

# 知的情報処理論 第一回レポート

28G23027

川原尚己

1.

- (a) Nearest Neighbor 法では, ある点 $\mathbf{x}$ のクラスは,  $\mathbf{x}$ と各クラスの代表点 $\mathbf{p}_i$ との距離が最小のクラスである. すなわち,  $\mathbf{x}$ の属するクラスを $C_x$ とすると,

$$C_x = \operatorname{argmin}_i |\mathbf{x} - \mathbf{p}_i| \quad (1)$$

であり,

$$C_x = \operatorname{argmax}_i (-|\mathbf{x} - \mathbf{p}_i|) \quad (2)$$

とも書ける.

$|\mathbf{x} - \mathbf{p}_i| \geq 0$ であるから, 2 乗しても大小が変化することはない,

$$\begin{aligned} C_x &= \operatorname{argmax}_i (-|\mathbf{x} - \mathbf{p}_i|^2) \\ &= \operatorname{argmax}_i (-|\mathbf{x}|^2 + 2\mathbf{p}_i\mathbf{x} - |\mathbf{p}_i|^2) \end{aligned} \quad (3)$$

としても結果は変わらない. (3)式において $-|\mathbf{x}|^2$ は $i$ に依存していないため $i$ に関しての大小の変化には関与しない.

$$\therefore C_x = \operatorname{argmax}_i (2\mathbf{p}_i\mathbf{x} - |\mathbf{p}_i|^2) \quad (4)$$

また,

$$C_x = \operatorname{argmax}_i (g_i(\mathbf{x})) \quad (5)$$

であるから,

$$g_i(\mathbf{x}) = 2\mathbf{p}_i\mathbf{x} - |\mathbf{p}_i|^2 \quad (6)$$

となる.

さらに,  $w_{i0} = -|\mathbf{p}_i|^2, (w_{i1}, w_{i2}, \dots, w_{id})^t = 2\mathbf{p}_i, \mathbf{x} = (x_1, x_2, \dots, x_d)^t, x_0 = 1$ とすると,

$$g_i(\mathbf{x}) = x_0 w_{i0} + (w_{i1}, w_{i2}, \dots, w_{id})(x_1, x_2, \dots, x_d)^t \quad (7)$$

と表せる. 新たに $\mathbf{x} = (x_0, x_1, \dots, x_d)^t, x_0 = 1, \mathbf{w}_i = (w_{i0}, w_{i1}, \dots, w_{id})^t, w_{i0} = -|\mathbf{p}_i|^2$ と定義すると,

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} \quad (8)$$

と表せる.

(8)式は $g_i(\mathbf{x})$ が $\mathbf{x}$ に対して線形であることを示している.

最後に, 重みベクトル $\mathbf{w}_i$ は, プロトタイプの座標 $\mathbf{p}_i$ を用いて以下のように表される:

$$\mathbf{w}_i = (w_{i0}, w_{i1}, \dots, w_{id})^t, w_{i0} = -|\mathbf{p}_i|^2, (w_{i1}, \dots, w_{id}) = 2\mathbf{p}_i \quad (9)$$

(b) (8)式及び(9)式に従って識別関数 $g_i(\mathbf{x})$  ( $i = 1, 2, 3$ )を求める.

$$\begin{aligned} g_1(\mathbf{x}) &= \mathbf{w}_1^t \mathbf{x} \\ &= (-|(8,4)^t|^2, 16, 8)(1, x_1, x_2)^t \\ &= -80 + 16x_1 + 8x_2 \end{aligned} \quad (10)$$

$$\begin{aligned} g_2(\mathbf{x}) &= \mathbf{w}_2^t \mathbf{x} \\ &= (-|(2,8)^t|^2, 4, 16)(1, x_1, x_2)^t \\ &= -68 + 4x_1 + 16x_2 \end{aligned} \quad (11)$$

$$\begin{aligned} g_3(\mathbf{x}) &= \mathbf{w}_3^t \mathbf{x} \\ &= (-|(4,2)^t|^2, 8, 4)(1, x_1, x_2)^t \\ &= -20 + 8x_1 + 4x_2 \end{aligned} \quad (12)$$

(c) 式(10), (11), 及び(12)を用いて決定境界 $g_{ij} \stackrel{\text{def}}{=} g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$  ( $i < j$ )は以下のように表される.

$$\begin{aligned} g_{12}(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\ &= (-80 + 16x_1 + 8x_2) - (-68 + 4x_1 + 16x_2) \\ &= -12 + 12x_1 - 8x_2 \\ \therefore g_{12}(\mathbf{x}) &= -12 + 12x_1 - 8x_2 = 0 \end{aligned} \quad (13)$$

同様に,  $g_{13}(\mathbf{x}), g_{23}(\mathbf{x})$ は,

$$g_{13}(\mathbf{x}) = -60 + 8x_1 + 4x_2 = 0 \quad (14)$$

$$g_{23}(\mathbf{x}) = -48 - 4x_1 + 12x_2 = 0 \quad (15)$$

となる.

