誤差逆伝播法 (Backpropagation)

目次

1 概要 (abstract)

誤差逆伝播法 (Backpropagation) はフィードフォワード型のニューラルネットワークの学習に用いられる方法である。確率的勾配法 (stochastic gradient descent) と呼ばれる方法の一つであり、広い意味では最小値を探索するアルゴリズムである。ニューラルネットワークの学習は、損失関数をパラメータの関数と考えたときに、損失関数を目的関数とした最小化問題であるということができる。最小化問題としてのニューラルネットワークに対して確率的勾配法を適用したのが誤差逆伝播法ということができる。

2 アルゴリズム (algorithm)

まず使用するニューラルネットワークを定義する。対象とするニューラルネットワークの層数を N とする。k $(k = 1, 2, \cdots, N)$ 層目のニューロン k $(i = 1, 2, \cdots, n_k)$ の出力値を $x_i^{(k)}$、$k+1$ 層目のニューロンの計算に用いられるパラメータを $w_i^{(k)}$ $i = 1, 2, \cdots, m_k$ とする。ここで n_k は k 層目のニューロン数、m_k は $k+1$ 層目のニューロンの出力値の計算に用いられるパラメータの数である。また、$k+1$ 層目のニューロン i の出力値を計算する関数を $f_i^{(k)}$ と

する。これらによって順伝播の計算は

$$x_i^{(k+1)} = f_i^{(k)}\left(x^{(k)}, w^{(k)}\right)$$

として書かれる。また、損失関数を L とすると、p 個目の教師パターン $t_i^{(p)}$ $(i = 1, 2, \cdots, n_N)$ とネットワークの出力 $x_i^{(N)}$ との誤差は

$$L\left(t^{(p)}, x^{(N)}\right)$$

となる。よって教師パターン $t^{(p)}$ に対する各重みの更新量 $\Delta w_i^{(k)}$ は

$$\Delta w_i^{(k)} = -\eta \frac{\partial L}{\partial w_i^{(k)}}$$

として書かれる。更新量は連鎖律によって

$$\Delta w_i^{(k)} = -\eta \partial^{(p)} x_j^{(k)} \partial^{(p)} w_i^{(k)}$$

と変形できる。ここで

$$\partial^{(p)} x_i^{(k)} = \begin{cases} \frac{\partial L}{\partial x_i^{(N)}} = \partial_{x_i} L\left(t^{(p)}, x^{(N)}\right), & k = N \\ \sum_j \frac{\partial L}{\partial x_j^{(k+1)}} \frac{\partial x_j^{(k+1)}}{\partial x_i^{(k)}} = \sum_j \partial^{(p)} x_j^{(k+1)} \cdot \partial_{x_i} f_j^{(k)}\left(x^{(k)}, w^{(k)}\right), & \text{otherwise} \end{cases}$$

$$\partial^{(p)} w_i^{(k)} = \sum_j \partial^{(p)} x_j^{(k+1)} \cdot \partial_{w_i} f_j^{(k)}\left(x^{(k)}, w^{(k)}\right)$$

$$\partial_{x_i} f_j^{(k)} = \frac{\partial f_j^{(k)}}{\partial x_i^{(k)}} = \frac{\partial x_j^{(k+1)}}{\partial x_i^{(k)}}$$

$$\partial_{w_i} f_j^{(k)} = \frac{\partial f_j^{(k)}}{\partial w_i^{(k)}} = \frac{\partial x_j^{(k+1)}}{\partial w_i^{(k)}}$$

である。

誤差逆伝播法の具体的な手順は以下の通りである。

Step1　教師パターンの組 $(x'^{(p)}, t^{(p)})$ を P 個用意する。学習率 η を定める。

Step2　ある p に対して

$$x_i^{(1)} = x_i'^{(p)}$$

とし、k を 2 から N まで増やしながら

$$x_i^{(k)} = f_i^{(k-1)}\left(x^{(k-1)}, w^{(k-1)}\right)$$

を計算する。

Step3　ネットワークの出力と教師パターンとの誤差から

$$\partial^{(p)} x_i^{(N)} = \partial_{x_i} L\left(t^{(p)}, x^{(N)}\right)$$

を計算し、k を $N-1$ から 1 まで減らしながら

$$\partial^{(p)} x_i^{(k)} = \sum_j \partial^{(p)} x_j^{(k+1)} \cdot \partial_{x_i} f_j^{(k)}\left(x^{(k)}, w^{(k)}\right)$$
$$\partial^{(p)} w_i^{(k)} = \sum_j \partial^{(p)} x_j^{(k+1)} \cdot \partial_{w_i} f_j^{(k)}\left(x^{(k)}, w^{(k)}\right)$$

