Лекція № 5

Прискорені варіанти методу зворотного поширення похибки

Проф. Куссуль Н.М.

5.1 Алгоритм навчання Delta-Bar-Delta — некумулятивний алгоритм без моментів

Алгоритм Delta-bar-Delta (DBD) - один з способів прискорення методу зворотного поширення похибки.

Евристики:

- Для кожного вагового коефіцієнту зв'язку використовується свій власний коефіцієнт навчання, що обчислюється на кожнім кроці роботи алгоритму.
- Якщо часткова похідна функції похибки протягом декількох кроків роботи алгоритму зберігає свій знак, коефіцієнт навчання для даного вагового коефіцієнту зв'язку збільшується.
- Якщо часткова похідна змінює знак на кожному кроці навчання, коефіцієнт швидкості навчання зменшується.

Перед початком процесу навчання вводиться початкове значення коефіцієнта навчання, однакове для усіх вагових коефіцієнтів зв'язків.

5.1 Алгоритм навчання Delta-Bar-Delta — некумулятивний алгоритм без моментів

Навчання здійснюється відповідно до алгоритму (пункти 1–7) з розділу 4.6, однак при переході до пункту 6 для кожного з образів навчальної вибірки здійснюється перерахування коефіцієнтів навчання

Позначимо $$g_{ij}(t) = \frac{\partial E}{\partial W_{ij}}(t) \quad (1)$$

Експоненціальне середнє значення градієнта для даного моменту часу

$$\tilde{g}_{ij}(t) = (1-\theta)\frac{\partial E}{\partial W_{ij}}(t) + \theta \cdot \tilde{g}_{ij}(t-1) \quad (2)$$

$$\Delta\eta_{ij}(t) = \begin{cases} k, & \text{якщо } \tilde{g}_{ij}(t-1)\cdot g_{ij}(t) > 0, \\ -\varphi\cdot\eta_{ij}(t-1), & \text{якщо } \tilde{g}_{ij}(t-1)\cdot g_{ij}(t) < 0, \\ 0 & \text{в інших випадках.} \end{cases} \quad (3)$$

$$k \in [0.03,\ 0.1],\ \varphi \in [0.1, 0.3],\ \theta \approx 0.7 \quad (4)$$

5.1 Алгоритм навчання Delta-Bar-Delta (продовження)

$$\eta_{i\,j}(t) = \eta_{i\,j}(t-1) + \Delta\eta_{i\,j}(t)$$

$$W_{i\,j}(t+1) = W_{i\,j}(t) - \eta_{ij}(t)\frac{\partial E}{\partial W_{ij}}.$$

Переваги та недоліки

+ Висока швидкість збіжності (на порядок менше епох, ніж у класичному методі)

+ Залежність лише від 3 параметрів (невисока чутливість - робастність)

- Значна обчислювальна складність кожної ітерації

5.2. Обґрунтування евристики зміни коефіцієнту навчання

Для простоти припустимо, що вихідний шар складається з 1 нейрона. Тоді середньоквадратична похибка на епосі складає $E(t)=\frac{1}{2}\sum_{i=1}^{t}(d_i-y_i)^2$ (5.4)

Похідна похибки за коефіцієнтом навчання η_{ij} $\frac{\partial E}{\partial \eta_{ij}}(t)=\frac{\partial E}{\partial y_i}(t)\cdot\frac{\partial y_i}{\partial S_i}(t)\cdot\frac{\partial S_i}{\partial \eta_{ij}}(t)$ (5.5)

де $S_i=\sum_j W_{ij}(t)y_j(t)$ — зважена сума i -го входу, $y_i(t)=f(S_i(t))$ — вихід нейрона.

Оскільки $W_{ij}(t)=W_{ij}(t-1)-\eta_{ij}(t)\cdot\frac{\partial E}{\partial W_{ij}}(t-1)$ (5.6) то $S_i(t)=\sum_j y_j(t)\left[W_{ij}(t-1)-\eta_{ij}(t)\cdot\frac{\partial E}{\partial W_{ij}}(t-1)\right]$ (5.7)

Диференціюючи (5.7) по $\eta_{ij}(t)$, одержимо $\frac{\partial S_i}{\partial \eta_{ij}}(t)=-y_j(t)\cdot\frac{\partial E}{\partial W_{ij}}(t-1)$ (5.8)