2. 以下の表 1 及び図 1 に学習過程と結果を示す.

学習係数 $\rho$	更新回数	重み( $w_0, w_1$ )	更新発生時の $x_1$
0.33	0	(0.4, -0.25)	
	1	(0.070, 0.410)	-2.0
	2	(-0.260, 0.443)	-0.1
	3	(0.070, 0.509)	0.2
	4	(-0.260, 0.542)	-0.1
	5	(0.070, 0.608)	0.2
	6	(-0.260, 0.641)	-0.1
	7	(0.070, 0.707)	0.2

表 1 学習過程と結果

表 1 ではパーセプトロンの学習係数, 初期重み及び重みの学習過程を表している. 「更新回数」の列の「0」は初期値を表している. 今回の測定では学習係数 $\rho$ を0.33, 初期重みを( $w_0, w_1$ ) = (0.4, -0.25)とした. 重みの更新回数は7回であった.

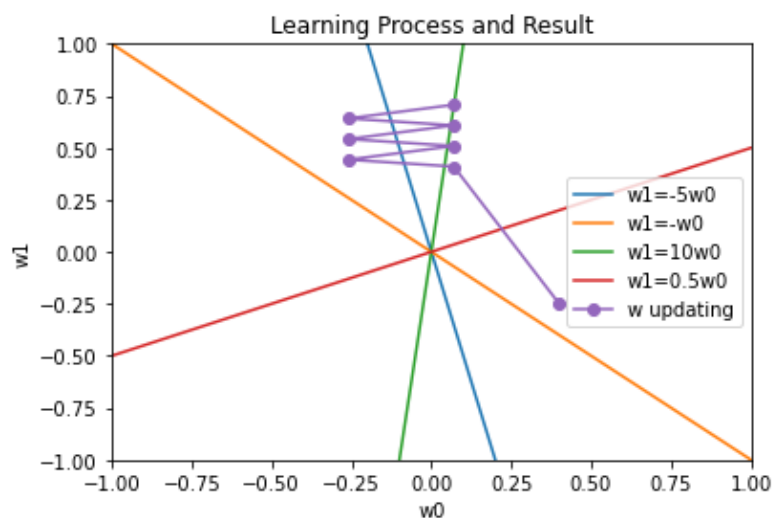


図 1 学習過程と結果の図示

図 1 には, 表 1 における学習過程の図示を行っている.

今回の終了条件は $w_1 > -5w_0$ かつ $w_1 > 10w_0$ である.

今回の学習ではまず $w_1 > 0.5w_0$ の条件にて更新がなされた後,  $w_1 > -5w_0$ か $w_1 > 10w_0$ において更新がなされている.

今回の実験で興味を引いたことは, 更新回数が学習係数 $\rho$ の値に敏感であったことである. 上記した $\rho = 0.33$ の場合では更新回数は7回であったが,  $\rho = 0.32$ の時には11回,  $\rho = 0.34$ では5回であった.  $\rho$ は正の全ての範囲をとり得るが, 0.02の差で 100%以

上の更新回数の変化があることは興味深い結果であった。そこで、 $\rho$ の変化に対し測定回数はどれほどの変化が起こるのかを測定するため、初期重み $w_0, w_1$ に対し $-1 \leq w_0, w_1 \leq 1$ の中でランダムに選択し、 $\rho$ を0.1から1まで0.01刻みで変化させながら100000回測定した。その結果から $\rho$ ごとの平均値と標準偏差を計算し、隣り合った $\rho$ に対する差分の絶対値を求めた。その結果を以下の図2に示す。ただし $w_0, w_1$ の初期値は終了条件内の値はとらないようにしている。

