

Sztuczne sieci neuronowe

Wykład 4: Algorytmy optymalizacji

Małgorzata Krętowska
Katedra Oprogramowania
e-mail: mmac@ii.pb.bialystok.pl

Plan wykładu

- Algorytmy gradientowe optymalizacji
 - Algorytm największego spadku
 - Algorytm zmiennej metryki
 - Algorytm gradientów sprzężonych
- Algorytmy doboru współczynnika uczenia
 - adaptacyjny dobór współczynnika uczenia
 - dobór współczynnika przez minimalizację kierunkową
 - reguła delta-bar-delta
 - metoda gradientów sprzężonych z regularyzacją
- Algorytmy heurystyczne
 - algorytm Quickprop
 - algorytm RPROP

Sztuczne sieci neuronowe2

Uczenie z nauczycielem

- Minimalizacja funkcji celu E
- Zakładając ciągłą funkcję aktywacji, minimalizacja odbywa się metodami gradientowymi
- W każdym kroku uczenia wyznacza się tzw. kierunek minimalizacji p ($W(k)$)
- Korekcja wag odbywa się według wzoru:

$$W(k+1) = W(k) + \eta p(W(k))$$

gdzie η jest współczynnikiem uczenia z przedziału $[0, 1]$.

Algorytmy gradientowe optymalizacji

Algorytmy gradientowe bazują na rozwinięciu w szereg Taylora funkcji celu $E(W)$ w najbliższym sąsiedztwie znanego rozwiązania $W = [w_1, w_2, \dots, w_n]^T$ (na starcie algorytmu jest to punkt początkowy W_0):

$$E(W + p) = E(W) + [g(W)]^T p + \frac{1}{2} p^T H(W) p + \dots$$

gdzie:

$$g(W) = \nabla E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \dots, \frac{\partial E}{\partial w_n} \right]^T$$
$$H(W) = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1 \partial w_1} & \dots & \frac{\partial^2 E}{\partial w_n \partial w_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 E}{\partial w_1 \partial w_n} & \dots & \frac{\partial^2 E}{\partial w_n \partial w_n} \end{bmatrix}$$

Algorytmy gradientowe optymalizacji

- Punkt $W=W_k$ jest punktem optymalnym funkcji $E(W)$, jeśli
 - $g(W_k)=0$
 - hesjan $H(W_k)$ jest dodatnio określony
- W praktyce (ze względu na skończoną dokładność obliczeń) zakłada się, że punkt W_k jest punktem optymalnym, jeżeli:

$$\begin{aligned} E(W_{k-1}) - E(W_k) &\leq \tau(1 + |E(W_k)|) \\ \|W_{k-1} - W_k\| &\leq \sqrt{\tau}(1 + \|W_k\|) \\ \|g(W_k)\| &\leq \sqrt[3]{\tau}(1 + |E(W_k)|) \end{aligned}$$

gdzie τ przyjęta dokładność obliczeń

Ogólny algorytm optymalizacji

Zakładamy $W_0=W_k$

- Test: jeżeli W_k spełnia warunki testowe jest punktem optymalnym to kończymy obliczenia, w przeciwnym przypadku pkt. 2
- Wyznaczanie wektora kierunku poszukiwań p_k w punkcie W_k .
- Minimalizacja kierunkowa funkcji $E(W)$ na kierunku p_k w celu wyznaczenia takiej wartości η_k , aby $E(W_k + \eta_k p_k) < E(W_k)$
- Wyznaczenie nowego rozwiązania $W_{k+1} = W_k + \eta_k p_k$ oraz odpowiadającej mu wartości $E(W_k)$, $g(W_k)$ (i ew. $H(W_k)$) i powrót do pkt 1.

Różnice: wyznaczanie kierunku poszukiwań p oraz kroku η .

Algorytm największego spadku

Ograniczenie do liniowego przybliżenia funkcji $E(W)$ w najbliższym sąsiedztwie znanego rozwiązania W :

$$E(W + p) = E(W) + [g(W)]^T p + O(h^2)$$

aby $E(W_{k+1}) < E(W_k)$ wystarczy aby $[g(W_k)]^T p < 0$

Wektor kierunkowy w metodzie największego spadku przyjmuje postać:

$$p_k = -g(W_k)$$

Algorytm największego spadku

$$W_{k+1} = W_k + \Delta W_k$$

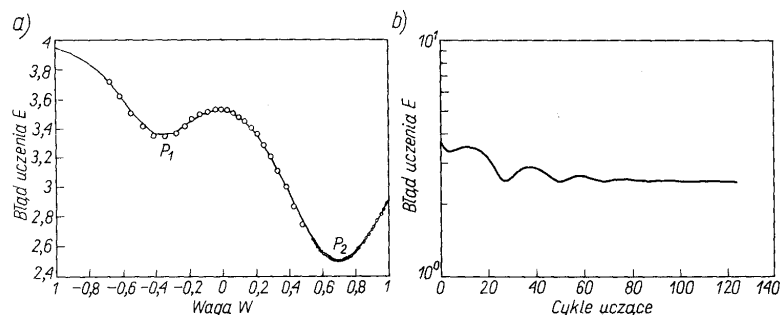
- Podejście klasyczne $\Delta W_k = \eta p_k$
- Metoda momentu $\Delta W_k = \eta p_k + \alpha(W_k - W_{k-1})$

