誤差逆伝播法

吉永 塁

2020 年 5 月 15 日

記号設定

- L 層のニューラルネットワークを考える.
- 第 l 層のユニット数を $v(l)$ とおく.
- 各ユニットの入出力について,

$$z_j^{(l)} = f(u_j^{(l)}) = f\left(\sum_{i=0}^{v(l-1)} w_{ji}^{(l)} z_i^{(l-1)}\right)$$

  - z は入力, u は出力を表す.
  - 右肩括弧内の数字は第何層かを, 下添字は何番目のユニットかを示す.
  - w は重み.
  - バイアスは $w_{j0}^{(l)} = b_j^{(l)}$, $z_0^{(l)} \equiv 1$ として重みの一部とする.
  - f は活性化関数.
- $\boldsymbol{w}$: 全てのパラメータ (重み, バイアス) を成分に持つベクトル.
- $\mathbf{1}_n \coloneqq \sum_{i=1}^{n} \boldsymbol{e}_i$: 全成分が 1 の n 次元ベクトル

連鎖律

定理 (連鎖律)

U, W がそれぞれ $\mathbb{R}^n, \mathbb{R}^m$ の開集合であり，二つの関数 $f : U \to \mathbb{R}^n$, $g : W \to \mathbb{R}^p$ が合成可能，つまり $f(U) \subset W$ であるとする．
いま，f が $\boldsymbol{x} \in U$ で微分可能で，g が $\boldsymbol{y} = f(\boldsymbol{x})$ で微分可能であるとすれば，合成関数 $\varphi = g \circ f$ は $\boldsymbol{x}$ で微分可能で，

$$\frac{\partial \varphi_r}{\partial x_j}(\boldsymbol{x}) = \sum_{i=1}^{m} \frac{\partial g_r}{\partial y_i}(\boldsymbol{y}) \frac{\partial f_i}{\partial x_j}(\boldsymbol{x})$$
$$(1 \leq r \leq p, 1 \leq j \leq n)$$

但し，添字 i は第 i 成分を表す．

証明略．

誤差逆伝播法

- 勾配降下法では重み更新において勾配 ∇E が必要．ここで誤差関数 $E(\boldsymbol{w})$ について，

$$\begin{cases} E(\boldsymbol{w}) = \displaystyle\sum_{n=1}^{N} E_n(\boldsymbol{w}) & \text{(バッチ学習)} \\ E(\boldsymbol{w}) = \dfrac{1}{|D_t|} \displaystyle\sum_{n \in D_t} E_n(\boldsymbol{w}) & \text{(ミニバッチ学習)} \end{cases}$$

より，何れも $\nabla E_n(\boldsymbol{w})$ を計算できれば良い．

- 勾配計算はそのまま行うと面倒．例えば，二乗誤差の場合，

$$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = {}^t(\boldsymbol{y}(\boldsymbol{x}_n) - \boldsymbol{d}_n) \frac{\partial \boldsymbol{y}}{\partial w_{ji}^{(l)}}$$

となり，$\boldsymbol{y}$ について，活性化関数のネストが深くなるので，連鎖律を繰り返し適用しなければいけない．

誤差逆伝播法

第 l 層の重み $w_{ji}^{(l)}$ について，

$$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \frac{\partial E_n}{\partial u_j^{(l)}} \frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}} \tag{1}$$

(1) 式右辺第一項について，

$$\frac{\partial E_n}{\partial u_j^{(l)}} = \sum_{k=0}^{v(l+1)} \frac{\partial E_n}{\partial u_k^{(l+1)}} \frac{\partial u_k^{(l+1)}}{\partial u_j^{(l)}} \tag{2}$$

ここで，記号

$$\delta_j^{(l)} := \frac{\partial E_n}{\partial u_j^{(l)}} \tag{3}$$

を導入する．δ は誤差と呼ばれる．

誤差逆伝播法

(2) 式右辺第二項について，$u_k^{(l+1)} = \sum_i w_{ki}^{(l+1)} f(u_i^{(l)})$ に注意して，

$$\frac{\partial u_k^{(l+1)}}{\partial u_j^{(l)}} = w_{kj}^{(l+1)} f'(u_j^{(l)})$$