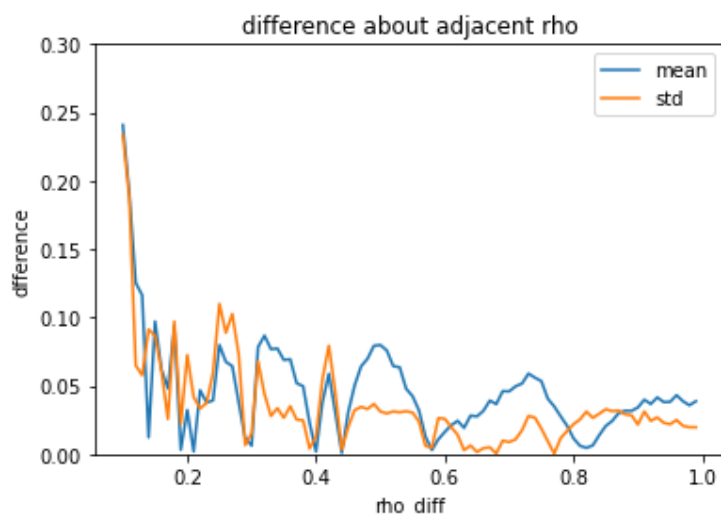


図2 測定回数の平均値と標準偏差

図2では、 $\rho_{diff}$ は $\rho$ が $\rho_{diff}+1$ のときと $\rho_{diff}$ のときの測定結果を表している。例えば $\rho_{diff}=0.20$ のとき、 $\rho=0.21$ と $\rho=0.20$ のときの差分の絶対値を表している。この実験結果では平均値や標準偏差の差分の絶対値をとっているため、値が小さいほど隣り合う $\rho$ と値が近く、大きいほど離れている。

実験前は、初期重みはランダムで決定するため、 $\rho$ の値によらずある程度ならかなグラフになると予想していたが、実際に測定してみると、予想通り全体的に差分は小さい( $\rho_{diff} \geq 0.2$ では平均値の差分は0.1以下)となっていた。しかしながら、 $\rho$ の値の変化に対しては敏感で、0に近い差分をとった直後に0.05を超える差分をとっていたりと、 $\rho$ に従った性質は $\rho$ が非常に小さい値をとるときには差分は大きくなるということを除けば決定的なものは見受けられなかった。ただ、この結果からは $\rho$ が大きくなるほどに差分の変化の急峻さが小さくなっているようにも見え、さらに大きな $\rho$ 、あるいはさらに細かい差分についての検証が必要である。

3. 以下の図3に多層ネットワークの概要図を示し、これに基づいて誤差逆伝播法を導出する。

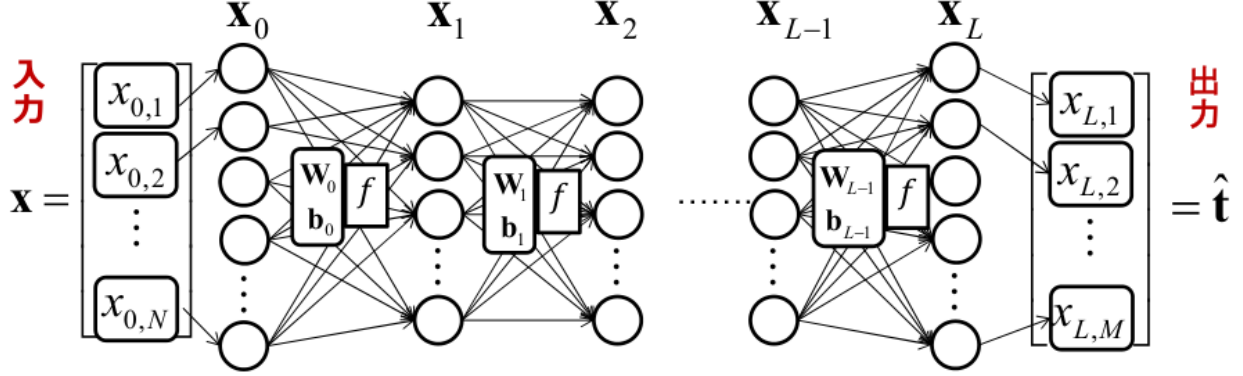


図3 多層ネットワーク概要図（講義資料より引用）

以下では、 $p$ 番目の学習パターンを $\mathbf{x}_p$ 、ある層における $i$ 番目のユニットをユニット $i$ と呼ぶ。入力 $\mathbf{x}_p$ に対し第 $k$ 層のユニット $i$ からの出力信号を $g_{ip}^{(k)}$ 、第 $(k+1)$ 層のユニット $j$ への信号の重みを $w_{ij}^{(k)}$ とする。また、第 $k$ 層のユニットの集合 $U_k$ を $U_k = \{1, 2, \dots\}$ とする。

このとき第 $k$ 層のユニット $i$ から第 $(k+1)$ 層のユニット $j$ への入力信号 $h_{jp}^{(k)}$ は、

$$h_{jp}^{(k)} = \sum_{i \in U_k} w_{ij}^{(k)} g_{ip}^{(k)} \quad (16)$$

と表され、第 $(k+1)$ 層のユニット $j$ の出力信号 $g_{jp}^{(k+1)}$ は、活性化関数 $f(\cdot)$ を用いて

$$g_{jp}^{(k+1)} = f(h_{jp}^{(k)}) \quad (17)$$

と表される。

出力層 $L$ では教師信号 $(b_{1p}, \dots, b_{Mp})$ が与えられるが、それだけではネットワーク全体の処理はわからないため、中間の隠れ層に教師信号を伝える方法を考える必要がある。

