УДК 519.85
DOI: 10.20310/1810-0198-2017-22-3-552-557

О ГРАДИЕНТЕ НЕЙРОСЕТЕВОЙ ФУНКЦИИ

© **Н. М. Мишачев, А. М. Шмырин**

Липецкий государственный технический университет
398600, Российская Федерация, г. Липецк, ул. Московская, 30
E-mail: nmish@lipetsk.ru, amsh@lipetsk.ru

В статье предлагается матричная формула для вычисления градиента нейросетевой функции $\nabla_W f(X;W)$ по вектору оптимизируемых параметров W.
Ключевые слова: нейросетевая функция; нейронная сеть; алгоритм обратного распространения ошибки; умножение Адамара

Нейронная сеть прямого распространения (см., например, [1]) представляет собой композицию чередующихся линейных и покоординатных нелинейных отображений, при этом матрицы линейных отображений считаются параметрами сети. В отличие от обычных регрессионных моделей, оптимизируемые параметры входят в нейронную сеть нелинейным образом. Для оптимизации (обучения) нейронной сети нужно уметь вычислять градиент сети по вектору параметров. Формулы для градиента были получены рядом авторов (обычно принято ссылаться на [2]) и соответствующий метод оптимизации был назван *алгоритмом обратного распространения ошибок* (backpropagation). В рекуррентном координатном виде эти формулы имеются во всех руководствах по нейронным сетям. Нам не удалось найти в литературе матричную версию записи этих формул и мы предлагаем такую версию в данной заметке.

1. Нейросетевые функции.

Везде далее векторы $X\in\mathbb{R}^n$ считаются столбцами $[x_1,\ldots,x_n]^T$. Покомпонентное произведение (произведение Адамара) матриц A и B одинаковой размерности обозначается $A\circ B$. *Покоординатное* отображение $\Sigma:\mathbb{R}^n\to\mathbb{R}^n$, определяемое формулой

$$\Sigma(X)=\Sigma([x_1,\ldots,x_n]^T)=[\sigma_1(x_1),\ldots,\sigma_k(x_n)]^T,$$

т. е. прямая сумма n функций $\sigma_i:\mathbb{R}^1\to\mathbb{R}^1$, обозначается двойной стрелкой $\Sigma:\mathbb{R}^n\rightrightarrows\mathbb{R}^n$. Действие такого отображения на вектор X можно считать «операторным» произведением Адамара столбцов $\Sigma=[\sigma_1,\ldots,\sigma_k]^T$ и $X=[x_1,\ldots,x_n]^T$, т. е. $\Sigma(X)=\Sigma\circ X$. *Нейросетевой функцией* $f:\mathbb{R}^n\to\mathbb{R}^1$ мы называем функцию вида

$$f:\mathbb{R}^n\xrightarrow{W_1}\mathbb{R}^{n_1}\rightrightarrows^{\Sigma_1}\mathbb{R}^{n_1}\xrightarrow{W_2}\mathbb{R}^{n_2}\rightrightarrows^{\Sigma_2}\mathbb{R}^{n_2}\xrightarrow{W_3}\ldots\xrightarrow{W_{k-1}}\mathbb{R}^{n_{k-1}}\rightrightarrows^{\Sigma_k\Gamma}\mathbb{R}^{n_{k-1}}\xrightarrow{W_k}\mathbb{R}^1\to^{\Sigma_k}\mathbb{R}^1 \quad (1)$$

где W_i – линейные отображения и $\Sigma_i=[\sigma_{i1},\ldots,\sigma_{in_i}]^T$ – покоординатные отображения. Функции $\sigma_{ij}:\mathbb{R}^1\to\mathbb{R}^1$ называются *функциями активации*; их выбор важен в теории нейронных сетей, но мы здесь можем считать, что это произвольные кусочно дифференцируемые функции. Заметим, что $\Sigma_k=[\sigma_{k1}]^T=\sigma_k$, так как $n_k=1$. Отображение f можно записать в виде

$$f(X)=f(X;W)=\Sigma_k(W_k\cdot\Sigma_{k-1}(W_{k-1}\cdot\Sigma_{k-2}\cdot\cdots\Sigma_2(W_2\cdot\Sigma_1(W_1\cdot X))\ldots)), \qquad (2)$$