よって式 (2) は，

$$\delta_j^{(l)} = f'(u_j^{(l)}) \sum_{k=0}^{v(l+1)} \delta_k^{(l+1)} w_{kj}^{(l+1)} \tag{4}$$

と表される．この式は，$\delta_j^{(l)}$ が $\delta_k^{(l+1)} (k = 0, \ldots, v(l+1))$ により計算できることを意味する．

- 微分計算は出力層から入力層へと順に計算できる．
- 順伝播とは逆向きに伝播するので誤差逆伝播と呼ばれる．

また，逆伝播の最初の値

$$\delta_j^{(L)} = \frac{\partial E_n}{\partial u_j^{(L)}}$$

は陽に計算される．
(1) 式右辺第二項について，$u_j^{(l)} = \sum_k w_{jk}^{(l)} z_k^{(l-1)}$ に注意して，

$$\frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}} = z_i^{(l-1)}$$

以上より，(1) 式は，

$$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} z_i^{(l-1)} \tag{5}$$

と表される．

誤差逆伝播法

以上より，訓練サンプル $(\boldsymbol{x}_n, \boldsymbol{d}_n)$ が与えられたとき，この訓練サンプルについての誤差関数 E_n の勾配は次の手順で得られる．

誤差逆伝播法による勾配計算

1. $\boldsymbol{z}^{(1)} = \boldsymbol{x}_n$ として各層の入出力 $\boldsymbol{u}^{(l)}, \boldsymbol{z}^{(l)} (l = 2, \ldots, L)$ を計算．
   このとき計算は入力層から順に行われる．[順伝播]
2. 出力層 (第 L 層) の誤差 $\delta_j^{(L)}$ を計算．
3. 中間層について，(4) 式より誤差 $\delta_j^{(l)} (l = 2, \ldots, L-1)$ を計算．
   このとき計算は出力層から順に行われる．[逆伝播]
4. (5) 式より各層の重み $w_{ji}^{(l)}$ に関する微分を計算．

複数個の訓練サンプルについての誤差の総和は

$$\frac{\partial E}{\partial w_{ji}^{(l)}} = \sum_n \frac{\partial E_n}{\partial w_{ji}^{(l)}}$$

より，単に加算すればよい．

誤差逆伝播法

計算量について考える．ネットワークの重みの総数を W とおく．
このとき，順伝播+逆伝播により勾配 ∇E を求めるのには，$\mathcal{O}(W)$ の計算が必要となる．

$\because$ 一般的に，(重み数) $>>$ (ユニット数) であることに注意する．
順伝播計算の大部分は，$u_j^{(l)} = \sum_i w_{ji}^{(l)} z_i^{(l-1)}$ であるので，計算量は W による．
逆伝播計算についても，$\delta_j^{(l)} = f'(u_j^{(l)}) \sum_k \delta_k^{(l+1)} w_{kj}^{(l+1)}$ より同様．

このことより，数値微分

$$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \frac{1}{\varepsilon}(E_n(w_{ji}^{(l)} + \varepsilon) - E_n(w_{ji}^{(l)})) + O(\varepsilon)$$

は，全ての重みについて $E_n(w_{ji}^{(l)} + \varepsilon)$ を計算する必要があるため，全体では $\mathcal{O}(W^2)$ となる．
数値微分は計算効率は良くないが，逆伝播の実装の確認として有用．