出力層 $L$ においては、その出力と教師信号を用いて、その誤差を以下のような評価関数 $J_p$ で表す。

$$J_p = \frac{1}{2} \sum_{i \in U_L} (g_{ip}^{(L)} - b_{ip})^2 \quad (18)$$

また、重み $w_{ij}^{(k)}$ の更新は $J_p$ と学習係数 $\rho$ を用いて

$$w_{ij}^{(k)'} = w_{ij}^{(k)} - \rho \frac{\partial J_p}{\partial w_{ij}^{(k)}} \quad (19)$$

となる。(19)式の $\frac{\partial J_p}{\partial w_{ij}^{(k)}}$ は連鎖律を用いて

$$\frac{\partial J_p}{\partial w_{ij}^{(k)}} = \frac{\partial J_p}{\partial h_{jp}^{(k)}} \frac{\partial h_{jp}^{(k)}}{\partial w_{ij}^{(k)}} \quad (20)$$

と書ける．(20)式の $\frac{\partial h_{jp}^{(k)}}{\partial w_{ij}^{(k)}}$ は式(16)より $g_{ip}^{(k)}$ に等しい． $\frac{\partial J_p}{\partial h_{jp}^{(k)}}$ を $h_{jp}^{(k)}$ に関する誤差の変化量として

$$\epsilon_{jp}^{(k)} = \frac{\partial J_p}{\partial h_{jp}^{(k)}} \quad (21)$$

とおくと，(20)式は以下のように書き換えられる．

$$\frac{\partial J_p}{\partial w_{ij}^{(k)}} = \epsilon_{jp}^{(k)} g_{ip}^{(k)} \quad (22)$$

また， $\epsilon_{jp}^{(k)}$ についても連鎖律を用い，

$$\epsilon_{jp}^{(k)} = \frac{\partial J_p}{\partial h_{jp}^{(k)}} = \frac{\partial J_p}{\partial g_{jp}^{(k+1)}} \frac{\partial g_{jp}^{(k+1)}}{\partial h_{jp}^{(k)}} \quad (23)$$

となる．式(23)の $\frac{\partial g_{jp}^{(k+1)}}{\partial h_{jp}^{(k)}}$ は式(17)を用いると以下のように書ける．

$$\frac{\partial g_{jp}^{(k+1)}}{\partial h_{jp}^{(k)}} = f' \left( h_{jp}^{(k)} \right) \quad (24)$$

以下では式(23)の $\frac{\partial J_p}{\partial g_{jp}^{(k+1)}}$ の層ごとのふるまいを考えていく．

ユニット $j$ が出力層 $L$ にあるときには式(18)を用いて，

$$\frac{\partial J_p}{\partial g_{jp}^{(L)}} = g_{jp}^{(L)} - b_{jp} \quad (25)$$

と書ける．一方，ユニット $j$ が中間層にあるとき，式(16)及び式(21)を用いて，連鎖律から

$$\begin{aligned} \frac{\partial J_p}{\partial g_{jp}^{(k+1)}} &= \sum_{m \in U_{k+1}} \frac{\partial J_p}{\partial h_{mp}^{(k+1)}} \frac{\partial h_{mp}^{(k+1)}}{\partial g_{jp}^{(k+1)}} \\ &= \sum_{m \in U_{k+1}} \epsilon_{mp}^{(k+1)} w_{jm}^{(k+1)} \end{aligned} \quad (26)$$

と書ける．以上より，式(23)は

$$\epsilon_{jp}^{(k)} = \begin{cases} \left( g_{jp}^{(L)} - b_{jp} \right) f' \left( h_{jp}^{(L)} \right) & (k = L) \\ \left( \sum_{m \in U_{k+1}} \epsilon_{mp}^{(k+1)} w_{jm}^{(k+1)} \right) f' \left( h_{jp}^{(k)} \right) & (k \neq L) \end{cases} \quad (27)$$

と書ける．これで誤差逆伝播法が導出できた．

式(27)を見ると，第 $k$ 層における誤差は第 $(k + 1)$ 層における誤差を求めることで与えられることがわかる．このように，入力から出力という方向とは逆の方向である，出力から入力という方向へ誤差を求めている手法であるため誤差逆伝播法と名付けられていると考えられる．