или в виде произведения (порядок действий – справа налево)

$$f(X) = f(X;W) = \Sigma_k \circ W_k \cdot \Sigma_{k-1} \circ W_{k-1} \cdot \Sigma_{k-2} \cdots \Sigma_2 \circ W_2 \cdot \Sigma_1 \circ W_1 \cdot X, \tag{3}$$

где $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$ и W_i – матрицы одноименных линейных отображений W_i. Эти матрицы мы считаем параметрами отображения $f(X;W)$, таким образом вектор параметров $W = (W_k, \ldots, W_1)$ состоит из $n_i \times n_{i-1}$ матриц W_i ($n_k = 1$, $n_0 = n$). Положим еще $N_1(X) = W_1 \cdot X$ и $N_{i+1}(X) = W_{i+1} \cdot \Sigma_i(N_i(X))$, тогда

$$f(X;W) = \underbrace{\Sigma_k(W_k \cdot \underbrace{\Sigma_{k-1}(W_{k-1} \cdot \Sigma_{k-2} \cdots \underbrace{\Sigma_2(W_2 \cdot \Sigma_1(\underbrace{W_1 \cdot X}_{N_1}))}_{N_2} \ldots)}_{N_{k-1}})}_{N_k}$$

и

$$f(X;W) = \Sigma_k(N_k) = \Sigma_k(W_k \cdot \Sigma_{k-1}(N_{k-1})) = \ldots ,$$

где N_i – столбец размерности n_i. Заметим, что при $k = 1$ и тождественной функции активации σ_1 нейросетевая функция является (однородной) линейной функцией n переменных.

2. Нейросетевые функции и нейронные сети.

Нейросетевая функция вида (1) задает *однородную* k-слойную нейронную сеть прямого распространения с n-мерным входом и одномерным выходом. Количество внутренних слоев равно $k-1$, при этом i-тый слой содержит n_i нейронов. Термин *однородная* в данном контексте не является общепринятым и указывает на то, что все промежуточные линейные отображения являются однородными. Чтобы получить из однородной сети обычную (неоднородную) сеть, в которой линейные отображения дополнены смещениями (bias), нужно потребовать, чтобы:

(a) все входные векторы имели вид $X = [x_1, \ldots, x_{n-1}, 1]^T$;

(b) последние строки всех матриц W_i с $i < k$ имели вид $[0, \ldots, 0, 1]$;

(c) последние функции σ_{in_i} в столбцах Σ_i с $i < k$ были тождественными.

В этом случае однородная сеть будет эквивалентна k-слойной неоднородной нейронной сети с $(n-1)$-мерным входом и одномерным выходом. Каждый внутренний слой такой сети будет содержать $n_i - 1$ «настоящих» нейронов и один (последний) «фиктивный», ответственный за смещения. При $k = 1$ и тождественной функции активации σ_1 описанный переход соответствует стандартному переходу от однородной множественной регрессии к обычной (неоднородной) путем добавления в множество независимых переменных (предикторов) дополнительного предиктора, тождественно равного единице.

3. Градиент $\nabla_W f(X;W)$ нейросетевой функции с $n = n_1 = \cdots = n_k = 1$.

Если $n = n_1 = \cdots = n_k = 1$, то все W_i - числа w_i, все Σ_i - функции σ_i и

$$f(x) = f(x;W) = \sigma_k(w_k \sigma_{k-1}(w_{k-1}\sigma_{k-2}(w_{k-2}\sigma_{k-3} \ldots \sigma_2(w_2\sigma_1(w_1 x)) \ldots))) . \tag{4}$$

В этом случае применение цепного правила дифференцирования не вызывает затруднений и для градиента

$$\nabla_W f = (\nabla_{w_k} f, \nabla_{w_{k-1}} f, \nabla_{w_{k-2}} f \ldots , \nabla_{w_1} f)$$

