The background of the slide features a close-up, slightly blurred photograph of a tree branch. The branch is light-colored, possibly birch, and has a textured bark. A small, cup-shaped bird's nest is visible, nestled in the fork of the branch. The lighting is soft and natural, suggesting an outdoor setting.

Машинное обучение

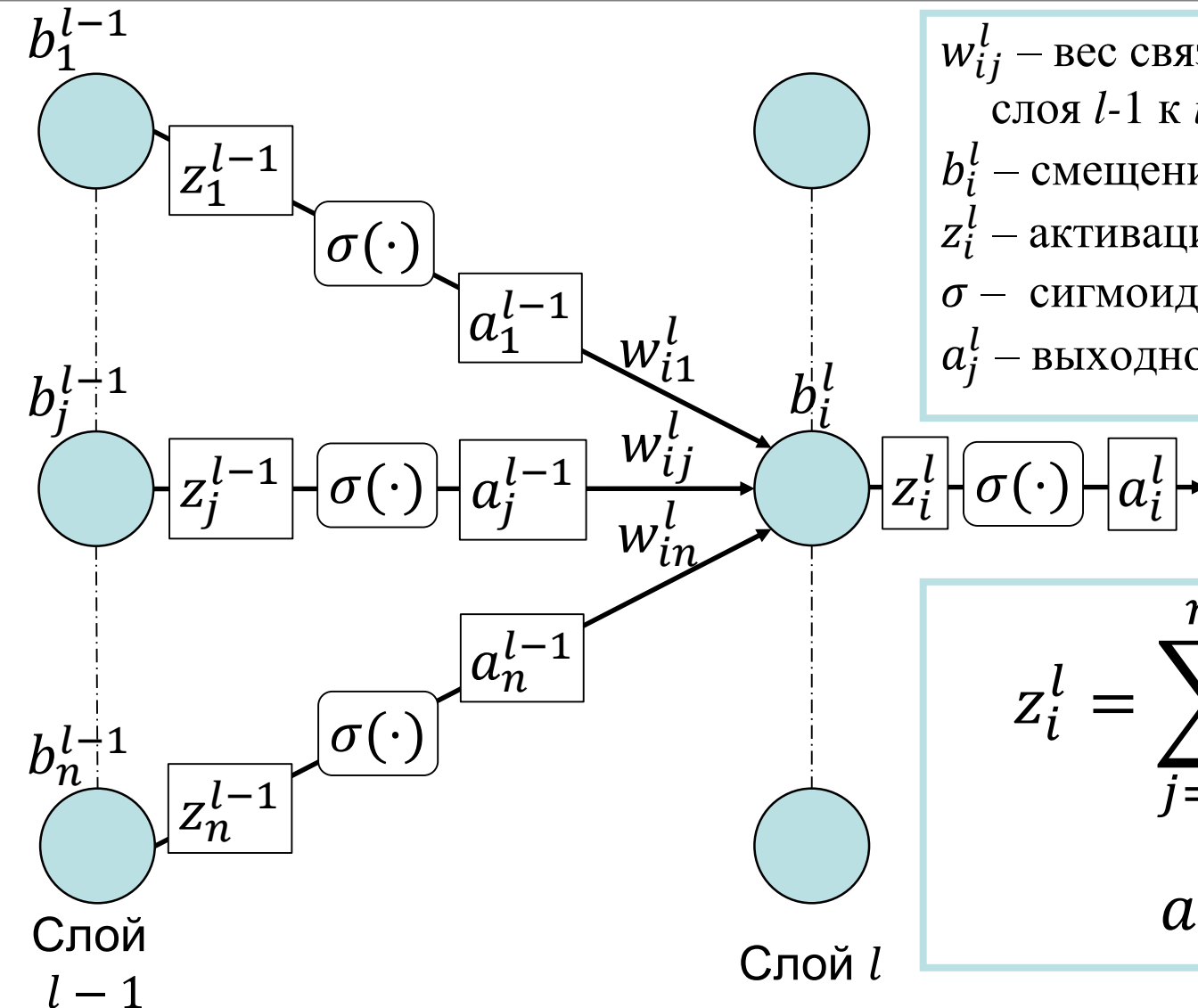
Метод обратного

распространения ошибки

(Error Back Propagation)

Лекция 3

Обозначения



w_{ij}^l – вес связи от j -того нейрона
 слоя $l-1$ к i -тому нейрону слоя l
 b_i^l – смещение
 z_i^l – активационный потенциал
 σ – сигмоидная функция активации
 a_j^l – выходной сигнал

$$z_i^l = \sum_{j=1}^n w_{ij}^l a_j^{l-1} + b_i^l$$

$$a_i^l = \sigma(z_i^l)$$

Обозначения

Везде далее

L – количество уровней нейронной сети

$$\vec{a}^L = \vec{a}(\vec{x})$$

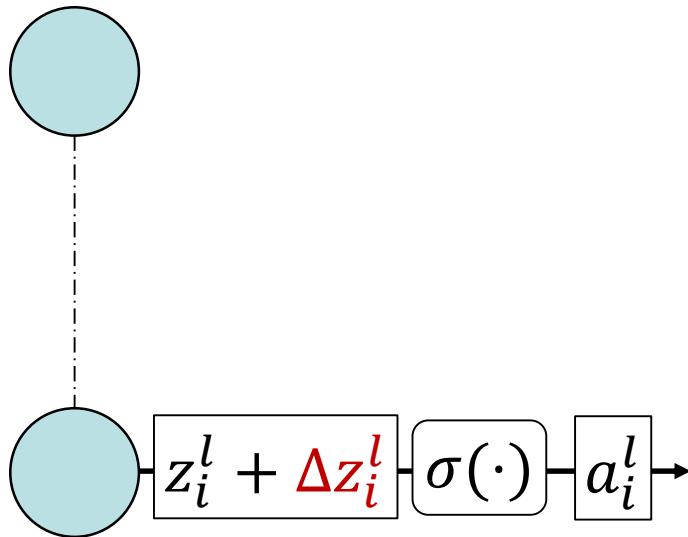
$$C = \mathbb{C}_{(\vec{x}, \vec{y})} = \frac{\|\vec{a}^L - \vec{y}\|^2}{2}$$

