は変換しない場合} \end{cases} & \text{if } j = s \\ 0 & \text{otherwise} \end{cases} \quad (40)$$

である。

まとめると

$$\frac{\partial o_s^n}{\partial w_{ij}^k} = \frac{\partial o_s^n}{\partial i_j^k} \times \frac{\partial i_j^k}{\partial w_{ij}^k} = \delta_j^{n,s} \times \begin{cases} o_i^{k-1} & \text{if } i \geq 1 \quad (w) \\ 1 & \text{if } i = 0 \quad (\theta) \end{cases} \quad (41)$$

$$= \sum_{q=1}^{n_{k+1}} \delta_q^{k+1,s} w_{jq}^{k+1} \times o_j^k(1-o_j^k) \times \begin{cases} o_i^{k-1} & \text{if } i \geq 1 \quad (w) \\ 1 & \text{if } i = 0 \quad (\theta) \end{cases} \quad (42)$$

となる。また、RSS の  $w_{ij}^k$  に関する微分 (gradient) は

$$\frac{\partial \text{RSS}}{\partial w_{ij}^k} = -2 \sum_{s=1}^m \sum_{a=1}^N (y_{as} - o_{as}^n) \times \frac{\partial o_{as}^n}{\partial w_{ij}^k} \quad (43)$$

となる。

### 3.1 行列表記

いま、 $\mathbf{O}^k$ ,  $k = 1, 2, \dots, n$  を  $N \times n_k$  の第  $k$  層の出力を要素として持つ行列とする。ただし、 $\mathbf{O}^1 = \mathbf{X}$  であり、最終層の第  $s$  列を  $o_{(s)}^n$  という  $N \times 1$  のベクトルで表す。また、 $\mathbf{W}^k$ ,  $k = 2, 3, \dots, n$  を第  $k-1$  層の出力から第  $k$  層の入力を

$$\mathbf{I}^k = \begin{bmatrix} \mathbf{1}_N, & \mathbf{O}^{k-1} \end{bmatrix} \mathbf{W}^k \quad (44)$$

という形で作るための  $(n_{k-1} + 1) \times n_k$  の重み行列とする。なお、 $\mathbf{O}^{k-1}$  行列の最初の列を  $\mathbf{1}$  と定義せずに、 $N \times n_{k-1}$  の行列  $\mathbf{O}^{k-1}$  の第 1 列の左側に  $\mathbf{1}$  を付け加えたものに  $\mathbf{W}^k$  を掛けて  $\mathbf{I}^k$  を定義していることに注意されたい。したがって、 $(n_{k-1} + 1) \times n_k$  の行列  $\mathbf{W}^k$  を行単位に

$$\mathbf{W}^k = (\mathbf{w}_0^k, \mathbf{w}_1^k, \mathbf{w}_2^k, \dots, \mathbf{w}_{n_{k-1}}^k)' \quad (45)$$

と定義すれば、第 0 行  $\mathbf{w}_0^{k'}$  は定数項 1 に掛かる係数、第 1 行  $\mathbf{w}_1^{k'}$  が  $\mathbf{O}^{k-1}$  の第 1 列、 $o_{(1)}^{k-1}$ 、に掛かる係数、第 2 行  $\mathbf{w}_2^{k'}$  が  $o_{(2)}^{k-1}$  に掛かる係数、という具合になる。また、 $\Delta_s^k$  を Eq.(39) で定義される  $\delta_j^{k,s}$ 、すなわち  $\delta_{aj}^{k,s} = \frac{\partial o_{as}^n}{\partial i_{aj}^k}$  をその  $a, j$  要素として持つ  $N \times n_k$  の行列と定義すると、Eq.(39) と Eq.(40) は

$$\Delta_s^k = \begin{cases} \begin{cases} o_{(s)}^n \odot (\mathbf{1}_N - o_{(s)}^n) \mathbf{e}'_s & \text{最終層も } \Psi \text{ で変換する場合} \\ \mathbf{1}_N & \text{最終層は変換しない場合} \end{cases} & \text{if } k = n \\ \mathbf{O}^k \odot (\mathbf{1}_N \mathbf{1}'_{n_k} - \mathbf{O}^k) \odot (\Delta_s^{k+1} \mathbf{W}^{k+1 \dagger}) & \text{if } k = n-1, \dots, 2 \\ k = 2, 3, \dots, n-1, \quad s = 1, 2, \dots, n_n & \end{cases} \quad (46)$$

と書くことが出来る。ただし、 $\odot$  は行列のアダマール積を表し、 $\mathbf{e}_s$  はその第  $s$  行のみが 1 となっている  $n_n \times 1$  のベクトルである。また、 $\mathbf{W}^{k+1 \dagger}$  は  $\mathbf{W}^{k+1}$  行列の第 0 行目を取り除いた

$n_k \times n_{k+1}$  行列を転置して得られる  $n_{k+1} \times n_k$  の行列である。(この  $\mathbf{W}$  の第 0 行目を取り除くことは Eq.(32) の和が  $q = 1$  から始まっていることに対応している。)

したがって、Eq.(41) に示す  $\frac{\partial o_{qs}^n}{\partial w_{ij}^k}$  の  $\mathbf{w}_i^k$  の要素  $w_{ij}^k$ ,  $j = 1, 2, \dots, n_k$  に関する部分をその  $a_j$  要素として持つ  $N \times n_k$  のヤコビアン行列は

$$\frac{\partial \mathbf{o}_{(s)}^n}{\partial \mathbf{w}_i^k} = \begin{cases} \left( \mathbf{o}_{(i)}^{k-1} \mathbf{1}'_{n_k} \right) \odot \Delta_s^k & \text{if } i \geq 1 \\ \Delta_s^k & \text{if } i = 0 \end{cases} \quad (47)$$