мы получаем

$$\begin{cases} \nabla_{w_k} f = \overbrace{\sigma'_k(N_k)\,\sigma_{k-1}(N_{k-1})} \\ \nabla_{w_{k-1}} f = \underbrace{\sigma'_k(N_k)\,w_k}\,\overbrace{\sigma'_{k-1}(N_{k-1})\,\sigma_{k-2}(N_{k-2})} \\ \nabla_{w_{k-2}} f = \underbrace{\sigma'_k(N_k)\,w_k}\,\underbrace{\sigma'_{k-1}(N_{k-1})\,w_{k-1}}\,\overbrace{\sigma'_{k-2}(N_{k-2})\,\sigma_{k-3}(N_{k-3})} \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \nabla_{w_1} f = \underbrace{\sigma'_k(N_k)\,w_k}\,\underbrace{\sigma'_{k-1}(N_{k-1})\,w_{k-1}}\,\underbrace{\sigma'_{k-2}(N_{k-2})\,w_{k-2}}\dots\overbrace{\sigma'_1(N_1)\,x} \end{cases} \tag{5}$$

(фигурные скобки указывают на периодичность в структуре формул) или, опуская для краткости аргументы N_i ,

$$\begin{cases} \nabla_{w_k} f = \overbrace{\sigma'_k\,\sigma_{k-1}} \\ \nabla_{w_{k-1}} f = \underbrace{\sigma'_k\,w_k}\,\overbrace{\sigma'_{k-1}\,\sigma_{k-2}} \\ \nabla_{w_{k-2}} f = \underbrace{\sigma'_k\,w_k}\,\underbrace{\sigma'_{k-1}\,w_{k-1}}\,\overbrace{\sigma'_{k-2}\,\sigma_{k-3}} \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \nabla_{w_1} f = \underbrace{\sigma'_k\,w_k}\,\underbrace{\sigma'_{k-1}\,w_{k-1}}\,\underbrace{\sigma'_{k-2}w_{k-2}}\dots\underbrace{\sigma'_2 w_2}\,\overbrace{\sigma'_1\,x}\,. \end{cases} \tag{6}$$

Чтобы написать рекуррентную формулу, положим, по-прежнему опуская в обозначениях аргументы,

$$\Delta_i = \Delta_{i+1}\,w_{i+1}\,\sigma'_i\,, \tag{7}$$

где $i=k,\dots,1$ и $\Delta_{k+1}=w_{k+1}=1$, и пусть $\sigma_0 = x$, тогда

$$\nabla_{w_i} f = \Delta_i\,\sigma_{i-1}\,. \tag{8}$$

4. Градиент $\nabla_W f(X;W)$ нейросетевой функции в общем случае.

В общем случае компоненты $\nabla_{W_i} f(X;W)$ градиента $\nabla_W f(X;W)$ нейросетевой функции могут быть записаны в виде, аналогичном (5). Компоненту $\nabla_{W_i} f(X;W)$ градиента, состоящую из $n_i \times n_{i-1}$ скалярных компонент, мы будем считать матрицей такого же размера, как и W_i . Применение цепного правила в таких обозначениях нельзя назвать совсем простой задачей. Для записи формул нам потребуются три вида поизведений матриц: обычное произведение $A \cdot B$ столбца A на строку B , произведение Адамара $A \circ B$ двух столбцов и «обращенное» обычное произведение $A \bullet B = B \cdot A$ столбца A на матрицу B . Как и в формулах (6), мы опускаем, для краткости, аргументы N_i у отображений $\Sigma_i(N_i)$ и $\Sigma'_i(N_i)$. Напомним, что N_i – это столбец размерности n_i . Итак,

Для нейросетевой функции

$$f(X;W) = \Sigma_k(W_k \cdot \Sigma_{k-1}(W_{k-1} \cdot \Sigma_{k-2} \cdot \dots \cdot \Sigma_2(W_2 \cdot \Sigma_1(W_1 \cdot X))\dots))$$