Мера влияния нейрона на величину ошибки

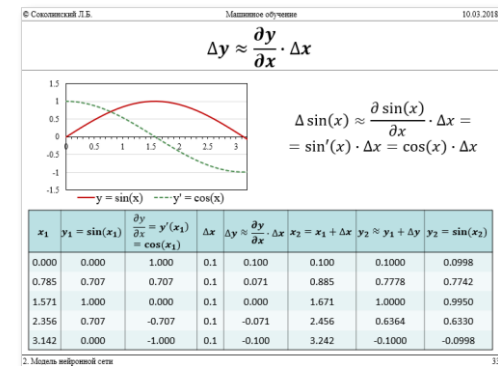
$$C = \frac{\|\vec{a}^L - \vec{y}\|^2}{2}$$

$$\Delta C \approx \frac{\partial C}{\partial z_i^l} \Delta z_i^l$$

$$\delta_i^l = \frac{\partial C}{\partial z_i^l}$$

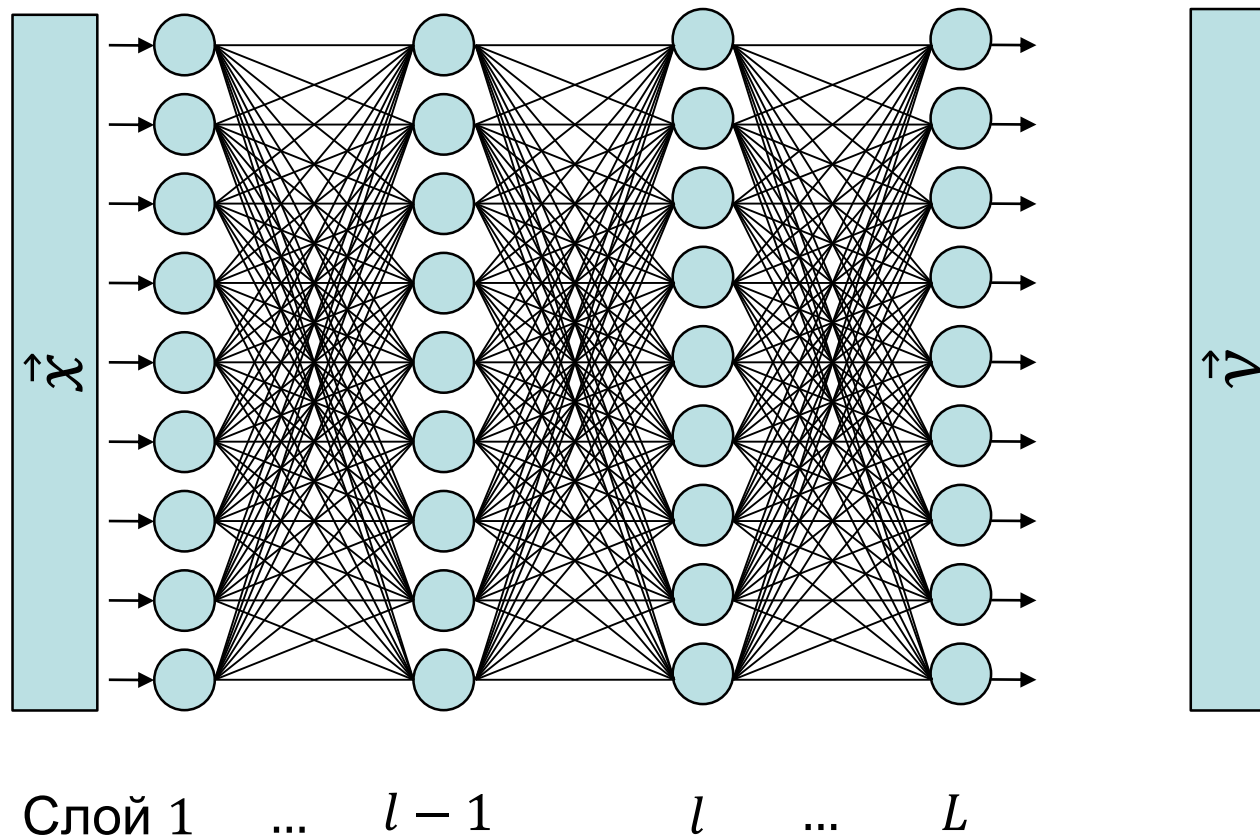


Слой l