З огляду на формули обчислення похідних згідно методу зворотного розповсюдження

$\delta_i=\frac{\partial E}{\partial S_i}=\frac{\partial E}{\partial y_i}\cdot\frac{\partial y_i}{\partial S_i}$ і $\frac{\partial E}{\partial W_{ij}}=-\delta_i y_j$ співвідношення (5.5) можна переписати у вигляді

$$\frac{\partial E}{\partial \eta_{ij}}(t)=\frac{\partial E}{\partial y_j}(t)\cdot\frac{\partial y_i}{\partial S_i}(t)\cdot\frac{\partial S_i}{\partial \eta_{ij}}(t)=$$

$$=\delta_i(t)\left(-y_j(t)\cdot\frac{\partial E}{\partial W_{ij}}(t-1)\right)=-\frac{\partial E}{\partial W_{ij}}(t)\cdot\frac{\partial E}{\partial W_{ij}}(t-1).\quad (5.9)$$

Таким чином, похідна функції похибки по коефіцієнту навчання η_{ij} — це взятий зі знаком "-" добуток поточної і попередньої похідних по ваговим коефіцієнтам .

У цих евристиках використовується похідна функції похибки за *коефіцієнтом навчання*, а не довільне необгрунтоване значення.

В алгоритмі Delta-Bar-Delta це співвідношення згладжується шляхом взяття експонентного середнього $\frac{\partial E}{\partial W_{ij}}(t)$

$$\bar{\delta}(t)=(1-\theta)\frac{\partial E}{\partial W_{ij}}(t)+\theta\cdot\bar{\delta}(t-1).\quad (5.10)$$

5.3. Алгоритм навчання Enhanced Delta-Bar-Delta (EDBD)

Основна властивість - автоматичний вибір фактора моменту (водночас з коефіцієнтом швидкості навчання) для кожного вагового коефіцієнту зв'язку мережі на кожному кроці навчання.

Обчислюється експоненціальне середнє значення градієнта для даного моменту часу:

$$\bar{g}_{ij}(t) = (1-\theta)\cdot g_{ij}(t) + \theta \cdot \bar{g}_{ij}(t-1) \qquad (5.11)$$

$$\Delta\eta_{ij}(t) = \begin{cases} k_1(L)\cdot e^{-y(L)\left|\bar{g}_{ij}(t)\right|}, & \bar{g}_{ij}(t-1)\cdot g_{ij}(t) > 0 \\ -k_2(L)\cdot \eta_{ij}^{k}(t-1), & \bar{g}_{ij}(t-1)\cdot g_{ij}(t) < 0 \\ 0 & \bar{g}_{ij}(t-1)\cdot g_{ij}(t) = 0 \end{cases} \qquad (5.12)$$

$$\Delta\mu_{ij}(t) = \begin{cases} k_1(M)\cdot e^{-y(M)\left|\bar{g}_{ij}(t)\right|}, & \bar{g}_{ij}(t-1)\cdot g_{ij}(t) > 0 \\ -k_2(M)\cdot \mu_{ij}^{k}(t-1), & \bar{g}_{ij}(t-1)\cdot g_{ij}(t) < 0 \\ 0 & \bar{g}_{ij}(t-1)\cdot g_{ij}(t) = 0 \end{cases}$$

Позначення в алгоритмі Enhanced Delta-Bar-Delta (EDBD)

k номер шару;

η_{ij}^{k} коефіцієнт швидкості навчання на кроці t^*;

μ_{ij}^{k} коефіцієнт моменту на кроці t^*;

$\Delta L_{ij}^{k}(t)$ інкремент коефіцієнта швидкості навчання на кроці t^*;

$\Delta M_{ij}^{k}(t)$ інкремент коефіцієнта моменту на кроці t^*;

$\overline{G}_{ij}^{k}(t)$ середнє значення градієнта на кроці t^*;

$Y(L)$ експоненціальний фактор для швидкості навчання;

$Y(M)$ експоненціальний фактор для моменту;

$k_1(L),\ k_2(L)$ фактори масштабування для швидкості навчання;

$k_1(M),\ k_2(M)$ фактори масштабування для моменту;

θ Фактор впливу попередньої величини градієнта;

Примітка. * — окремий коефіцієнт для кожного зв'язку.