имеют место формулы

$$
\begin{cases}
\nabla_{W_k} f = \overbrace{\Sigma'_k \cdot \Sigma^T_{k-1}} \\
\nabla_{W_{k-1}} f = \underbrace{\Sigma'_k \bullet W^T_k} \circ \overbrace{\Sigma'_{k-1} \cdot \Sigma^T_{k-2}} \\
\nabla_{W_{k-2}} f = \underbrace{\Sigma'_k \bullet W^T_k} \circ \underbrace{\Sigma'_{k-1} \bullet W^T_{k-1}} \circ \overbrace{\Sigma'_{k-2} \cdot \Sigma^T_{k-3}} \\
\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\
\nabla_{W_1} f = \underbrace{\Sigma'_k \bullet W^T_k} \circ \underbrace{\Sigma'_{k-1} \bullet W^T_{k-1}} \circ \underbrace{\Sigma'_{k-2} \bullet W_{r-2}} \dots \underbrace{\Sigma'_2 \bullet W^T_2} \circ \overbrace{\Sigma'_1 \cdot X^T} .
\end{cases}
\tag{9}
$$

З а м е ч а н и я.

1. Фигурные скобки указывают на периодичность в структуре формул, а не на порядок действий. Порядок действий в (9)– слева направо.

2. Произведение $\Sigma'_k \bullet W^T_k$ можно заменить на $\Sigma'_k \cdot W^T_k$ (поскольку Σ'_k – скаляр) и знак $\bullet$ используется в этом случае только для того, чтобы подчеркнуть периодичность в структуре формул.

3. Чтобы написать рекуррентную формулу, положим, по-прежнему опуская в обозначениях аргументы,

$$\Delta_i = \Delta_{i+1} \bullet W^T_{i+1} \circ \Sigma'_i = (W^T_{i+1} \cdot \Delta_{i+1}) \circ \Sigma'_i\,, \tag{10}$$

где $i = k, \dots, 1$ и $\Delta_{k+1} = W^T_{k+1} = 1$, и пусть $\Sigma_0 = X$, тогда

$$\nabla_{W_i} f = \Delta_i \cdot \Sigma^T_{i-1}\,. \tag{11}$$

4. Δ_i – это n_i -мерный столбец и Σ^T_{i-1} – это n_{i-1} -мерная строка, и потому $\nabla_{W_i} f$ – это $n_i \times n_{i-1}$ -матрица.

5. В рекуррентной версии формул удается заменить «обращенное» произведение матриц $\bullet$ обычным.

5. Доказательство формул (9).

Формулы (9) немедленно следуют из равенств

$$\nabla_{W_r} f = (\nabla_{\Sigma_r} f) \circ \Sigma'_r \cdot \Sigma_{r-1} \tag{12}$$

$$\nabla_{\Sigma_r} f = (\nabla_{\Sigma_{r+1}} f) \circ \Sigma'_{r+1} \bullet\ W^T_{r+1}\,, \tag{13}$$

которые мы и будем проверять. Нам потребуются следующее правило вычисления градиента, согласованное с нашим соглашением о матричной записи производных по матрицам: если A – строка и B – столбец, то

$$\nabla_A (A \cdot B) = B^T\,. \tag{14}$$

Далее, обозначим через W^i_r i -ю строку матрицы $W_r = \{W^i_r\}_i$ и через Σ^i_r – i -тый элемент столбца $\Sigma_r = \{\Sigma^i_r\}_i$. В частности, $\nabla_{W_r} = \{\nabla_{W^i_r}\}_i$ и $\nabla_{\Sigma_r} = \{\nabla_{\Sigma^i_r}\}_i$. Проверим теперь равенство (12):

$$\nabla_{W_r} f = \{\nabla_{W^i_r} f\}_i = \{\nabla_{\Sigma^i_r} f \cdot \nabla_{W^i_r} \Sigma^i_r\}_i = \{\nabla_{\Sigma^i_r} f \cdot \nabla_{W^i_r} (\Sigma^i_r\,(W^i_r \cdot \Sigma_{r-1}))\}_i =$$

$$= \{\nabla_{\Sigma^i_r} f \cdot \Sigma^{i\,\prime}_r \cdot \Sigma^T_{r-1}\}_i = (\nabla_{\Sigma_r} f) \circ \Sigma'_r \cdot \Sigma^T_{r-1}\,.$$

Равенство (13) запишем покоординатно

$$\{\nabla_{\Sigma^i_r} f\}_i = \{(\nabla_{\Sigma_{r+1}} f) \circ \Sigma'_{r+1} \bullet\ (W^T_{r+1})^i\}_i$$