Идея метода обратного распространения ошибки

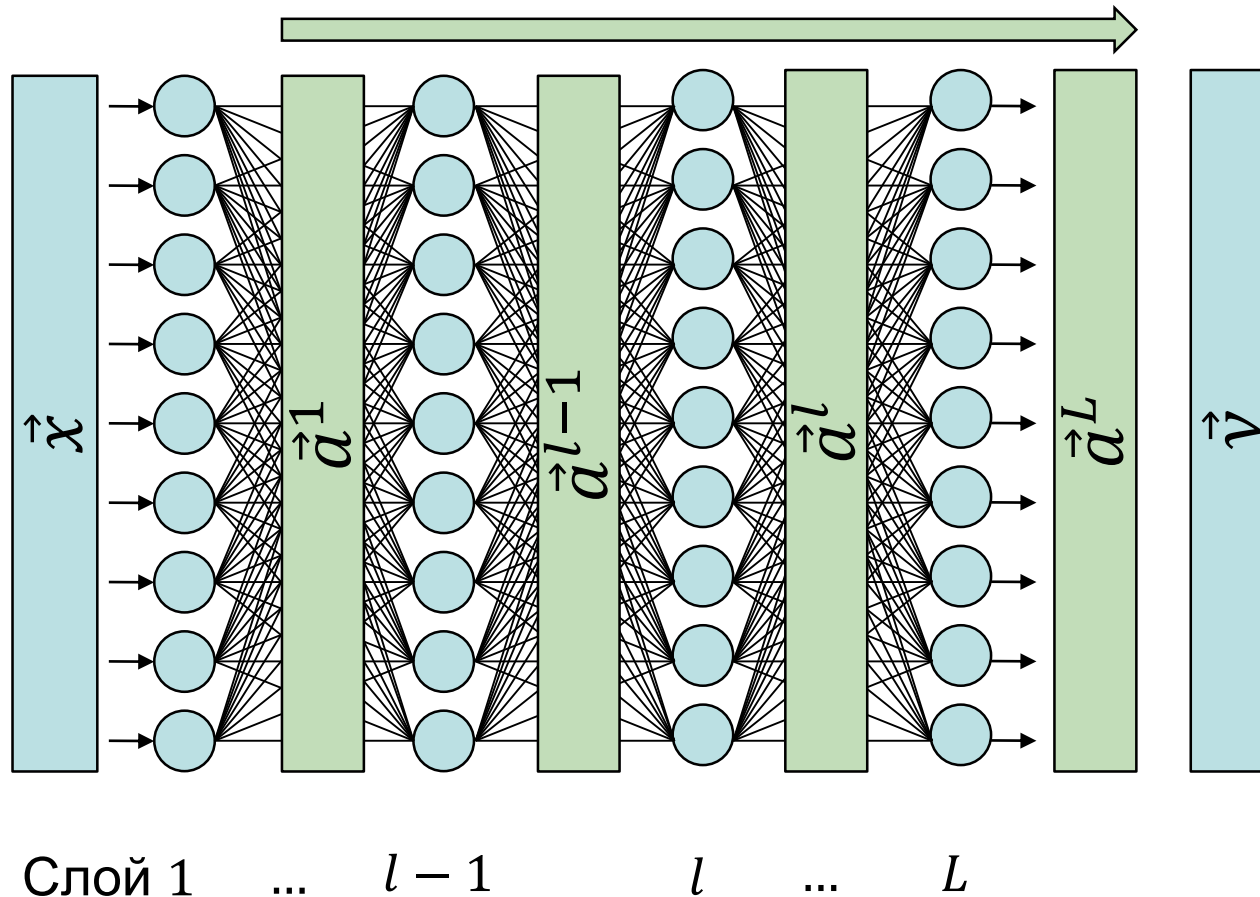
Шаг 1. Взять образец (\vec{x}, \vec{y}) и подать сигнал \vec{x} на вход нейронной сети



Идея метода обратного распространения ошибки

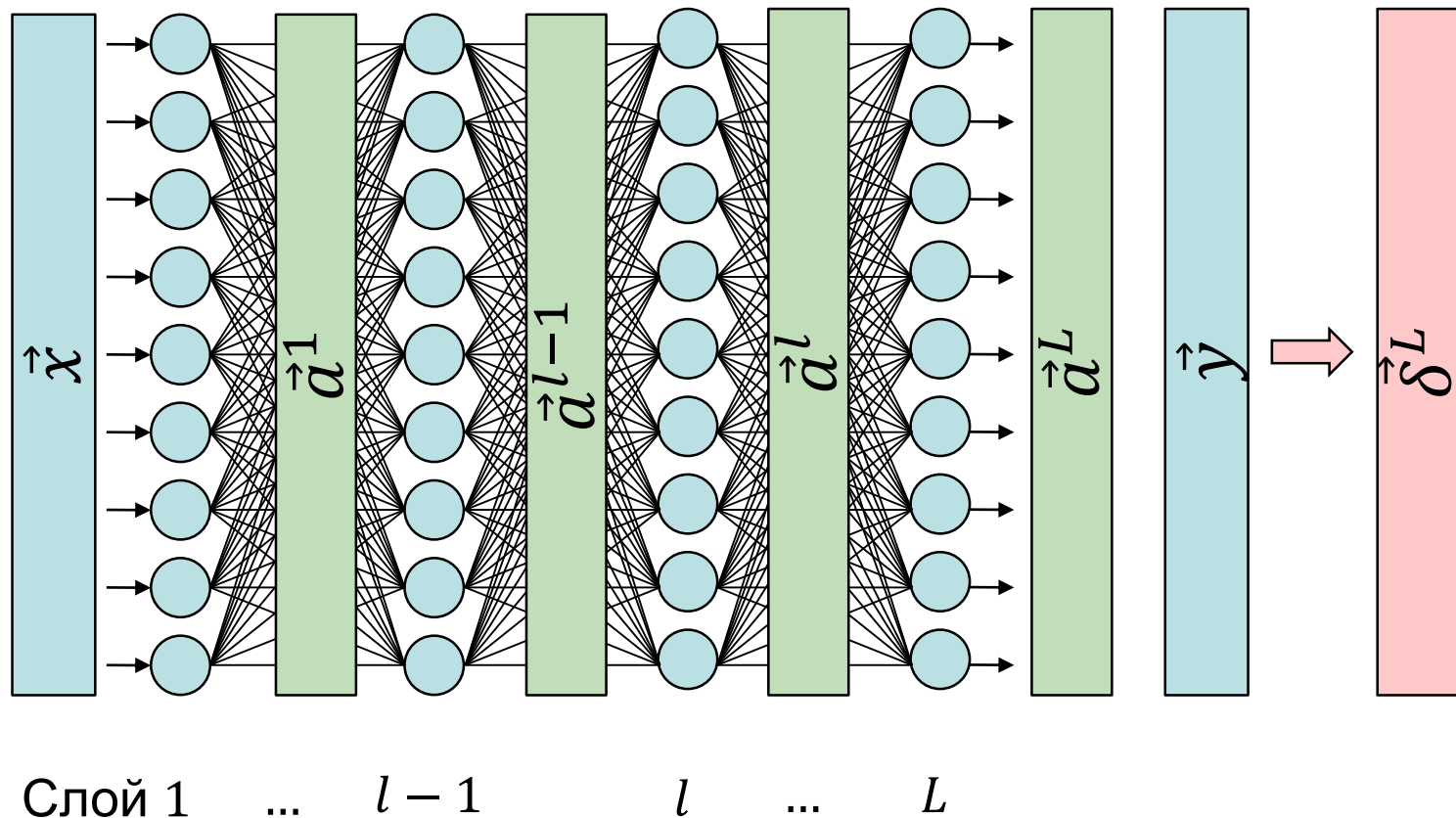
Шаг 2. Последовательно вычислить выходные сигналы \vec{a}^l для каждого слоя:

$$\vec{z}^l = W^l \cdot \vec{a}^{l-1T} + \vec{b}^l$$



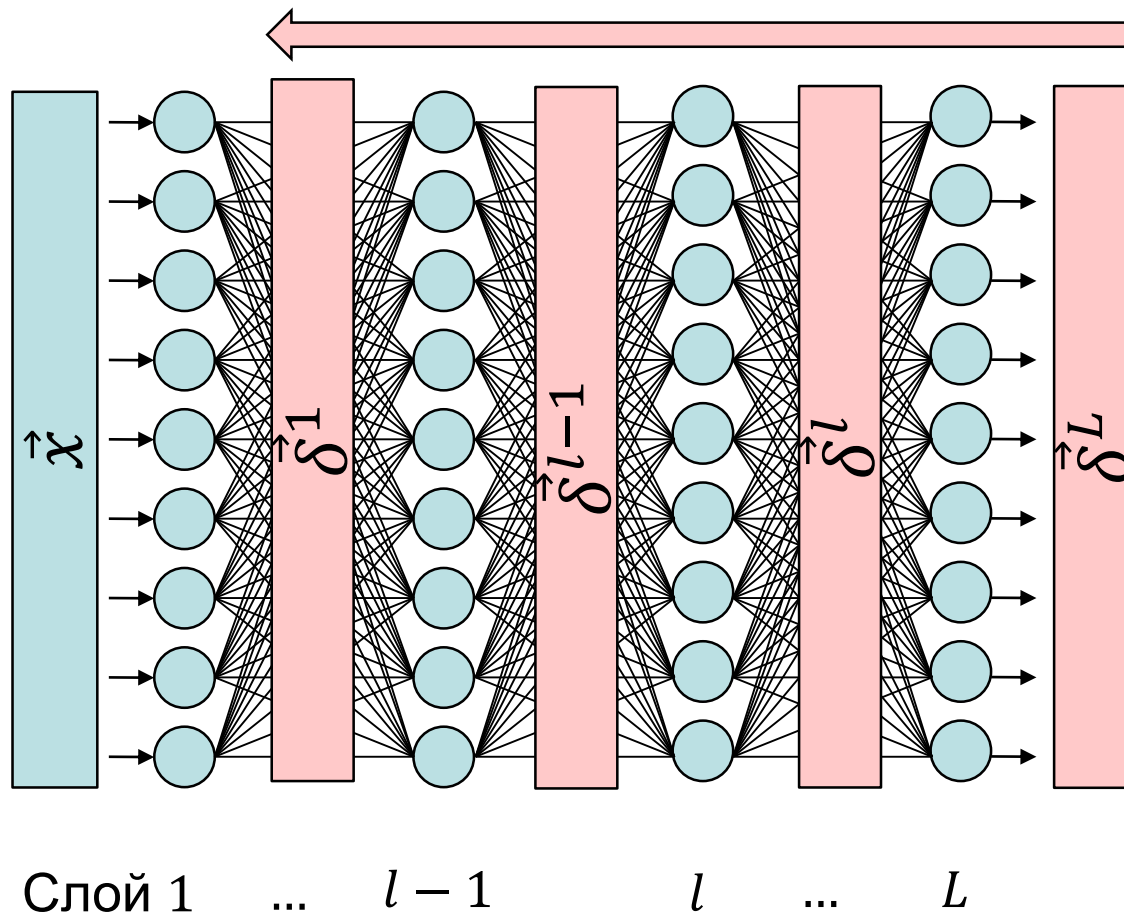
Идея метода обратного распространения ошибки

Шаг 3. Вычислить меру влияния δ^L нейронов выходного слоя L на ошибку C



Идея метода обратного распространения ошибки

Шаг 4. Вычислить в обратном порядке меру влияния δ^l на ошибку для каждого слоя



Идея метода обратного распространения ошибки

Шаг 5. Используя $\vec{\delta}^l$, вычислить $\nabla C(W^l)$ и $\nabla C(\vec{b}^l)$ для каждого слоя l

Шаг 6. Выполнив шаги 1-5 для всей подвыборки V_i , вычислить

$$\nabla C_{V_i}(\vec{w}) := \frac{1}{|V_i|} \sum_{(x,y) \in V_i} \nabla C_{(x,y)}(\vec{w})$$

$$\nabla C_{V_i}(\vec{b}) := \frac{1}{|V_i|} \sum_{(x,y) \in V_i} \nabla C_{(x,y)}(\vec{b})$$

