2012 年度春学期　画像情報処理

第１４回付録：数学的説明

誤差逆伝播法

一般の第 k 層と第 $(k+1)$ 層について，誤差逆伝播法を記号と式を使って書くと，次のようになります。図 A1 も参照してください。

いま，第 k 層の j 番目のニューロンが受け持つ誤差の量を δ_j^k とし，第 $(k+1)$ 層の l 番目のニューロンが受け持っている誤差の量を δ_l^{k+1} とします。また，第 k 層の j 番目のニューロンから $(k+1)$ 層の l 番目のニューロンへの結合の重みを $w_{j\ l}^{k\ (k+1)}$ で表します。さらに，第 k 層の j 番目のニューロンの状態を x_j^k で表します。

さて，上の「考え方」1. の条件は，「δ_j^k のうち第 $(k+1)$ 層の l 番目のニューロンから割り振られる量」が $w_{j\ l}^{k\ (k+1)}$ に比例することを意味します。また，「第 k 層の j 番目のニューロンに第 $(k-1)$ 層から送られてくる刺激の総和」S_j^k は

$$S_j^k = \sum_i w_j^{(k-1)}{}_l^k x_i^{k-1} \tag{A1}$$

と表せますから，「考え方」2. の条件は，「δ_j^k のうち第 $(k+1)$ 層の l 番目のニューロンから割り振られる量」が，dx_j^k/dS_j^k にも比例することを意味します。これらの誤差が第 $(k+1)$ 層の各ニューロンから割り振られて，その合計が δ_j^k となるわけですから

$$\delta_j^k = \sum_l \left(w_{j\ l}^{k\ k+1} \frac{dx_j^k}{dS_j^k} \right) \delta_l^{k+1} \tag{A2}$$

となります。ここで，（しきい関数の代用のシグモイド関数などの）非線形関数を $f()$ とすると

$$x_j^k = f(S_j^k) \tag{A3}$$

ですから，

$$\frac{dx_j^k}{dS_j^k} = f'(S_j^k) \tag{A4}$$

となり，(A2) 式は

$$\begin{aligned} \delta_j^k &= \sum_l \left(w_{j\ l}^{k\ k+1} f'(S_j^k) \right) \delta_l^{k+1} \\ &= \left[\sum_l (w_{j\ l}^{k\ k+1} \delta_l^{k+1}) \right] f'(S_j^k) \end{aligned} \tag{A5}$$

と表されます。この式によって，第 k 層の j 番目のニューロンが受け持つ誤差の量が，第 k 層の各ニューロンが持つ誤差から決まります。$k=n$ のとき，すなわち出力層については第 $(k+1)$ 層がありませんが，出力層での誤差とは所望の出力と現実の出力との差ですから，出力層の l 番目のニューロンにおける所望の出力を y_j とすると

$$\delta_l^n = (x_l^n - y_l) f'(S_l^n) \tag{A6}$$

となります。

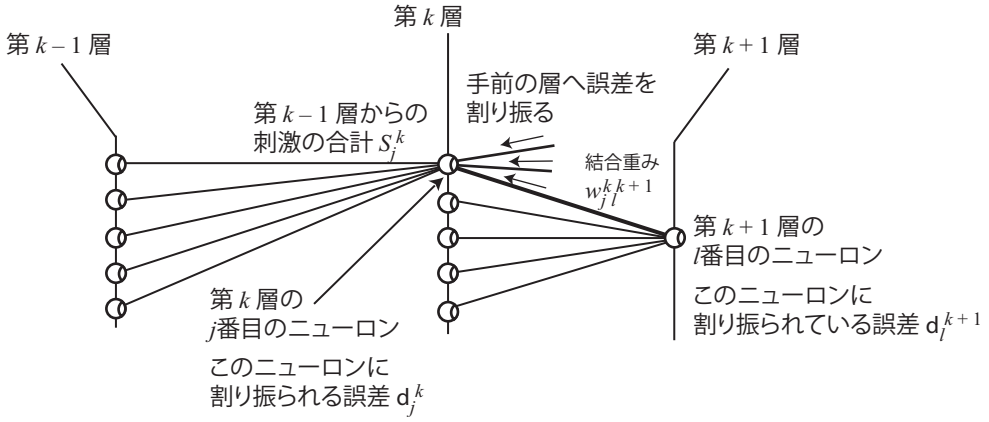

図 A1: 誤差逆伝播法

(A5)(A6) 式で，各層の各ニューロンが受け持つ誤差が決まりました。そこで，これらの誤差に本文 (4) 式の δ-ルールを適用して，新しい結合重み $w'_{j\ l}^{k\ k+1}$ を

$$w'_{j\ l}^{k-1\ k} = w_{j\ l}^{k-1\ k} - \varepsilon\delta_l^k x_j^{k-1} \tag{A7}$$