Uwagi:

- na płaskich odcinkach $\Delta W_k = \frac{\eta}{1-\alpha} p_k$
(dla $\alpha=0.9$ oznacza to 10 krotne przyspieszenie procesu uczenia)
- pozwala na wyjście z minimów lokalnych
- należy kontrolować wartość E

Algorytm największego spadku

Wykres wpływu działania momentu na proces uczenia



- Metoda „weight decay” $\Delta W_k = \eta p_k - \beta W_k$
zabezpiecza przez zbyt dużym wzrostem wag

Algorytm zmiennej metryki (quasi-Newtona)

Kwadratowe przybliżenie funkcji $E(W)$ w sąsiedztwie znanego rozwiązania W_k :

$$E(W + p) = E(W) + [g(W)]^T p + \frac{1}{2} p^T H(W) p + O(h^3)$$

kierunek p jest wyznaczony ze wzoru:

$$p_k = -[H(W_k)]^{-1} g(W_k)$$

Problemy:

- wymóg dodatniej określoności hesjanu w każdym kroku

Rozwiązanie

- zastosowanie przybliżenia hesjanu przy użyciu metody zmiennej metryki

Algorytm zmiennej metryki (quasi-Newtona)

Przybliżenie hesjanu polega na modyfikacji hesjanu z kroku poprzedniego o pewną poprawkę, która powoduje, że aktualna wartość hesjanu $G(W_k)$ przybliży krzywiznę funkcji celu E zgodnie z zależnością:

$$G(W_k)(W_k - W_{k-1}) = g(W_k) - g(W_{k-1})$$

Na podstawie powyższego założenia można otrzymać wzory określające hesjan w kroku k -tym:

$$V_k = V_{k-1} \left[1 + \frac{r_k^T V_{k-1} r_k}{s_k^T r_k} \right] \frac{s_k s_k^T}{s_k^T r_k} - \frac{s_k r_k^T V_{k-1} + V_{k-1} r_k s_k^T}{s_k^T r_k}$$

gdzie $s_k = W_k - W_{k-1}$;

$r_k = g(W_k) - g(W_{k-1})$;

$V_1 = [G(W_1)]^{-1}$.

Algorytm zmiennej metryki (quasi-Newtona)

- wartość startowa $V_0=1$
- pierwsza iteracja zgodnie z algorytmem największego spadku
- odtwarzana macierz hesjanu jest w każdym kroku dodatnio określona (stąd $g(W_k)=0$ odpowiada rozwiązaniu problemu optymalizacji)
- metoda uważana za jedną z najlepszych metod optymalizacji funkcji wielu zmiennych

Wady:

- stosunkowo duża złożoność obliczeniowa (n^2 elementów hesjanu)
- duże wymagania co do pamięci przy przechowywaniu macierzy hesjanu

Metoda gradientów sprzężonych

- rezygnacja z bezpośredniej informacji o hesjanie
- nowy kierunek poszukiwań ma być ortogonalny i sprzężony z poprzednim kierunkami p_0, p_1, \dots, p_{k-1} , stąd:

$$p_k = -g(W_k) + \sum_{j=0}^{k-1} \beta_{kj} p_j$$

co można uprościć do postaci:

$$p_k = -g(W_k) + \beta_{k-1} p_{k-1}$$

współczynnik sprzężenia ($\beta_k = \beta_k(W_k)$):

$$\beta_{k-1} = \frac{g_k^T (g_k - g_{k-1})}{g_{k-1}^T g_{k-1}}$$

Zbiór wektorów p_i jest wzajemnie sprzężony względem macierzy H , jeżeli

$$p_i^T H p_j = 0, i \neq j$$

Metoda gradientów sprzężonych

- metoda mniej skuteczna od metody zmiennej metryki, ale bardziej skuteczna niż metoda największego spadku
- stosuje się ją do optymalizacji przy bardzo dużej liczbie zmiennych
- ze względu na błędy zaokrągleń w trakcie ztraca się własność ortogonalności między wektorami kierunków minimalizacji. Po wykonaniu n iteracji przeprowadza się jej ponowny start (w I kroku zgodnie z algorytmem największego spadku)