© Соколинский Л.Б. Машинное обучение 10.03.2018

Стохастический градиентный спуск

1. $\vec{w} := \overline{rnd}$; $\vec{b} := \overline{rnd}$; $epoch := 1$;
2. Последовательное разбиение V на непересекающиеся подвыборки V_1, \dots, V_M
3. $i := 0$;
4. $i := i + 1$;
5. $\nabla C_{V_i}(\vec{w}) := \frac{1}{|V_i|} \sum_{(x,y) \in V_i} \nabla C_{(x,y)}(\vec{w})$;
6. $\nabla C_{V_i}(\vec{b}) := \frac{1}{|V_i|} \sum_{(x,y) \in V_i} \nabla C_{(x,y)}(\vec{b})$;
7. $\vec{w} := \vec{w} - \eta \nabla C_{V_i}(\vec{w})$;
8. $\vec{b} := \vec{b} - \eta \nabla C_{V_i}(\vec{b})$;
9. **if** $i < M$ **goto** 4; -----
10. $shuffle(V)$; $epoch := epoch + 1$;
11. **if** $epoch \leq 10$ **goto** 2; -----

Цикл по подвыборкам
Цикл по эпохам обучения

2. Модель нейронной сети 34

Мера влияния нейронов выходного слоя L на величину ошибки

$$\vec{\delta}^L = (\vec{a}^L - \vec{y}) \circ \vec{\sigma}'(\vec{z}^L) \quad \text{BP1'}$$

© Соколинский Л.Б. Машинное обучение 10.03.2018

Производная вектор-функции

$$\vec{\sigma}'(\vec{z}) = [\sigma'(z_1), \dots, \sigma'(z_n)]$$

3. Метод обратного распространения ошибки 30

© Соколинский Л.Б. Машинное обучение 10.03.2018

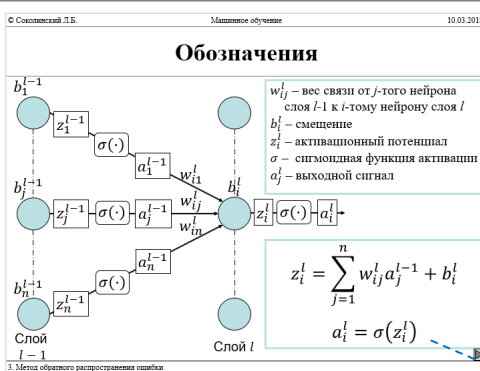
**Произведение Адамара
(покомпонентное умножение)**

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \circ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 * y_1 \\ \vdots \\ x_n * y_n \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \circ \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 * 4 \\ 2 * 5 \\ 3 * 6 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \\ 18 \end{bmatrix}$$

3. Метод обратного распространения ошибки 31

Доказательство BP1'



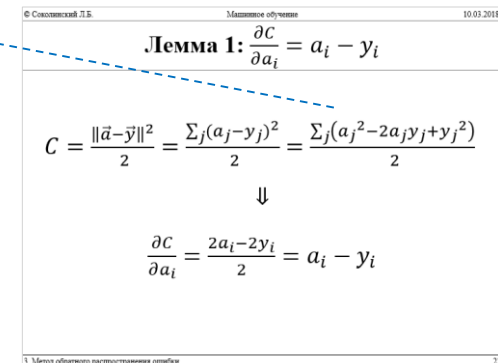
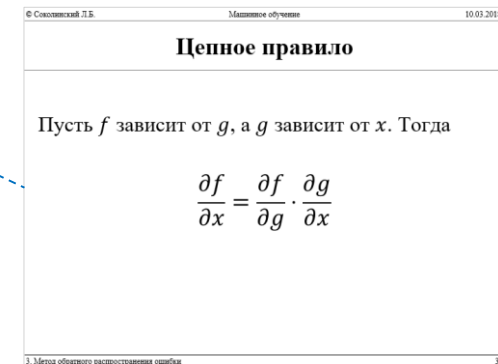
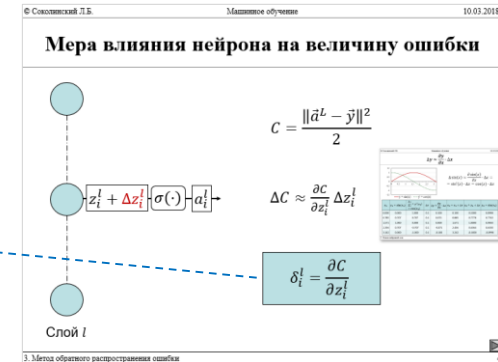
$$\delta_i^L = \frac{\partial C}{\partial z_i^L} =$$

$$= \frac{\partial C}{\partial a_i^L} \cdot \frac{\partial a_i^L}{\partial z_i^L} =$$

$$= (a_i^L - y_i) \cdot \sigma'(z_i^L)$$

То же самое в векторной форме

$$\vec{\delta}^L = (\vec{a}^L - \vec{y}) \circ \vec{\sigma}'(\vec{z}^L)$$



Обобщение BP1'

$$\vec{\delta}^L = \nabla C(\vec{a}) \circ \vec{\sigma}'(\vec{z}^L)$$

BP1

© Соколинский Л.Б. Машинное обучение 10.03.2018

Градиент

$$\nabla C(\vec{a}) = \left[\frac{\partial C}{\partial a_1}, \dots, \frac{\partial C}{\partial a_n} \right]$$

3. Метод обратного распространения ошибки 29

© Соколинский Л.Б. Машинное обучение 10.03.2018

Мера влияния нейронов выходного слоя L на величину ошибки

$$\vec{\delta}^L = (\vec{a}^L - \vec{y}) \circ \vec{\sigma}'(\vec{z}^L) \quad \text{BP1'}$$

Применим вектор-функцию

$$\sigma'(z) = [\sigma'(z_1), \dots, \sigma'(z_n)]$$

Применим Аппроксимацию Липшица

$$\left[\begin{array}{c} \sigma'(z_1) \\ \sigma'(z_2) \\ \vdots \\ \sigma'(z_n) \end{array} \right] \approx \left[\begin{array}{c} \sigma'(z_1) \\ \sigma'(z_2) \\ \vdots \\ \sigma'(z_n) \end{array} \right] \cdot \left[\begin{array}{c} \sigma'(z_1) \\ \sigma'(z_2) \\ \vdots \\ \sigma'(z_n) \end{array} \right]$$