のように結合の重みを修正します。ε は本文 (4) 式と同じ学習係数です。

誤差逆伝播法が最急降下になっていることの証明

ある入出力例における，所望の出力例と現実の出力（第 n 層でのニューロンの状態）の誤差の 2 乗和を

$$E = \frac{1}{2}\sum_l (x_l^n - y_l)^2 \tag{A8}$$

と定義します。このとき，E は各結合重みの関数です。いま，結合重み $w_{j\ l}^{k-1\ k}$ 方向の E の微分を求めると

$$\frac{\partial E}{\partial w_{j\ l}^{k-1\ k}} = \frac{\partial E}{\partial S_l^k}\cdot\frac{\partial S_l^k}{\partial w_{j\ l}^{k-1\ k}} \tag{A9}$$

となり，また本文 (A1) 式から

$$\frac{\partial S_l^k}{\partial w_{j\ l}^{k-1\ k}} = \frac{\partial}{\partial w_{j\ l}^{k-1\ k}}\sum_i w_{i\ l}^{k-1\ k} x_i^{k-1} = x_j^{k-1} \tag{A10}$$

となります。そこで

$$\delta_l^k = \frac{\partial E}{\partial S_l^k} \tag{A11}$$

とおいて，(A10)(A11) 式を (A9) 式に代入すると

$$\frac{\partial E}{\partial w_{j\ l}^{k-1\ k}} = \delta_l^k x_j^{k-1} \tag{A12}$$

となります。この式を，結合重みを修正する方法を表す (A7) 式，すなわち

$$w'_{j\ l}^{k-1\ k} = w_{j\ l}^{k-1\ k} - \varepsilon\delta_l^k x_j^{k-1}$$

と比較すると，(A7) 式は，結合重みを $\partial E/\partial w_j^{k-1}{}_l^k$ 方向，すなわち $-\mathrm{grad}E$ の方向に変化させています。すなわち，(A7) 式による結合重みの修正は，E の最急降下になっていることがわかります。

さて，(A11) 式の δ は，本文の「考え方」1., 2. から導いた δ と同じものであることを示しましょう。出力層以外，すなわち $k \neq n$ のとき，式 (A11) を

$$\frac{\partial E}{\partial S_j^k} = \sum_l \frac{\partial E}{\partial S_l^{k+1}} \cdot \frac{\partial S_l^{k+1}}{\partial x_j^k} \cdot \frac{\partial x_j^k}{\partial S_j^k} \tag{A13}$$

と変形します。一方，本文の (A4) 式から

$$\frac{dx_j^k}{dS_j^k} = f'(S_j^k) \tag{A14}$$

です。また，(A1) 式の添え字を付け替えると

$$S_l^{k+1} = \sum_i w_i^k{}_l^{k+1} x_i^k \tag{A15}$$

ですから，

$$\frac{\partial S_l^{k+1}}{\partial x_j^k} = w_j^k{}_l^{k+1} \tag{A16}$$

となります。(A14) 式，(A16) 式を (A13) 式に代入すると

$$\frac{\partial E}{\partial S_j^k} = \sum_l \frac{\partial E}{\partial S_l^{k+1}} \cdot w_j^k{}_l^{k+1} \cdot f'(S_j^k) \tag{A17}$$

となり，(A11) 式の δ の定義から

$$\delta_j^k = \frac{\partial E}{\partial S_j^k},\ \delta_j^{k+1} = \frac{\partial E}{\partial S_l^{k+1}} \tag{A18}$$

です。これを (A17) 式に代入すると

$$\begin{aligned} \delta_j^k &= \sum_l \delta_j^{k+1} \cdot w_j^k{}_l^{k+1} \cdot f'(S_j^k) \\ &= \left[\sum_l (w_j^k{}_l^{k+1} \delta_j^{k+1}) \right] f'(S_j^k) \end{aligned} \tag{A19}$$

となり，本文の δ の定義と同じになります。

出力層では，すなわち $k = n$ のときは，(A11) 式を

$$\frac{\partial E}{\partial S_l^n} = \frac{\partial E}{\partial x_l^n} \cdot \frac{dx_l^n}{dS_l^n} \tag{A20}$$

と変形します。誤差２乗和の定義である (A8) 式から

$$\begin{aligned} \frac{\partial E}{\partial x_l^n} &= \frac{\partial}{\partial x_l^n} \left\{ \frac{1}{2} \sum_l (x_l^n - y_l)^2 \right\} \\ &= x_l^n - y_l \end{aligned} \tag{A21}$$

となります。これを (A20) 式に代入すると

$$\frac{\partial E}{\partial S_l^n} = (x_l^n - y_l) \cdot \frac{dx_l^n}{dS_l^n} \tag{A22}$$