Metody doboru współczynnika uczenia

Po określeniu właściwego kierunku p_k minimalizacji, należy dobrać odpowiednią wartość współczynnika uczenia, aby nowy punkt W_{k+1} leżał możliwie najbliżej minimum funkcji $E(W)$ na kierunku p_k

$$W_{k+1} = W_k + \eta_k p_k$$



Stały współczynnik uczenia

- Stały współczynnik uczenia
 - stosuje się głównie w połączeniu z metodą największego spadku
 - sposób najmniej efektywny, gdyż nie uzależnia wartości współczynnika od wektora gradientu oraz kierunku poszukiwań p w danej iteracji
 - algorytm ma skłonność utykania w minimach lokalnych
 - często dobór współczynnika odbywa się oddzielnie dla każdej warstwy, przyjmując

$$\eta \leq \min \left(\frac{1}{n_i} \right)$$

gdzie n_i liczba wejść i -tego neuronu w warstwie

Adaptacyjny dobór współczynnika uczenia

- zmiany współczynnika uczenia dopasowują się do aktualnych zmian wartości funkcji celu w czasie. Wartość błędu ε w i-tej iteracji:

$$\varepsilon = \sqrt{\sum_{j=1}^M (y_j - d_j)^2}$$

określa strategię zmian wartości współczynnika uczenia.

- Przyspieszenie procesu uczenia uzyskuje się poprzez ciągłe zwiększanie współczynnika η sprawdzając jednocześnie czy nie zacznie wzrastać w porównaniu z błędem obliczonym przy poprzedniej wartości η

Adaptacyjny dobór współczynnika uczenia

Adaptacja współczynnika uczenia:

$$\eta_{i+1} = \begin{cases} \eta_i \rho_d & \text{gdy } \varepsilon_i > k_w \varepsilon_{i-1} \\ \eta_i \rho_i & \text{gdy } \varepsilon_i \leq k_w \varepsilon_{i-1} \end{cases}$$

gdzie:

ε_{i-1} , ε_i - błąd odpowiednio w (i-1)-szej iteracji oraz w i-tej iteracji

η_{i-1} , η_i - współczynnik uczenia w kolejnych iteracjach

k_w - dopuszczalny współczynnik wzrostu błędu

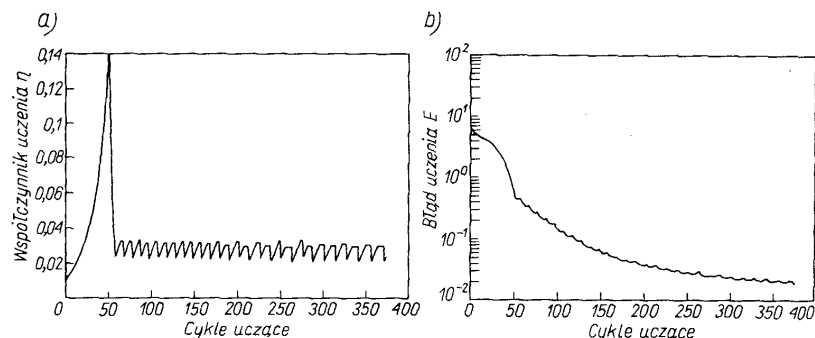
ρ_d - współczynnik zmniejszania wartości

ρ_i - współczynnik zwiększający wartość

Przykładowe wartości współczynników: $k_w = 1,04$; $\rho_d = 0,7$; $\rho_i = 1,05$

Adaptacyjny dobór współczynnika uczenia

Wpływ adaptacyjnego doboru współczynnika uczenia na proces uczenia



Dobór współczynnika uczenia przez minimalizację kierunkową

- Polega na minimalizacji kierunkowej funkcji celu na wyznaczonym wcześniej kierunku p_k .
- Cel: takie dobranie wartości η_k aby nowy punkt $W_{k+1} = W_k + \eta_k p_k$ odpowiadał minimum funkcji celu na danym kierunku
 - Jeżeli η_k odpowiada dokładnie minimum funkcji na danym kierunku p_k to pochodna kierunkowa w punkcie $W_{k+1} = W_k + \eta_k p_k$ musi być równa 0
- W praktyce wyznaczony punkt W_{k+1} odpowiada tylko w przybliżeniu rzeczywistemu punktowi minimalnemu na danym kierunku.

Dobór współczynnika uczenia przez minimalizację kierunkową

W celu „regulacji” dokładności wyznaczenia współczynnika uczenia wprowadza się współczynnik $0 < \gamma_2 < 1$, który stanowi ułamek pochodnej funkcji celu na kierunku p_k w punkcie wyjściowym W_k .

Algorytm pozwalający na wyznaczenie optymalnej wartości η_k przeprowadza się dopóty, dopóki spełnione są następujące warunki:

$$[g(W_k + \eta_k p_k)]^T p_k \geq \gamma_2 [g(W_k)]^T p_k$$

oraz

$$E(W_k + \eta_k p_k) - E(W_k) \geq \gamma_1 \eta_k [g(W_k)]^T p_k$$

przyjęcie $0 \leq \gamma_1 \leq \gamma_2 < 1$ gwarantuje jednoczesne spełnienie obu tych warunków.