3. Метод обратного распространения ошибки 30

© Соколинский Л.Б. Машинное обучение 10.03.2018

Лемма 1: $\frac{\partial C}{\partial a_i} = a_i - y_i$

$$C = \frac{\|\vec{a} - \vec{y}\|^2}{2} = \frac{\sum_j (a_j - y_j)^2}{2} = \frac{\sum_j (a_j^2 - 2a_j y_j + y_j^2)}{2}$$

↓

$$\frac{\partial C}{\partial a_i} = \frac{2a_i - 2y_i}{2} = a_i - y_i$$

3. Метод обратного распространения ошибки 31

Мера влияния нейронов слоя l на величину ошибки

$$\vec{\delta}^l = \left((W^{l+1})^T \vec{\delta}^{l+1} \right) \circ \vec{\sigma}'(\vec{z}^l) \quad \text{BP2}$$

© Соколинский Л.Б. Машинное обучение 10.03.2018

Производная вектор-функции

$$\vec{\sigma}'(\vec{z}) = [\sigma'(z_1), \dots, \sigma'(z_n)]$$

3. Метод обратного распространения ошибки 30

© Соколинский Л.Б. Машинное обучение 10.03.2018

**Произведение Адамара
(покомпонентное умножение)**

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \circ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 * y_1 \\ \vdots \\ x_n * y_n \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \circ \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 * 4 \\ 2 * 5 \\ 3 * 6 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \\ 18 \end{bmatrix}$$

3. Метод обратного распространения ошибки 31

Доказательство BP2

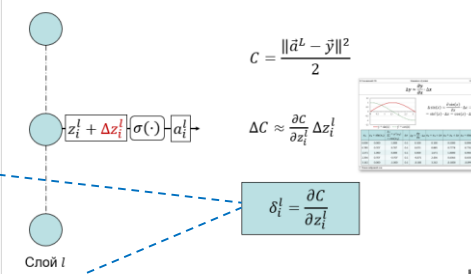
$$\begin{aligned}
 \delta_j^l &= \frac{\partial C}{\partial z_j^l} = \\
 &= \sum_i \frac{\partial C}{\partial z_i^{l+1}} \cdot \frac{\partial z_i^{l+1}}{\partial z_j^l} = \\
 &= \sum_i \frac{\partial z_i^{l+1}}{\partial z_j^l} \cdot \frac{\partial C}{\partial z_i^{l+1}} = \sum_i w_{ij}^{l+1} \sigma'(z_j^l) \cdot \delta_i^{l+1} = \\
 &= \sum_i w_{ij}^{l+1} \delta_i^{l+1} \sigma'(z_j^l)
 \end{aligned}$$

То же самое в векторной форме

$$\vec{\delta}^l = \left((W^{l+1})^T \vec{\delta}^{l+1} \right) \circ \vec{\sigma}'(\vec{z}^l)$$

© Соколинский Л.Б. Машинное обучение 10.03.2018

Мера влияния нейрона на величину ошибки



$$C = \frac{\|\vec{a}^l - \vec{y}\|^2}{2}$$

$$\Delta C \approx \frac{\partial C}{\partial z_i^l} \Delta z_i^l$$

$$\delta_i^l = \frac{\partial C}{\partial z_i^l}$$

Слой l

3. Метод обратного распространения ошибки

© Соколинский Л.Б. Машинное обучение 10.03.2018

Ценное правило для многих переменных

Пусть f зависит от $\sum g_i$, и все g_i зависят от x . Тогда

$$\frac{\partial f}{\partial x} = \sum_i \frac{\partial f}{\partial g_i} \cdot \frac{\partial g_i}{\partial x}$$

3. Метод обратного распространения ошибки 32

© Соколинский Л.Б. Машинное обучение 10.03.2018

Лемма 2: $\frac{\partial z_i^{l+1}}{\partial z_j^l} = w_{ij}^{l+1} \sigma'(z_j^l)$

$$z_i^{l+1} = \sum_j w_{ij}^{l+1} a_j^l + b_i^{l+1} = \sum_j w_{ij}^{l+1} \sigma(z_j^l) + b_i^{l+1}$$

$$\Downarrow$$

$$\frac{\partial z_i^{l+1}}{\partial z_j^l} = w_{ij}^{l+1} \sigma'(z_j^l)$$

3. Метод обратного распространения ошибки 34

Формула для вычисления градиента по смещению

$$\nabla C(\vec{b}^l) = \vec{\delta}^l$$

ВРЗ

Доказательство ВРЗ

$$\frac{\partial C}{\partial b_i^l} = \frac{\partial C}{\partial z_i^l} \cdot \frac{\partial z_i^l}{\partial b_i^l} =$$

$$= \delta_i^l \cdot \frac{\partial z_i^l}{\partial b_i^l} =$$

$$= \delta_i^l$$

То же самое в векторной форме

$$\nabla C(\vec{b}) = \vec{\delta}^l$$

© Соколинский Л.Б. Машинное обучение 10.03.2018

Цепное правило

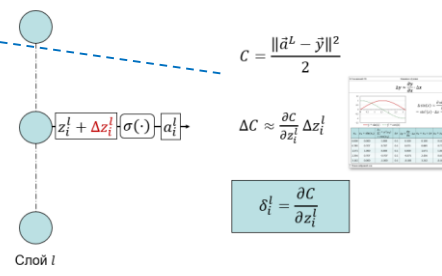
Пусть f зависит от g , а g зависит от x . Тогда

