

多層パーセプトロンにおける 順伝播・誤差逆伝播計算の行列形式

宮本 崇

2017/4/3

1 はじめに

多層パーセプトロンにおける順伝播・逆伝播計算を，行列形式で定式化する．以下に，この文章で用いる記号の定義を表 1 にまとめる．

Symbols	Meaning
N_b	ミニバッチサイズ
L	ニューラルネットワークにおける全層数
$N_u^{(l)}$	l 層のユニットサイズ
$\mathbf{x}_n = {}^t\{x_{1n}, \dots, x_{N_u^{(1)}n}\}$	ミニバッチ内のサンプルの一つ
$\mathbf{y}_n = {}^t\{y_{1n}, \dots, y_{N_u^{(L)}n}\}$	\mathbf{x}_n に対するシステムの出力
$\mathbf{d}_n = {}^t\{d_{1n}, \dots, d_{N_u^{(L)}n}\}$	\mathbf{x}_n に対応する目標出力
E_n	\mathbf{x}_n に対する誤差関数
$E = \sum_{k=1}^{N_b} E_n$	ミニバッチ全体に対する誤差関数
$\mathbf{v}_n^{(l)} = {}^t\{v_{1n}^{(l)}, \dots, v_{N_u^{(l)}n}^{(l)}\}$	\mathbf{x}_n をシステムへの入力としたときの l 層への入力ベクトル
$\mathbf{z}_n^{(l)} = {}^t\{z_{1n}^{(l)}, \dots, z_{N_u^{(l)}n}^{(l)}\}$	\mathbf{x}_n をシステムへの入力としたときの l 層からの出力ベクトル
$\phi^{(l)}(\cdot)$	l 層における活性化関数
$\mathbf{W}^{(l)} = [w_{ij}^{(l)}], i \leq N_u^{(l)}, j \leq N_u^{(l-1)}$	$l-1$ 層- l 層間の線形変換における重み
$\mathbf{b}^{(l)} = {}^t\{b_1^{(l)}, \dots, b_{N_u^{(l)}}^{(l)}\}$	$l-1$ 層- l 層間の線形変換におけるバイアス
$\delta_n^{(l)} = {}^t\{\delta_{1n}^{(l)}, \dots, \delta_{N_u^{(l)}n}^{(l)}\}$	l 層におけるデルタ $\delta_{in}^{(l)}$ を並べたベクトル
$\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_{N_b}\}$	\mathbf{x}_n を列ベクトルに並べた行列
$\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_{N_b}\}$	\mathbf{y}_n を列ベクトルに並べた行列
$\mathbf{D} = \{\mathbf{d}_1, \dots, \mathbf{d}_{N_b}\}$	\mathbf{d}_n を列ベクトルに並べた行列
$\mathbf{V}^{(l)} = \{\mathbf{v}_1^{(l)}, \dots, \mathbf{v}_{N_b}^{(l)}\}$	$\mathbf{v}_n^{(l)}$ を列ベクトルに並べた行列
$\mathbf{Z}^{(l)} = \{\mathbf{z}_1^{(l)}, \dots, \mathbf{z}_{N_b}^{(l)}\}$	$\mathbf{z}_n^{(l)}$ を列ベクトルに並べた行列
$\mathbf{\Delta}^{(l)} = \{\delta_1^{(l)}, \dots, \delta_{N_b}^{(l)}\}$	$\delta_n^{(l)}$ を列ベクトルに並べた行列
$\partial \mathbf{W}^{(l)} = \left[\frac{\partial E}{\partial w_{ij}^{(l)}} \right]$	重み $w_{ij}^{(l)}$ に関する誤差関数 E の偏微分を成分 (i, j) に持つ行列
$\partial \mathbf{b}^{(l)} = {}^t\left\{ \frac{\partial E}{\partial b_1^{(l)}}, \dots, \frac{\partial E}{\partial b_{N_u^{(l)}}^{(l)}} \right\}$	重み $b_i^{(l)}$ に関する誤差関数 E の偏微分を i 成分に持つベクトル
$\mathbf{1}_N$	1 を N 個並べてできるベクトル