5.3. Enhanced Delta-Bar-Delta (EDBD) – продовження

$$\eta_{ij}^{k} = \eta_{ij}^{k}(t-1) + \Delta\eta_{ij}(t)$$
$$\mu_{ij}^{k} = \mu_{ij}^{k}(t-1) + \Delta\mu_{ij} \qquad (5.13)$$

$$\eta_{ij}^{k}(t) > L_{\max} \Rightarrow \eta_{ij}^{k}(t) = L_{\max}$$
$$\mu_{ij}^{k}(t) > M_{\max} \Rightarrow \mu_{ij}^{k}(t) = M_{\max}$$

$$\Delta W_{ij}(t) = -\eta_{ij}(t)\frac{\partial E}{\partial \overline{W}_{ij}}(t) + \mu_{ij}(t)\Delta\overline{W}_{ij}(t-1) \qquad (5.14)$$

Переваги та недоліки DBD та EDBD

- **DBD**
  - **+** Адаптивне налаштування коефіцієнту навчання
  - **+** Невелика кількість параметрів (3)
- **EDBD**
  - **+** За рахунок використання фактора моменту даний алгоритм є більш ефективним, ніж його спрощений аналог Delta-Bar-Delta.
  - **-** Однак робота даного алгоритму визначається досить великим числом параметрів, вибір яких повинен здійснюватися з урахуванням особливостей кожної розв'язуваної задачі.

Пружне поширення — Resilient Propagation (RPROP)

- Головна відмінність цього методу полягає в тому, що налаштування вагових коефіцієнтів визначається **тільки знаком** градієнтів, а не їхніми амплітудами.
- Це пояснюється тим, що **амплітуда** градієнта залежить від **значення функції похибки** і може значно мінятися від одного кроку до іншого.
- Крок навчання в алгоритмі пружного поширення RPROP (Resilient Propagation) визначається для кожного вагового коефіцієнта в кожен момент часу.

5.4. Пружне поширення — Resilient Propagation

$$\Delta_{ij}(t)=\begin{cases}\eta^{+}\cdot\Delta_{ij}(t-1), & \text{якщо } \dfrac{\partial E}{\partial W_{ij}}(t-1)\cdot\dfrac{\partial E}{\partial W_{ij}}(t)>0,\\ \eta^{-}\cdot\Delta_{ij}(t-1), & \text{якщо } \dfrac{\partial E}{\partial W_{ij}}(t-1)\cdot\dfrac{\partial E}{\partial W_{ij}}(t)<0,\\ \Delta_{ij}(t-1), & \text{якщо } \dfrac{\partial E}{\partial W_{ij}}(t-1)\cdot\dfrac{\partial E}{\partial W_{ij}}(t)=0.\end{cases} \qquad 0<\eta^{-}<1<\eta^{+}$$

$$\Delta W_{ij}(t)=\begin{cases}-\Delta_{ij}(t-1), & \text{якщо } \dfrac{\partial E}{\partial W_{ij}}(t)>0,\\ +\Delta_{ij}(t-1), & \text{якщо } \dfrac{\partial E}{\partial W_{ij}}(t)<0,\\ 0, & \text{якщо } \dfrac{\partial E}{\partial W_{ij}}(t)=0.\end{cases}$$

$$\Delta_{\min}=10^{-6},\quad \Delta_{\max}=50 \qquad \eta^{-}=0.5,\qquad \eta^{+}=1.2$$

Переваги та недоліки

+ Нечутливість до великих абсолютних значень градієтна
+ Висока ефективність в задачі класифікації двох спіралей
+ Хороша реалізація в Java-бібліотеці

- Необхідність обчислення градієнта
- Неробастність (чутливість до вибору параметрів)
- Нестабільна робота в різних задачах

5.5. Алгоритм навчання Quick Propagation

Алгоритм навчання Quick Propagation (QP) є одним з найпоширеніших способів прискорення методу зворотного розповсюдження похибки. На відміну від інших модифікацій методу, він *не використовує адаптивне налаштування* коефіцієнтів швидкості навчання, а базується на методі Ньютона мінімізації квадратичних функцій. Подібно до алгоритму зворотного поширення похибки, даний метод є локальним, тобто при налаштуванні параметрів мережі ваговий коефіцієнт кожного зв'язку розглядається незалежно від зміни інших. Метод Quick Propagation ґрунтується на припущенні про те, що функцію похибки можна апроксимувати увігнутою параболою.

Алгоритм QP визначається квадратичною залежністю:

$$\Delta W_{ij}(t) = \frac{g_{ij}(t)}{g_{ij}(t-1) - g_{ij}(t)} \cdot \Delta W_{ij}(t-1), \qquad (5.17)$$

де $g_{ij}(t) = \frac{\partial E}{\partial W_{ij}}(t)$ — похідна похибки по ваговому коефіцієнту, $\frac{g_{ij}(t) - g_{ij}(t-1)}{\Delta W_{ij}(t-1)}$ — кінцево-різницева апроксимація другої похідної.