Minimalizacja kierunkowa

- Metody bezgradientowe
 - informacje o wartościach funkcji celu
 - wyznaczanie minimum poprzez kolejne podziały założonego na wstępie zakresu wartości wektora W
- Metody gradientowe
 - wykorzystują zarówno wartość funkcji jak też jej pochodną wzdłuż wektora kierunku p_k .
 - znaczne przyspieszenie wyznaczenia minimum na danym kierunku (informacja o kierunku spadku)

Przykład metody bezgradientowej

- Metoda bazuje na aproksymacji funkcji celu na kierunku p_k , a następnie wyznacza minimum otrzymanej w ten sposób funkcji jednej zmiennej η
- Wielomian aproksymujący:

$$P(\eta) = a_2 \eta^2 + a_1 \eta + a_0$$

gdzie a_2, a_1, a_0 - współczynniki wielomianu określone w każdym cyklu optymalizacyjnym

- Wyznaczanie współczynników wielomianu
 - wybór trzech dowolnych punktów W_1, W_2, W_3 leżących na kierunku p_k , tzn. $W_1 = W + \eta_1 p_k$; $W_2 = W + \eta_2 p_k$; $W_3 = W + \eta_3 p_k$; (W - poprzednie rozwiązanie);
 - $E_1 = E(W_1)$; $E_2 = E(W_2)$; $E_3 = E(W_3)$; wówczas
- $$P(\eta_1) = E_1; P(\eta_2) = E_2; P(\eta_3) = E_3;$$
- Rozwiązując układ równań otrzymujemy współczynniki wielomianu
- Porównując pochodną wielomianu do zera otrzymujemy $\eta_{\min} = (-a_1 / 2a_2)$
 - Po określeniu są sprawdzane warunki. Jeśli algorytm ma być kontynuowany to wybiera się kolejne punkty leżące na kierunku p_k w pobliżu punktu $W + \eta_{\min} p_k$.

Inne metody doboru współczynnika uczenia

- Reguła delta-bar-delta
 - jest metodą adaptacyjną opracowaną dla kwadratowej definicji funkcji celu i metody największego spadku
 - każdej wadze W_{ij} jest przyporządkowany indywidualnie dobrany współczynnik uczenia
 - Wada: duża złożoność obliczeniowa
 - Zaleta: przyspieszenie procesu uczenia i zwiększenie prawdopodobieństwa osiągnięcia minimum globalnego
- Metoda gradientów sprzężonych z regularyzacją
 - odmiana zwykłej metody gradientów sprzężonych łącząca jednocześnie wyznaczanie kierunku p oraz optymalnego kroku

Algorytm Quickprop

- odmiana algorytmu gradientowego zawiera elementy metody newtonowskiej i wiedzy heurystycznej
- zawiera elementy zabezpieczające przed utknięciem w płytkim minimum lokalnym (ze względu na nasycenie neuronu)
- Zmiana wagi w k-tym kroku

$$\Delta W_{ij}(k) = -\eta_k \left[\frac{\partial E(W(k))}{\partial W_{ij}} + \mathcal{W}_{ij}(k) \right] + \alpha_{ij}^k \Delta W_{ij}(k-1)$$

- Zalety: szybka zbieżność dla większości trudnych problemów
- kilkusetkrotne przyspieszenie procesu uczenia (w porównaniu z algorytmem największego spadku)
- małe prawdopodobieństwo utknięcia w minimum lokalnym

Algorytm RPROP (ang. *Resilient backPROPagation*)

$$\Delta W_{ij}(k) = -\eta_{ij}^{(k)} \operatorname{sgn} \left[\frac{\partial E(W(k))}{\partial W_{ij}} \right]$$

$$\eta_{ij}^{(k)} = \begin{cases} \min(a\eta_{ij}^{(k-1)}, \eta_{\max}) & \text{dla } S_{ij}(k)S_{ij}(k-1) > 0 \\ \max(b\eta_{ij}^{(k-1)}, \eta_{\min}) & \text{dla } S_{ij}(k)S_{ij}(k-1) < 0 \\ \eta_{ij}^{(k-1)} & \text{w pozostałych przypadkach} \end{cases}$$

gdzie

$$S_{ij}(k) = \frac{\partial E(W(k))}{\partial W_{ij}}$$

– a=1.2; b=0.5

η_{\min} ; η_{\max} - minimalna i maksymalna wartość współczynnika uczenia (10^{-6} ; 50)

Zalety

- przyspieszenie procesu uczenia w obszarach gdzie nachylenie funkcji celu jest niewielkie