表 1 多層パーセプトロンに関する記号の定義

2 多層パーセプトロン

2.1 順伝播

ミニバッチ内のあるサンプル x_n について、順伝播の計算は以下のように書くことができる。

$$\mathbf{v}_n^{(l)} = \mathbf{W}^{(l)} \mathbf{z}_n^{(l-1)} + \mathbf{b}^{(l)} \quad (l \geq 2) \quad (1)$$

$$\mathbf{z}_n^{(l)} = \begin{cases} \mathbf{x}_n & (l = 1) \\ \phi^{(l)}(\mathbf{v}_n^{(l)}) & (\text{otherwise}) \end{cases} \quad (2)$$

ここで、 $\phi^{(l)}(\mathbf{v}_n^{(l)})$ は、ベクトル $\mathbf{v}_n^{(l)}$ の各要素に対して活性化関数 $\phi(\cdot)$ を作用させる操作を意味するものとする。

式 (1) は前の層からの出力を線形変換したものが次の層への入力となることを、式 (2) は層への入力を活性化関数により非線形変換したものが出力となることを表している。システム全体の出力 \mathbf{y}_n は L 層の出力 $\mathbf{z}_n^{(L)}$ となる。

式 (1) や式 (2) の全てのサンプルについての計算は、次の行列形式にまとめることができる。

$$\mathbf{V}^{(l)} = \mathbf{W}^{(l)} \mathbf{Z}^{(l-1)} + \mathbf{b}^{(l)} \cdot \mathbf{1}_{N_b} \quad (l \geq 2) \quad (3)$$

$$\mathbf{Z}^{(l)} = \begin{cases} \mathbf{X} & (l = 1) \\ \phi^{(l)}(\mathbf{V}^{(l)}) & (\text{otherwise}) \end{cases} \quad (4)$$

2.2 逆伝播

次に、ミニバッチ X に対する誤差関数 E について、システムのパラメタ $w_{ij}^{(l)}$ に関する偏微分を以下のように求める。

$$\frac{\partial E}{\partial w_{ij}^{(l)}} = \frac{1}{N} \sum_{n=1}^{N_b} \frac{\partial E_n}{\partial w_{ij}^{(l)}} = \frac{1}{N} \sum_{n=1}^{N_b} \frac{\partial E_n}{\partial v_{in}^{(l)}} \cdot \frac{\partial v_{in}^{(l)}}{\partial w_{ij}^{(l)}} = \frac{1}{N} \sum_{n=1}^{N_b} \delta_{in}^{(l)} \cdot z_{jn}^{(l-1)} \quad (5)$$

ここで、 $\frac{\partial E_n}{\partial v_{in}^{(l)}} = \delta_{in}^{(l)}$ とおいた。上式は、次の行列形式に表すことができる。

$$\partial \mathbf{W}^{(l)} = \frac{1}{N} \mathbf{\Delta}^{(l)} \cdot \mathbf{Z}^{(l-1)} \quad (6)$$

同様に、パラメタ $b_i^{(l)}$ に関する偏微分も求めることができる。

$$\frac{\partial E}{\partial b_i^{(l)}} = \frac{1}{N} \sum_{n=1}^{N_b} \frac{\partial E_n}{\partial b_i^{(l)}} = \frac{1}{N} \sum_{n=1}^{N_b} \frac{\partial E_n}{\partial v_{in}^{(l)}} \cdot \frac{\partial v_{in}^{(l)}}{\partial b_i^{(l)}} = \frac{1}{N} \sum_{n=1}^{N_b} \delta_{in}^{(l)} \cdot 1 \quad (7)$$

$$\therefore \partial \mathbf{b}^{(l)} = \frac{1}{N} \mathbf{\Delta}^{(l)} \cdot \mathbf{1}_{N_b} \quad (8)$$