を計算する。

Step4　いくつかの p に対して Step2 から Step3 を繰り返し行い、重みの更新量

$$\Delta w_i^{(k)} = -\eta \sum_p \partial^{(p)} w_i^{(k)}$$

を計算し重みを更新する。

Step5　繰り返し回数の上限を越えた場合やネットワークの出力と教師パターンの誤差

$$E = \sum_p L\left(t^{(p)}, x^{(N)}\right)$$

が一定の値以下になった場合に終了する。そうでない場合 Step2 へ戻る。

3　全結合層 (fully-connected layer)

ニューロン数 n の層からニューロン数 m の層への全結合層の順伝播は

$$y_j = f(x, w) = \sum_i x_j w_{ij}$$

と書くことができる。ここで i は前の層のニューロンの番号、j は次の層のニューロンの番号、w_{ij} は i 番目のニューロンから j 番目のニューロンへの結合荷重である。前節の式とてらしあわせると

$$\partial_{x_i} f_j(x, w) = \frac{\partial}{\partial x_i} \sum_k x_k w_{kj} = \begin{cases} w_{ij}, & i = k \\ 0, & \text{otherwise} \end{cases}$$
$$\partial_{w_{ik}} f_j(x, w) = \frac{\partial}{\partial w_{ik}} \sum_l x_l w_{lj} = \begin{cases} x_i, & i = l, j = k \\ 0, & \text{otherwise} \end{cases}$$

となる。

4　畳み込み層 (convolution layer)

前層と後層のニューロンが n 次元で構成され、各次元 i でのサイズが μ_i であるとする。このとき各次元 i に対してストライドを s_i、パディングのサイズを p_i、フィルタのサイズを f_i とすると、出力される信号長 ν_i は

$$\nu_i = \frac{\mu_i + 2p_i - f_i}{s_i} + 1$$

として計算される。ここで ν_i が整数とならないときを許容しない場合と、切り捨て等の処理を行って処理を継続する場合がある。入力 $x_{\gamma_1\gamma_2\cdots\gamma_n}$ が与えられたとき、計算に用いられるのはパディングによって埋められた値も含まれる。よって、任意の i に対して $-p_i \leq \gamma_i < \mu_i + p_i$ として

$$x'_{\gamma_1\gamma_2\cdots\gamma_n} = \begin{cases} x_{\gamma_1\gamma_2\cdots\gamma_n}, & \forall i, 0 \leq \gamma_i < \mu_i \\ v_{\gamma_1\gamma_2\cdots\gamma_n}, & \text{otherwise} \end{cases}$$

と書かれる $x'_{\gamma_1\gamma_2\cdots\gamma_n}$ を用いて計算されるとする。ここで v はパディングによって埋められる値である。入力 $x_{\gamma_1\gamma_2\cdots\gamma_n}$ とフィルタ $w_{\alpha_1\alpha_2\cdots\alpha_n}$ のよって出力 $y_{\beta_1\beta_2\cdots\beta_n}$ は

$$y_{\beta_1\beta_2\cdots\beta_n} \sum_{\alpha_1\alpha_2\cdots\alpha_n} x'_{\gamma_1\gamma_2\cdots\gamma_n} w_{\alpha_1\alpha_2\cdots\alpha_n}$$

で計算され、各 γ_i は α_i と β_i によって

$$\gamma_i = s_i\beta_i + f_i - p_i - \alpha_i - 1$$

と書かれる。

畳み込み層の重みと入力による偏導関数は

$$\partial_{x_{\gamma_1\gamma_2\cdots\gamma_n}} f_{\beta_1\beta_2\cdots\beta_n}(x, w) = \begin{cases} w_{\alpha_1\alpha_2\cdots\alpha_n}, & \forall i, \frac{\gamma_i - f_i + p_i + 1}{s_i} \leq \beta_i < \frac{\gamma_i + p_i + 1}{s_i} \\ 0, & \text{otherwise} \end{cases}$$

$$\partial_{w_{\alpha_1\alpha_2\cdots\alpha_n}} f_{\beta_1\beta_2\cdots\beta_n}(x, w) = \begin{cases} x'_{\gamma_1\gamma_2\cdots\gamma_n}, & \forall i, \frac{\alpha_i - f_i + 1}{s_i} \leq \beta_i < \frac{\mu_i + \alpha_i - f_i + 2p_i + 1}{s_i} \\ 0, & \text{otherwise} \end{cases}$$

となる。