ĆW 6. Stochastic Gradient Descent (SGD)

Cel stochastycznej metody gradientu (SGD) dla BP, XOR:

$$E(s_i, w_{ij}) = \sum_{p=1}^{4} E_p(s_i, w_{ij}) \to \text{minimum lokalne, gdzie}\, E_p(s_i, w_{ij}) = \frac{1}{2}(y(p) - z(p))^2$$

Algorytm SGD
Ustalić stałe $c > 0$, $\varepsilon > 0$ i $\beta > 0$.
($c = 0.01(\sim 1.0)$, $\varepsilon = 0.000(00)1$, $\beta = 1.0(\sim 3.0)$)

(i)
Startować od

$$\begin{array}{lll} w_{11}^{\text{old}} = 0.0 & w_{12}^{\text{old}} = 1.0 & w_{13}^{\text{old}} = 2.0 \\ w_{21}^{\text{old}} = 0.0 & w_{22}^{\text{old}} = 1.0 & w_{23}^{\text{old}} = 2.0 \\ s_1^{\text{old}} = 0.0 & s_2^{\text{old}} = 1.0 & s_3^{\text{old}} = 2.0 \end{array}$$

(ii)
Losować p **(**$1 \leq p \leq 4$**).**
Niech

$$s_i^{\text{new}} = s_i^{\text{old}} - c\,\frac{\partial E_p}{\partial s_i}(s_i^{\text{old}}, 1 \leq i \leq 3; w_{ij}^{\text{old}}, 1 \leq i \leq 2, 1 \leq j \leq 3) \quad (1 \leq i \leq 3)$$

$$w_{ij}^{\text{new}} = w_{ij}^{\text{old}} - c\,\frac{\partial E_p}{\partial w_{ij}}(s_i^{\text{old}}, 1 \leq i \leq 3; w_{ij}^{\text{old}}, 1 \leq i \leq 2, 1 \leq j \leq 3) \quad (1 \leq i \leq 2, 1 \leq j \leq 3).$$

(iii)
while

$$\max\left\{|s_i^{\text{new}} - s_i^{\text{old}}|,\ |w_{ij}^{\text{new}} - w_{ij}^{\text{old}}|\right\} > \varepsilon$$

do
Losować p **(**$1 \leq p \leq 4$**).**
Niech

$$s_i^{\text{old}} = s_i^{\text{new}} \quad (1 \leq i \leq 3)$$

$$w_{ij}^{\text{old}} = w_{ij}^{\text{new}} \quad (1 \leq i \leq 2, 1 \leq j \leq 3)$$

$$s_i^{\text{new}} = s_i^{\text{old}} - c\,\frac{\partial E_p}{\partial s_i}(s_i^{\text{old}}, 1 \leq i \leq 3; w_{ij}^{\text{old}}, 1 \leq i \leq 2, 1 \leq j \leq 3) \quad (1 \leq i \leq 3)$$

$$w_{ij}^{\text{new}} = w_{ij}^{\text{old}} - c\,\frac{\partial E_p}{\partial w_{ij}}(s_i^{\text{old}}, 1 \leq i \leq 3; w_{ij}^{\text{old}}, 1 \leq i \leq 2, 1 \leq j \leq 3) \quad (1 \leq i \leq 2, 1 \leq j \leq 3).$$

(iv)
Wyświetlić $y(p) \in \mathbb{R}$ $(1 \leq p \leq 4)$, $s_i^{\text{new}}, 1 \leq i \leq 3$; $w_{ij}^{\text{new}}, 1 \leq i \leq 2, 1 \leq j \leq 3$ dla ostatniego kroku.
Koniec algorytmu

• Wejścia (input): $u(p) = (u_1(p), u_2(p), u_3(p)) \in \{0,1\}^3 \subset \mathbb{R}^3$, $\quad u_3(p) \equiv 1 \quad (1 \leq p \leq 4)$
$u(1) = (0,0,1), u(2) = (1,0,1), u(3) = (0,1,1), u(4) = (1,1,1)$
• Sygnały nauczyciela (XOR): $z(p) \in \{0,1\} \subset \mathbb{R} \quad (1 \leq p \leq 4)$
$z(1) = 0, z(2) = 1, z(3) = 1, z(4) = 0 \quad$ (XOR)

• Sygnały z środkowej warstwy: $x(p) = (x_1(p), x_2(p), x_3(p)) \in \mathbb{R}^3, \quad x_3(p) \equiv 1 \quad (1 \le p \le 4)$
$f(x) = \frac{1}{1+e^{-\beta x}}$

$$x_i(p) = f\left(\sum_{j=1}^{3} w_{ij}^{\mathrm{old}} u_j(p)\right) \quad (1 \le i \le 2)$$

• Wyjścia (output): $y(p) \in \mathbb{R} \quad (1 \le p \le 4)$

$$y(p) = f\left(\sum_{i=1}^{3} s_i^{\mathrm{old}} x_i(p)\right)$$

• Gradienty
$f'(x) = \frac{\beta e^{-\beta x}}{(1+e^{-\beta x})^2} = \beta f(x)\{1 - f(x)\}$

$$\frac{\partial E_p}{\partial s_i}(s_i^{\mathrm{old}}, 1 \le i \le 3; w_{ij}^{\mathrm{old}}, 1 \le i \le 2, 1 \le j \le 3) = (y(p) - z(p)) f'\left(\sum_{k=1}^{3} s_k^{\mathrm{old}} x_k(p)\right) x_i(p)$$

$(1 \le i \le 3)$

$$\frac{\partial E_p}{\partial w_{ij}}(s_i^{\mathrm{old}}, 1 \le i \le 3; w_{ij}^{\mathrm{old}}, 1 \le i \le 2, 1 \le j \le 3)$$
$$= (y(p) - z(p)) f'\left(\sum_{k=1}^{3} s_k^{\mathrm{old}} x_k(p)\right) s_i^{\mathrm{old}} f'\left(\sum_{\ell=1}^{3} w_{i\ell}^{\mathrm{old}} u_\ell(p)\right) u_j(p)$$

$(1 \le i \le 2, 1 \le j \le 3)$

Zadanie. (Propagacja wsteczna błędu z SGD)
Implementować powyżej podany algorytm SGD dla propagacji wstecznej, który realizuje uczenie parametrów (wag) dla XORa.
Próbować i porównać różne parametry c, ε, β.

Notacja. (Propozycja)
$u_j(p) \rightsquigarrow u[p][j],\ x_i(p) \rightsquigarrow x[p][i],\ y(p) \rightsquigarrow y[p],\ z(p) \rightsquigarrow z[p]$
$s_i^{\mathrm{old}} \rightsquigarrow s\_old[i],\ s_i^{\mathrm{new}} \rightsquigarrow s\_new[i],\ w_{ij}^{\mathrm{old}} \rightsquigarrow w\_old[i][j],\ w_{ij}^{\mathrm{new}} \rightsquigarrow w\_new[i][j]$
$\frac{\partial E_p}{\partial s_i} \rightsquigarrow DEp\_s[i][p],\ \frac{\partial E_p}{\partial w_{ij}} \rightsquigarrow DEp\_w[i][j][p]$
$f \rightsquigarrow f,\ f' \rightsquigarrow Df,\ e^{-\beta x} \rightsquigarrow math.exp((-1) * beta * x)$ (Python?)
$c \rightsquigarrow c,\ \varepsilon \rightsquigarrow epsilon,\ \beta \rightsquigarrow beta$