誤差逆伝播法

ここまでの計算を行列形式で表す．但し，バイアスと重みを区別して扱う．

- $\boldsymbol{X} = \left[\boldsymbol{x}_1 \cdots \boldsymbol{x}_N\right]$: $\boldsymbol{x}_n$ は n 番目の訓練サンプル.
- $\boldsymbol{U}^{(l)} = \left[\boldsymbol{u}_1^{(l)} \cdots \boldsymbol{u}_N^{(l)}\right]$: $\boldsymbol{u}_n^{(l)}$ は $\boldsymbol{x}_n$ についての ${}^t\left[u_1^{(l)} \cdots u_{v(l)}^{(l)}\right]$
- $\boldsymbol{Z}^{(l)} = \left[\boldsymbol{z}_1^{(l)} \cdots \boldsymbol{z}_N^{(l)}\right]$: $\boldsymbol{z}_n^{(l)}$ は $\boldsymbol{u}_n^{(l)}$ と同様.
- $\boldsymbol{W}^{(l)} = \left[w_{ji}^{(l)}\right]_{\substack{1 \le j \le v(l) \\ 1 \le i \le v(l-1)}}$: ユニット間に接続がなければ 0.
- $\boldsymbol{b}^{(l)} = {}^t\left[b_1 \cdots b_{v(l)}\right]$

以上より，順伝播計算は $\boldsymbol{U}^{(1)} = \boldsymbol{X}$ として，

$$\boldsymbol{U}^{(l)} = \boldsymbol{W}^{(l)}\boldsymbol{Z}^{(l-1)} + \boldsymbol{b}^{(l)}\,{}^t\boldsymbol{1}_N \tag{6}$$

$$\boldsymbol{Z}^{(l)} = f^{(l)}\left(\boldsymbol{U}^{(l)}\right) \tag{7}$$

と表される $(l = 2, \ldots, L)$．但し，活性化関数 $f^{(l)}$ は各成分に適応されることとする.

逆伝播について，

- $\boldsymbol{Y} = \left[\boldsymbol{y}_1 \cdots \boldsymbol{y}_N\right]$: $\boldsymbol{y}_n$ は $\boldsymbol{x}_n$ についての出力.
- $\boldsymbol{D} = \left[\boldsymbol{d}_1 \cdots \boldsymbol{d}_N\right]$: $\boldsymbol{d}_n$ は $\boldsymbol{x}_n$ についての出力目標.
- $\boldsymbol{\Delta}^{(l)} = \left[\boldsymbol{\delta}_1^{(l)} \cdots \boldsymbol{\delta}_N^{(l)}\right]$: $\boldsymbol{\delta}_n^{(l)}$ は $\boldsymbol{x}_n$ についての ${}^t\left[\delta_1^{(l)} \cdots \delta_{v(l)}^{(l)}\right]$

とおく．以上より，逆伝播計算は $\boldsymbol{\Delta}^{(L)} = \boldsymbol{D} - \boldsymbol{Y}$ として，

$$\boldsymbol{\Delta}^{(l)} = f^{(l)\prime}(\boldsymbol{U}^{(l)}) \odot ({}^t\boldsymbol{W}^{(l+1)}\boldsymbol{\Delta}^{(l+1)}) \tag{8}$$

と表される $(l = L-1, \ldots, 2)$．但し，$\odot$ はアダマール積.

誤差逆伝播法

誤差関数の勾配計算について，

- $\partial \boldsymbol{W}^{(l)} = \left[\frac{\partial E}{\partial w_{ji}^{(l)}}\right]_{\substack{1\leq j\leq v(l)\\ 1\leq i\leq v(l-1)}}$， $\partial \boldsymbol{b}^{(l)} = {}^{t}\left[\frac{\partial E}{\partial b_1^{(l)}} \cdots \frac{\partial E}{\partial b_{v(l)}^{(l)}}\right]$

とおく．以上より，

$$\partial \boldsymbol{W}^{(l)} = \frac{1}{N}\boldsymbol{\Delta}^{(l)}\,{}^{t}\boldsymbol{Z}^{(l-1)}$$
$$\partial \boldsymbol{b}^{(l)} = \frac{1}{N}\boldsymbol{\Delta}^{(l)}\boldsymbol{1}_N$$

となり，パラメータ更新は，

$$\boldsymbol{W}^{(l)} \leftarrow \boldsymbol{W}^{(l)} - \varepsilon\,\partial \boldsymbol{W}^{(l)} \tag{9}$$
$$\boldsymbol{b}^{(l)} \leftarrow \boldsymbol{b}^{(l)} - \varepsilon\,\partial \boldsymbol{b}^{(l)} \tag{10}$$

と表される．重みの更新について，重み減衰やモメンタムを用いる場合には，(9)(10) 式右辺第二項を適当に修正する．