и проверим равенство соответствующих координат:

$$\nabla_{\Sigma_r^i} f = \langle \nabla_{\Sigma_{r+1}} f \,, \nabla_{\Sigma_r^i} \Sigma_{r+1} \rangle = \langle \nabla_{\Sigma_{r+1}} f \,, \nabla_{\Sigma_r^i} (\Sigma_{r+1}(W_{r+1}\Sigma_r)) \rangle =$$

$$= \langle \nabla_{\Sigma_{r+1}} f \,, \Sigma'_{r+1} \circ (W_{r+1}^T)^i \rangle = \langle \nabla_{\Sigma_{r+1}} f \circ \Sigma'_{r+1} \,, (W_{r+1}^T)^i \rangle = (\nabla_{\Sigma_{r+1}} f) \circ \Sigma'_{r+1} \bullet (W_{r+1}^T)^i$$

(здесь $\langle \; , \; \rangle$ – скалярное произведение).

СПИСОК ЛИТЕРАТУРЫ

1. *Хайкин С.* С. Нейронные сети: полный курс. М.: Вильямс, 2006.
2. *Rumelhart D.E., Hinton G.E., Williams R.J.* Learning Internal Representations by Error Propagation // Parallel Distributed Processing, Cambridge: MA, MIT Press. 1986. V. 1. P. 318–362.

БЛАГОДАРНОСТИ: Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 16-07-00854_а).

Поступила в редакцию 2 марта 2017 г

Мишачёв Николай Михайлович, Липецкий государственный технический университет, г. Липецк, Российская Федерация, кандидат физико-математических наук, доцент кафедры высшей математики, e-mail: nmish@lipetsk.ru

Шмырин Анатолий Михайлович, Липецкий государственный технический университет, г. Липецк, Российская Федерация, доктор технических наук, профессор, зав. кафедрой высшей математики, e-mail: amsh@lipetsk.ru

UDC 519.85
DOI: 10.20310/1810-0198-2017-22-3-552-557

ON THE GRADIENT OF NEURONETWORK FUNCTION

© N. M. Mishachev, A. M. Shmyrin

Lipetsk State Technical University
30 Moskovskaya St., Lipetsk, Russian Federation, 398600
E-mail: nmish@lipetsk.ru, amsh@lipetsk.ru

The paper proposes a matrix formula for the gradient of neuronetwork function $\nabla_W f(X;W)$ with respect to the parameter vector W .

Key words: neuronetwork function; neural network; backpropagation algorithm; Hadamard product

REFERENCES

1. *Haykin* C. Neural Networks: A Comprehensive Foundation. M.: Viliams, 2006.
2. *Rumelhart D.E., Hinton G.E., Williams R.J.* Learning Internal Representations by Error Propagation // Parallel Distributed Processing, Cambridge: MA, MIT Press. 1986. V. 1. P. 318–362.

ACKNOWLEDGEMENTS: The present research is supported by the Russian Fund for Basic Research (project № 16-07-00854_a).

Received 2 March 2017

Mishachev Nikolay Mikhailovich, Lipetsk State Technical University, Lipetsk, the Russian Federation, Candidate of Physics and Mathematics, Associate Professor of the Higher Mathematics Department, e-mail:nmish@lipetsk.ru

Shmyrin Anatoliy Mikhailovich, Lipetsk State Technical University, Lipetsk, the Russian Federation, Doctor of Techniques, Professor, the Head of the Higher Mathematics Department, e-mail: amsh@lipetsk.ru

Информация для цитирования:

Мишачев Н.М., Шмырин А.М. О градиенте нейросетевой функции // Вестник Тамбовского университета. Серия Естественные и технические науки. Тамбов, 2017. Т. 22. Вып. 3. С. 552–557. DOI: 10.20310/1810-0198-2017-22-3-552-557

Mishachev N.M., Shmyrin A.M. O gradiente neyrosetevoy funktsii [On the gradient of neuronetwork function]. *Vestnik Tambovskogo universiteta. Seriya Estestvennye i tekhnicheskie nauki – Tambov University Reports. Series: Natural and Technical Sciences*, 2017, vol. 22, no. 3, pp. 552–557. DOI: 10.20310/1810-0198-2017-22-3-552-557 (In Russian)