であり、第  $k$  層の重み行列  $\mathbf{W}^k$  のすべての要素に関するヤコビアン行列は Eq.(47) を  $i = 0, 1, 2, \dots, n_k$  と変えて横に並べた  $N \times (n_{k-1} + 1)n_k$  の行列

$$\begin{aligned} \frac{\partial \mathbf{o}_{(s)}^n}{\partial \text{vec}(\mathbf{W}^{k'})} &= \left[ \frac{\partial \mathbf{o}_{(s)}^n}{\partial w_0^k}, \frac{\partial \mathbf{o}_{(s)}^n}{\partial w_1^k}, \frac{\partial \mathbf{o}_{(s)}^n}{\partial w_2^k}, \dots, \frac{\partial \mathbf{o}_{(s)}^n}{\partial w_{n_k}^k} \right] \quad (48) \\ &= \left[ \Delta_s^k, \left( \mathbf{o}_{(1)}^{k-1} \mathbf{1}'_{n_k} \right) \odot \Delta_s^k, \left( \mathbf{o}_{(2)}^{k-1} \mathbf{1}'_{n_k} \right) \odot \Delta_s^k, \dots, \left( \mathbf{o}_{(n_{k-1})}^{k-1} \mathbf{1}'_{n_k} \right) \odot \Delta_s^k \right] \quad (49) \end{aligned}$$

となる。ただし、 $\text{vec}(\mathbf{W}^{k'})$  は  $\mathbf{W}^k$  行列の各行を横に並べて作られた横ベクトルの転置である。また、

$$\text{RSS} = \sum_{s=1}^m (\mathbf{y}_{(s)} - \mathbf{o}_{(s)}^n)' (\mathbf{y}_{(s)} - \mathbf{o}_{(s)}^n) \quad (50)$$

$$= (\text{vec}(\mathbf{Y}) - \text{vec}(\mathbf{O}^n))' (\text{vec}(\mathbf{Y}) - \text{vec}(\mathbf{O}^n)) \quad (51)$$

であるから、全ての  $n_n$  個の変数に関する  $(N \times n_n) \times (n_{k-1} + 1)n_k$  のヤコビアン行列は

$$\frac{\partial \text{vec}(\mathbf{O}^n)}{\partial \text{vec}(\mathbf{W}^{k'})} = \left[ \left( \frac{\partial \mathbf{o}_{(1)}^n}{\partial \text{vec}(\mathbf{W}^{k'})} \right)', \left( \frac{\partial \mathbf{o}_{(2)}^n}{\partial \text{vec}(\mathbf{W}^{k'})} \right)', \dots, \left( \frac{\partial \mathbf{o}_{(n_n)}^n}{\partial \text{vec}(\mathbf{W}^{k'})} \right)' \right]' \quad (52)$$

となり、これらを層の添字  $k$  を変えて横に並べたものが全ての重み行列に関するヤコビアン行列となる。

なお、 $\frac{\partial \mathbf{o}_{(s)}^n}{\partial \text{vec}(\mathbf{W}^{k'})}$  行列の列を commutation matrix  $\mathbf{K}(n_k, n_{k-1} + 1)$  を用いて並べ替えると

$$\frac{\partial \mathbf{o}_{(s)}^n}{\partial \text{vec}(\mathbf{W}^k)} = \frac{\partial \mathbf{o}_{(s)}^n}{\partial \text{vec}(\mathbf{W}^{k'})} \mathbf{K}(n_k, n_{k-1} + 1) \quad (53)$$

が得られる。ただし、 $e_{a,b}$  をその第  $b$  要素の値のみが 1 でそれ以外の要素は 0 である  $a$  次元の縦ベクトルとして

$$\mathbf{K}(r, c) = \sum_{i=1}^r \sum_{j=1}^c (e_{r,i} e'_{c,j}) \otimes (e_{c,j} e'_{r,i}) \quad (54)$$

と書ける。(commutation matrix に関しては Harville, 1997, や wikipedia を参照のこと。) したがって、 $\frac{\partial \text{vec}(\mathbf{O}^n)}{\partial \text{vec}(\mathbf{W}^k)}$  は上記の  $\frac{\partial \mathbf{o}_{(s)}^n}{\partial \text{vec}(\mathbf{W}^{k'})}$  を縦に並べて得ることが出来る。

最後に、RSS の  $\mathbf{W}^k$  に関する微分 (gradient) は以下の  $(n_{k-1} + 1)n_k \times 1$  のベクトル

$$\frac{\partial \text{RSS}}{\partial \text{vec}(\mathbf{W}^{k'})} = -2 \left[ \frac{\partial \text{vec}(\mathbf{O}^n)}{\partial \text{vec}(\mathbf{W}^{k'})} \right]' (\text{vec}(\mathbf{Y}) - \text{vec}(\mathbf{O}^n)) \quad (55)$$

となる。また、

$$\frac{\partial \text{RSS}}{\partial \text{vec}(\mathbf{W}^k)} = \mathbf{K}^{(n_k, n_{k-1} + 1)} \frac{\partial \text{RSS}}{\partial \text{vec}(\mathbf{W}^{k'})} \quad (56)$$

である。

## References

- Mayekawa, Shin-ichi. (1995) An efficient algorithm for feed-forward neural network regression analysis. *Behaviormetrika*. Vol.22, No. 1, 67-90
- Commutation matrix. (n.d.). In *Wikipedia* Retrieved from [https://en.wikipedia.org/w/index.php?title=Commutation\\_matrix&oldid=1027627556](https://en.wikipedia.org/w/index.php?title=Commutation_matrix&oldid=1027627556)
- Harville,D.A. (1997) *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag New York, Inc.