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x}$$

3. Метод обратного распространения ошибки 31

© Соколинский Л.Б. Машинное обучение 10.03.2018

Мера влияния нейрона на величину ошибки



$$C = \frac{\|a^l - y\|^2}{2}$$

$$\Delta C \approx \frac{\partial C}{\partial z_i^l} \Delta z_i^l$$

$$\delta_i^l = \frac{\partial C}{\partial z_i^l}$$

Слой l

3. Метод обратного распространения ошибки 32

© Соколинский Л.Б. Машинное обучение 10.03.2018

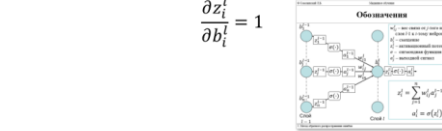
Лемма 3: $\frac{\partial z_i^l}{\partial b_i^l} = 1$

$$z_i^l = \sum_j w_{ij}^l a_j^{l-1} + b_i^l$$

$$\Downarrow$$

$$\frac{\partial z_i^l}{\partial b_i^l} = 1$$

Обозначения



3. Метод обратного распространения ошибки 33

Формула для вычисления градиента по весам

$$\nabla C(W^l) = \delta^{\vec{l}T} \circ \begin{pmatrix} \vec{a}^{l-1} \\ \vdots \\ \vec{a}^{l-1} \end{pmatrix} \quad \text{BP4}$$

Доказательство ВР4

$$\begin{aligned} \frac{\partial C}{\partial w_{ij}^l} &= \frac{\partial C}{\partial z_i^l} \cdot \frac{\partial z_i^l}{\partial w_{ij}^l} = \\ &= \delta_i^l \cdot \frac{\partial z_i^l}{\partial w_{ij}^l} = \\ &= \delta_i^l \cdot a_j^{l-1} \end{aligned}$$

То же самое в векторной форме

$$\nabla C(W^l) = \vec{\delta}^{lT} \circ \begin{pmatrix} \vec{a}^{l-1} \\ \vdots \\ \vec{a}^{l-1} \end{pmatrix}$$

© Соколинский Л.Б. Машинное обучение 10.03.2018

Цепное правило

Пусть f зависит от g , а g зависит от x . Тогда

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x}$$

3. Метод обратного распространения ошибки 31

© Соколинский Л.Б. Машинное обучение 10.03.2018

Мера влияния нейрона на величину ошибки

$C = \frac{\|\vec{a}^l - \vec{y}\|^2}{2}$

$\Delta C \approx \frac{\partial C}{\partial z_i^l} \Delta z_i^l$

$\delta_i^l = \frac{\partial C}{\partial z_i^l}$

Слой l

3. Метод обратного распространения ошибки 31

© Соколинский Л.Б. Машинное обучение 10.03.2018

Лемма 4: $\frac{\partial z_i^l}{\partial w_{ij}^l} = a_j^{l-1}$

$$z_i^l = \sum_j w_{ij}^l a_j^{l-1} + b_i^l$$

$$\Downarrow$$

$$\frac{\partial z_i^l}{\partial w_{ij}^l} = a_j^{l-1}$$

Обозначения

$z_i^l = \sum_j w_{ij}^l a_j^{l-1} + b_i^l$

$a_i^l = \sigma(z_i^l)$

3. Метод обратного распространения ошибки 36

Формулы обратного распространения ошибки

$$\vec{\delta}^L = \nabla C(\vec{a}) \circ \vec{\sigma}'(\vec{z}^L) \quad (\text{BP1})$$

$$\vec{\delta}^l = \left((W^{l+1})^T \vec{\delta}^{l+1} \right) \circ \vec{\sigma}'(\vec{z}^l) \quad (\text{BP2})$$

$$\nabla C(\vec{b}^l) = \vec{\delta}^l \quad (\text{BP3})$$

$$\nabla C(W^l) = \vec{\delta}^{lT} \circ \begin{pmatrix} \vec{a}^{l-1} \\ \vdots \\ \vec{a}^{l-1} \end{pmatrix} \quad (\text{BP4})$$

Алгоритм обратного распространения ошибки

1. **Вход x :** Установить соответствующие значения активации a^1 для входного уровня
2. **Прямое распространение:** Для $l = 2, \dots, L$ последовательно вычислить $\vec{z}^l = W^l(\vec{a}^{l-1}) + \vec{b}^l$ и $\vec{a}^l = \vec{\sigma}(\vec{z}^l)$
3. **Вычислить:** $\vec{\delta}^L = \nabla C(\vec{a}) \circ \vec{\sigma}'(\vec{z}^L)$
4. **Обратное распространение:** Для $l = L - 1, \dots, 2$ последовательно вычислить $\vec{\delta}^l = \left((W^{l+1})^T \vec{\delta}^{l+1} \right) \circ \vec{\sigma}'(\vec{z}^l)$
5. **Выход:** Для $l = 2, \dots, L$ вычислить

$$\nabla C_{(x,y)}(W^l) = \vec{\delta}^{lT} \circ \begin{pmatrix} \vec{a}^{l-1} \\ \vdots \\ \vec{a}^{l-1} \end{pmatrix}$$

$$\nabla C_{(x,y)}(\vec{b}^l) = \vec{\delta}^l$$

Конец лекции 3

Вспомогательные слайды

Лемма 1: $\frac{\partial C}{\partial a_i} = a_i - y_i$

$$C = \frac{\|\vec{a} - \vec{y}\|^2}{2} = \frac{\sum_j (a_j - y_j)^2}{2} = \frac{\sum_j (a_j^2 - 2a_j y_j + y_j^2)}{2}$$