となり，(A11) 式の δ の定義から

$$\delta_l^n = \frac{\partial E}{\partial S_l^n} \tag{A23}$$

で，また (A14) 式から

$$\frac{dx_l^n}{dS_l^n} = f'(S_l^n) \tag{A24}$$

ですから，(A23) 式，(A24) 式を (A22) 式に代入すると

$$\delta_l^n = (x_l^n - y_l) f'(S_l^n) \tag{A25}$$

となり，本文の (A6) 式の δ の定義と同じになります。

Lagrange の未定乗数法による，マージン最大化問題の解法

この方法では，下のように評価関数を定義します。

$$L(\boldsymbol{w}, b, \alpha_i) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_i \alpha_i [y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1] \tag{A26}$$

ここで $\alpha_i \geq 0$ は未定乗数です。$\boldsymbol{w}$ と b が最適値になるとき，

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{w}} &= \boldsymbol{w} - \sum_i \alpha_i y_i \boldsymbol{x}_i \\ \frac{\partial L}{\partial b} &= -\sum_i \alpha_i y_i \end{aligned} \tag{A27}$$

はいずれも 0 となります。そこで，(A27) 式の微分を 0 とおくと，

$$\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i \tag{A28}$$

$$\sum_i \alpha_i y_i = 0 \tag{A29}$$

が得られます。ここで (A26) 式から

$$L(\boldsymbol{w}, b, \alpha_i) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_i \alpha_i y_i \boldsymbol{w}^T\boldsymbol{x}_i - b\sum_i \alpha_i y_i \sum_i \alpha_i \tag{A30}$$

となるので，この式に (A28) 式と (A29) 式を代入すると

$$\begin{aligned} L(\boldsymbol{w}, b, \alpha_i) =& \frac{1}{2}\left(\sum_i \alpha_i y_i \boldsymbol{x}_i\right)^T \left(\sum_j \alpha_j y_j \boldsymbol{x}_j\right) \\ & - \sum_i \alpha_i y_i \left(\sum_j \alpha_j y_j \boldsymbol{x}_j\right)^T \boldsymbol{x}_i + \sum_i \alpha_i \\ =& -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_i \alpha_i \end{aligned} \tag{A31}$$

となります。(A26) 式の第２項の影響は最小にしなければならず，かつ L は最大にしなければなりません。したがって，この最適化問題は次の２次計画問題に帰着されます。

$$\begin{aligned} maximize \quad & -\frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j + \sum_i \alpha_i \\ subject\ to \quad & \sum_i \alpha_i y_i = 0, \alpha_i \geq 0 \end{aligned} \tag{A32}$$

２次計画問題をコンピュータで解くためには，多数市販されているプログラムパッケージが利用できます。

カーネル関数と境界面

カーネル関数 $K(\boldsymbol{x},\boldsymbol{x}')$ は，

$$K(\boldsymbol{x},\boldsymbol{x}') = \Phi(\boldsymbol{x})^T\Phi(\boldsymbol{x}') \tag{A33}$$

という関係を満たすものです。この式は，カーネル関数が，変換 Φ で変換された高次元空間で測られた距離に相当するものであることを意味しています。

この結果，通常の距離のかわりにカーネル関数を用いてマージンを測り，このマージンをもとに最適化を行なって分離超平面を求めると，もとの空間では，この境界は「曲がった」境界になります。変換後の境界は

$$\boldsymbol{w}^T\Phi(\boldsymbol{x}) + b = 0 \tag{A34}$$

で表されます。(A28) 式を，$\boldsymbol{x}$ を $\Phi(\boldsymbol{x})$ におきかえたうえで (A34) 式に代入すると，

$$\sum_i \alpha_i y_i \Phi(\boldsymbol{x}_i^T)\Phi(\boldsymbol{x}) + b = \sum_i \alpha_i y_i K(\boldsymbol{x}_i^,\boldsymbol{x}) + b = 0 \tag{A35}$$

となります。(A32) 式の最適化関数も，変換後の形は $\boldsymbol{x}_i^T\boldsymbol{x}_j$ を $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ におきかえることで得られます。このことは，境界を求めるのに必要な計算はすべて $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ を用いて可能で，変換後の空間や変換 Φ が実際にどういうものかは，知る必要はない，ということを示しています。

正定値の二次形式となっているカーネル関数 K は，(A33) 式の条件を満たすことが知られています。このようなカーネル関数の例には，つぎのようなものがあります。

$$K(\boldsymbol{x},\boldsymbol{x}') = (\boldsymbol{x}^T\boldsymbol{x}' + 1)^p \quad \text{（多項式カーネル）} \tag{A36}$$

$$K(\boldsymbol{x},\boldsymbol{x}') = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{\sigma^2}) \quad \text{（ガウシアンカーネル）} \tag{A37}$$