次に、 $\mathbf{\Delta}^{(l)}$ の各成分 $\delta_{in}^{(l)}$ は、次のように求めることができる。

$$\begin{aligned} \delta_{in}^{(l)} &= \frac{\partial E_n}{\partial v_{in}^{(l)}} = \sum_{k=1}^{N_u^{(l+1)}} \frac{\partial E_n}{\partial v_{kn}^{(l+1)}} \cdot \frac{\partial v_{kn}^{(l+1)}}{\partial v_{in}^{(l)}} \\ &= \sum_{k=1}^{N_u^{(l+1)}} \delta_{kn}^{(l+1)} \cdot \frac{\partial \sum_m (w_{km}^{(l+1)} \cdot \phi^{(l)}(v_{mn}^{(l)}) + b_m^{(l)})}{\partial v_{in}^{(l)}} \\ &= \sum_{k=1}^{N_u^{(l+1)}} \delta_{kn}^{(l+1)} \cdot w_{ki}^{(l+1)} \cdot \phi'^{(l)}(v_{in}^{(l)}) \\ &= \phi'^{(l)}(v_{in}^{(l)}) \cdot \sum_{k=1}^{N_u^{(l+1)}} w_{ki}^{(l+1)} \cdot \delta_{kn}^{(l+1)} \end{aligned} \quad (9)$$

式 (9) は, $l+1$ 層の $\delta_{in}^{(l+1)}$ から l 層の $\delta_{in}^{(l)}$ を逐次的に求めることができることを意味している. L 層における $\delta_{in}^{(L)}$ のみ, 定義に従って次のように計算する.

$$\delta_{in}^{(L)} = \frac{\partial E_n}{\partial v_{in}^{(L)}} = \frac{\partial E_n}{\partial z_{in}^{(L)}} \cdot \frac{\partial z_{in}^{(L)}}{\partial v_{in}^{(L)}} = \frac{\partial E_n}{\partial y_{in}} \cdot \phi'^{(L)}(v_{in}^{(L)}) \quad (10)$$

この式の右辺の値は, 問題に応じた E_n や $\phi^{(L)}(\cdot)$ の具体的な定義によって異なるが, 回帰問題において標準的に用いられる 2 乗和誤差・恒等写像の組み合わせや, 多クラス分類におけるクロスエントロピー誤差・ソフトマックス関数の組み合わせでは, いずれも $y_{in} - d_{in}$ となる. したがって, 以上の $\delta_{in}^{(L)}$ に関する計算を行列形式に整理すると, 以下のようになる.

$$\Delta^{(l)} = \begin{cases} \mathbf{Y} - \mathbf{D} & (l = L) \\ \phi^{(l)'}(\mathbf{V}^{(l)}) \circ ({}^t \mathbf{W}^{(l+1)} \Delta^{(l+1)}) & (otherwise) \end{cases} \quad (11)$$

ここで, \circ は行列間, もしくはベクトル間の要素積 (アダマール積) を表すものとする.

2.3 計算のまとめ

以上より, 多層パーセプトロンにおける順伝播・逆伝播の計算は次のようにまとめられる.

順伝播

$$\mathbf{V}^{(l)} = \mathbf{W}^{(l)} \mathbf{Z}^{(l-1)} + \mathbf{b}^{(l)} \cdot {}^t \mathbf{1}_{N_b} \quad (l \geq 2) \quad (12)$$

$$\mathbf{Z}^{(l)} = \begin{cases} \mathbf{X} & (l = 1) \\ \phi^{(l)}(\mathbf{V}^{(l)}) & (otherwise) \end{cases} \quad (13)$$

逆伝播

$$\Delta^{(l)} = \begin{cases} \mathbf{Y} - \mathbf{D} & (l = L) \\ \phi^{(l)'}(\mathbf{V}^{(l)}) \circ ({}^t \mathbf{W}^{(l+1)} \Delta^{(l+1)}) & (otherwise) \end{cases} \quad (14)$$

$$\partial \mathbf{W}^{(l)} = \frac{1}{N} \Delta^{(l)} \cdot {}^t \mathbf{Z}^{(l-1)} \quad (15)$$

$$\partial \mathbf{b}^{(l)} = \frac{1}{N} \Delta^{(l)} \cdot \mathbf{1}_{N_b} \quad (16)$$

このようにして求められた, 層ごとのパラメタによる偏微分値 $\partial \mathbf{W}^{(l)}$ を元に, 勾配降下法などの更新則に基づいて $\mathbf{W}^{(l)}$ を更新する. 勾配降下法を例にとると, η を学習係数として, 次のようにパラメタを更新する.

$$\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} - \eta \cdot \partial \mathbf{W}^{(l)} \quad (17)$$