⇓

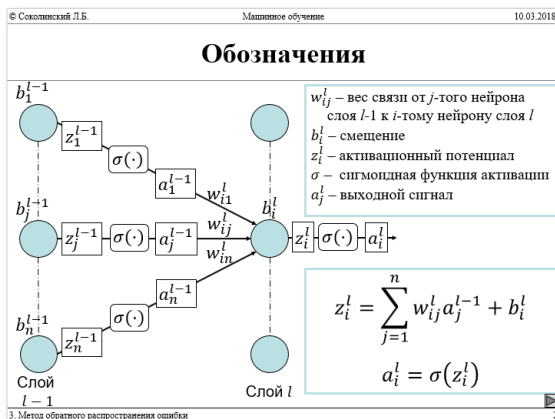
$$\frac{\partial C}{\partial a_i} = \frac{2a_i - 2y_i}{2} = a_i - y_i$$

$$\text{Лемма 2: } \frac{\partial z_i^{l+1}}{\partial z_j^l} = w_{ij}^{l+1} \sigma'(z_j^l)$$

$$z_i^{l+1} = \sum_j w_{ij}^{l+1} a_j^l + b_i^{l+1} = \sum_j w_{ij}^{l+1} \sigma(z_j^l) + b_i^{l+1}$$

⇓

$$\frac{\partial z_i^{l+1}}{\partial z_j^l} = w_{ij}^{l+1} \sigma'(z_j^l)$$



Транспонирование

$$[x_1 \dots x_n]^T = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$\begin{bmatrix} w_{11} & \dots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \dots & w_{mn} \end{bmatrix}^T = \begin{bmatrix} w_{11} & \dots & w_{m1} \\ \vdots & \ddots & \vdots \\ w_{1n} & \dots & w_{mn} \end{bmatrix}$$

Произведение Адамара (покомпонентное умножение)

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \circ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 * y_1 \\ \vdots \\ x_n * y_n \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \circ \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 * 4 \\ 2 * 5 \\ 3 * 6 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \\ 18 \end{bmatrix}$$

Градиент

$$\nabla C(\vec{a}) = \left[\frac{\partial C}{\partial a_1}, \dots, \frac{\partial C}{\partial a_n} \right]$$

Производная вектор-функции

$$\vec{\sigma}'(\vec{z}) = [\sigma'(z_1), \dots, \sigma'(z_n)]$$

Цепное правило

Пусть f зависит от g , а g зависит от x . Тогда

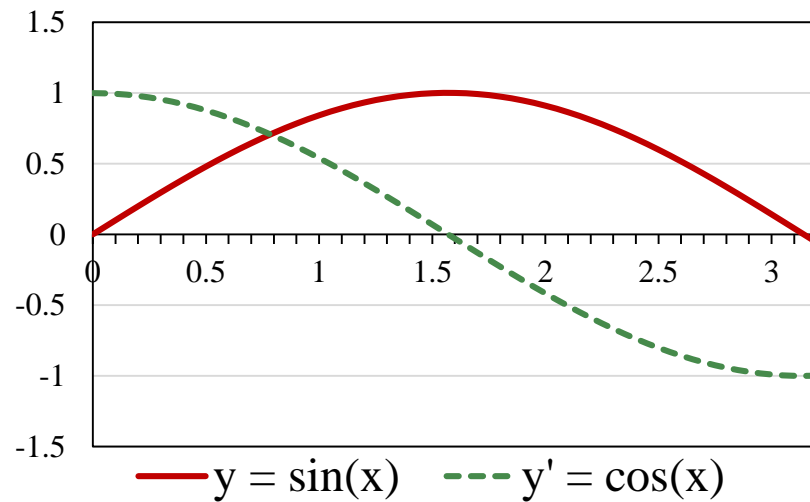
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x}$$

Цепное правило для многих переменных

Пусть f зависит от $\sum g_i$, и все g_i зависят от x .
Тогда

$$\frac{\partial f}{\partial x} = \sum_i \frac{\partial f}{\partial g_i} \cdot \frac{\partial g_i}{\partial x}$$

$$\Delta y \approx \frac{\partial y}{\partial x} \cdot \Delta x$$



$$\begin{aligned} \Delta \sin(x) &\approx \frac{\partial \sin(x)}{\partial x} \cdot \Delta x = \\ &= \sin'(x) \cdot \Delta x = \cos(x) \cdot \Delta x \end{aligned}$$

x_1	$y_1 = \sin(x_1)$	$\frac{\partial y}{\partial x} = y'(x_1) = \cos(x_1)$	Δx	$\Delta y \approx \frac{\partial y}{\partial x} \cdot \Delta x$	$x_2 = x_1 + \Delta x$	$y_2 \approx y_1 + \Delta y$	$y_2 = \sin(x_2)$
0.000	0.000	1.000	0.1	0.100	0.100	0.1000	0.0998
0.785	0.707	0.707	0.1	0.071	0.885	0.7778	0.7742
1.571	1.000	0.000	0.1	0.000	1.671	1.0000	0.9950
2.356	0.707	-0.707	0.1	-0.071	2.456	0.6364	0.6330
3.142	0.000	-1.000	0.1	-0.100	3.242	-0.1000	-0.0998

Стохастический градиентный спуск

1. $\vec{w} := \overrightarrow{rnd}$; $\vec{b} := \overrightarrow{rnd}$; $epoch := 1$;
2. Последовательное разбиение V на непересекающиеся подвыборки V_1, \dots, V_M
3. $i := 0$;
4. $i := i + 1$;
5. $\nabla \mathbb{C}_{V_i}(\vec{w}) := \frac{1}{|V_i|} \sum_{(x,y) \in V_i} \nabla \mathbb{C}_{(x,y)}(\vec{w})$;
6. $\nabla \mathbb{C}_{V_i}(\vec{b}) := \frac{1}{|V_i|} \sum_{(x,y) \in V_i} \nabla \mathbb{C}_{(x,y)}(\vec{b})$;
7. $\vec{w} := \vec{w} - \eta \nabla \mathbb{C}_{V_i}(\vec{w})$;
8. $\vec{b} := \vec{b} - \eta \nabla \mathbb{C}_{V_i}(\vec{b})$;
9. **if** $i < M$ **goto** 4;
10. $shuffle(V)$; $epoch := epoch + 1$;
11. **if** $epoch \leq 10$ **goto** 2;

Цикл по подвыборкам

Цикл по эпохам обучения