Позначимо $\beta = \frac{g_{ij}(t)}{g_{ij}(t-1) - g_{ij}(t)}$

Формула (5.17) апроксимує метод Ньютона мінімізації одновимірної функції $f(x)$: $\Delta x = -\frac{f'(x)}{f''(x)}$

Можливі три випадки:

1. Градієнти на поточному та попередньому кроках мають однаковий знак і $g_{ij}(t-1) > g_{ij}(t)$, тоді $\beta > 0$ і ваговий коефіцієнт змінюється в тому ж напрямку.
2. Знаки поточної і попередньої похідної протилежні, значить, на попередньому кроці “проскочили” мінімум і потрібно рухатися в протилежному напрямку.
3. Градієнти на поточному та попередньому кроках мають однаковий знак і $|g_{ij}(t)| \geq |g_{ij}(t-1)|$. Тоді функція похибки погано апроксимується параболою або парабола не є увігнутою.

Щоб уникнути надмірної зміни вагових коефіцієнтів у випадку (3), вводиться коефіцієнт максимального зростання

$$\mu \approx 1.75: \quad \Delta W_{ij}(t) = \mu \cdot \Delta W_{ij}(t-1).$$

У випадках (1) і (3) вводиться ще один доданок $\eta \cdot g_{ij}(t)$, що відповідає кроку в напрямку простого антиградієнта, якщо $\Delta W_{ij}(t-1) = 0$

У випадку (2) це доповнення ігнорується, оскільки $g_{ij}(t) \neq 0$ і знаки похідних відрізняються. Цей випадок добре задовольняє припущення про квадратичність функції похибки.

5.6. Варіант QP, реалізований у системі Thinks

У системі моделювання НМ Thinks реалізовано інший варіант алгоритму QP. Навчання здійснюється відповідно до алгоритму (пункти 1–6) з розділу 4.6, однак при переході до пункту 5 для кожного з образів навчальної вибірки обчислюється інкремент (величина зміни) кожного вагового коефіцієнту зв'язку мережі за наступними формулами:

1. Якщо знак градієнта функції похибки на поточному $\frac{\partial E}{\partial W_{ij}}(t)$ і попередньому $\frac{\partial E}{\partial W_{ij}}(t-1)$ кроках алгоритму навчання для даного вагового коефіцієнту зв'язку збігаються, але по абсолютній величині $\frac{\partial E}{\partial W_{ij}}(t)$ перевершує значення градієнта на попередньому кроці $\frac{\partial E}{\partial W_{ij}}(t-1): \left|\frac{\partial E}{\partial W_{ij}}(t)\right| > \left|\frac{\partial E}{\partial W_{ij}}(t-1)\right|$, то величина зміни вагового коефіцієнту зв'язку обчислюється за формулою $\frac{\partial E}{\partial W_{ij}}(t) = \mu \frac{\partial E}{\partial W_{ij}}(t-1)$ де $\mu > 1$ — коефіцієнт прискорення навчання, що зазвичай приймають рівним $\mu = 1.75$. Обмеження кроку навчання забезпечує коректність поведінки алгоритму при невиконанні апріорної умови про апроксимацію функції похибки увігнутою параболою.
2. Якщо знак градієнта функції похибки на поточному $\frac{\partial E}{\partial W_{ij}}(t)$ і попередньому $\frac{\partial E}{\partial W_{ij}}(t-1)$ кроках алгоритму навчання для даного вагового коефіцієнту зв'язку збігаються, але по абсолютній величині $\frac{\partial E}{\partial W_{ij}}(t-1)$ незначно перевершує значення градієнта на поточному кроці $\frac{\partial E}{\partial W_{ij}}(t)$, або знаки градієнта функції похибки на поточному $\frac{\partial E}{\partial W_{ij}}(t)$ і попередньому $\frac{\partial E}{\partial W_{ij}}(t-1)$ кроках алгоритму навчання для даного вагового коефіцієнту зв'язку протилежні, то величина зміни вагового коефіцієнту зв'язку обчислюється за формулою $\Delta W^n(t) = \beta \cdot \Delta W^n(t)$ де $\beta = \frac{\frac{\partial E}{\partial W_{ij}}(t)}{\frac{\partial E}{\partial W_{ij}}(t-1) - \frac{\partial E}{\partial W_{ij}}(t)}$
3. Щоб забезпечити ненульовий крок алгоритму в потрібному напрямку у випадку $\Delta W^n(t) = 0$, до величини зміни вагового коефіцієнту зв'язку у випадку 2 додається ще один доданок, що забезпечує крок за алгоритмом градієнтного спуску: $\Delta W^n(t) = \Delta W^n(t) + \eta \cdot \frac{\partial E}{\partial W_{ij}}(t)$ де η — коефіцієнт швидкості навчання.