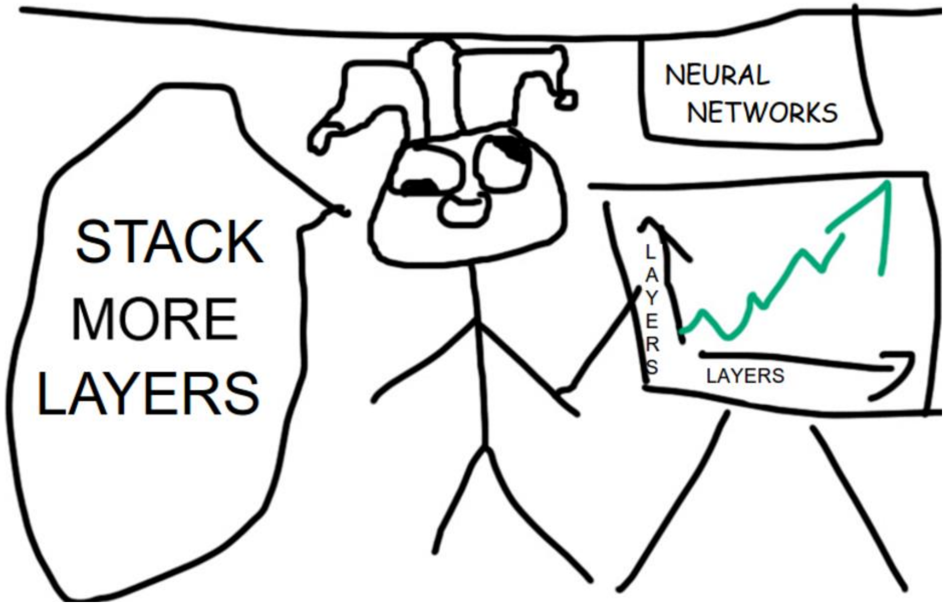

WE NEED TO GO
DEEPER


MACHINE LEARNING
MACHINE LEARNING EVERYWHERE!
memegenerator.net

ХАКАТОНЫ


NEURAL
NETWORKS
STACK
MORE
LAYERS
LAYERS
LAYERS


KAGGLE COMPETITION
LIGHTGBM
imgflip.com

Sendy Logistics Challenge

- https://zindi.africa/competitions/sendy-logistics-challenge
- Дата проведения: до 26 ноября 2019.
- Цель хакатона – предсказать время прибытия мотоцикла в пункт назначения
- Данные: Описание заказа ,дата и время приема, подтверждения и отгрузки заказа, данные пути, данные перевозчика
- Дополнительные данные: Africa GeoPortal
- Советы по старту : 1.Мержим нужные колонки из датасетов в один. 2.Работаем над фичами (посчитать статистику, корреляции и т.д.) 3. Logreg, Gradient boosting – отбор фичей. 4. Lstm для фичей основанных на временных данных. 5. Работа с данными Africa GeoPortal.
- Почитать:https://towardsdatascience.com/artificial-intelligence-in-supply-chain-management-predictive-analytics-for-demand-forecasting-80d2d512f155
- https://github.com/siddharth185/Predicting-Delivery-Time
- https://blog.postmates.com/estimating-delivery-times-a-case-study-in-practical-machine-learning-e70f677e736c

Understanding Clouds from Satellite Images

Немного математики

Дифференциальные вычисления

Дифференциальное исчисление сводится к нахождению изменения одной величины в результате изменения другой. В данном случае нас интересует, **как скорость изменяется со временем.**

Вышеизложенное можно записать в следующей математической форме:

$$\frac{\delta s}{\delta t} = 0$$

Функции функций

Представьте, что в функции

$$f = y^2$$

переменная **y** сама является функцией:

$$y = x^3 + x$$

Вот как выглядит новая закономерность.

$$\boxed{\frac{\delta f}{\delta x} = \frac{\delta f}{\delta y} \cdot \frac{\delta y}{\delta x}}$$

Это очень мощный результат, который называется **цепным правилом.**

Предположим, имеется функция

$$f = 2xy + 3x^2z + 4z$$

Вот как выглядит окончательный ответ:

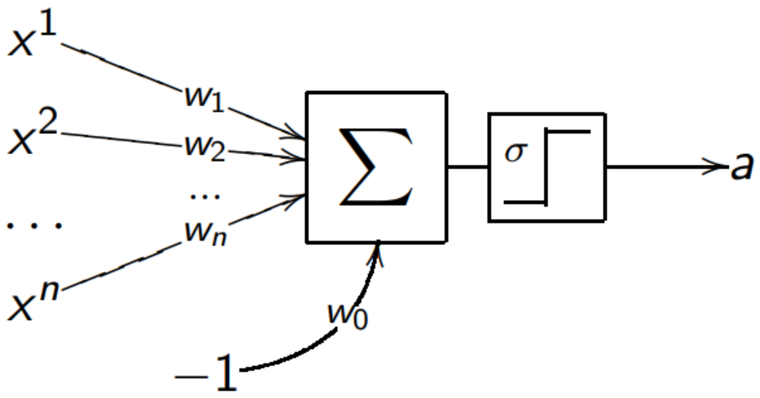
$$\frac{\delta f}{\delta x} = 2y + 6xz$$

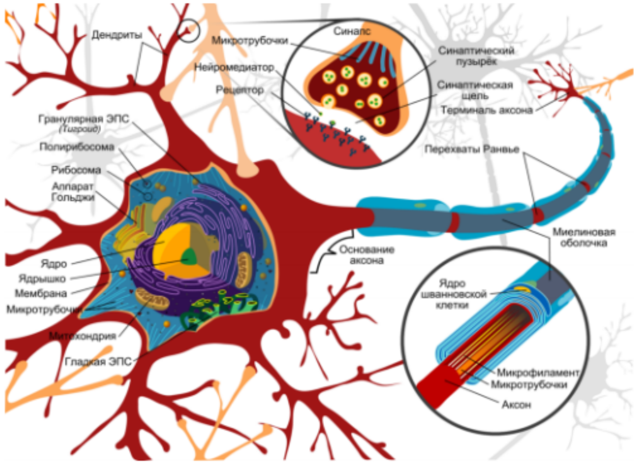
Нейронные сети прямого распространения (feed forward networks)

$f_j \colon X \to \mathbb{R}$, $j = 1, \dots, n$ — числовые признаки;

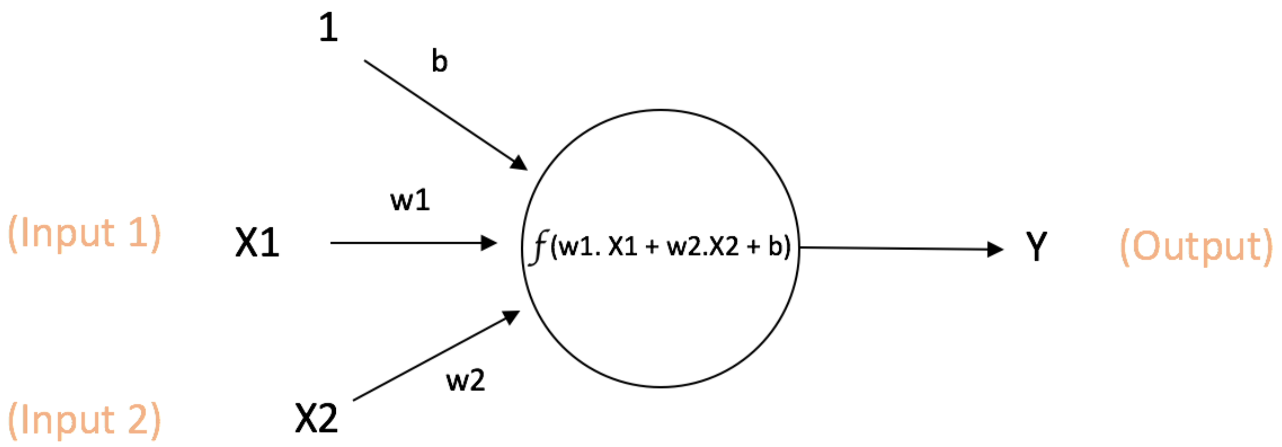
$$a(x, w) = \sigma(\langle w, x \rangle) = \sigma\left(\sum_{j=1}^{n} w_j f_j(x) - w_0 \right),$$

где $w_0, w_1, \dots, w_n \in \mathbb{R}$ — веса признаков;
$\sigma(z)$ — функция активации, например, $\operatorname{sign}(z)$, $\frac{1}{1+e^{-z}}$, $(z)_+$





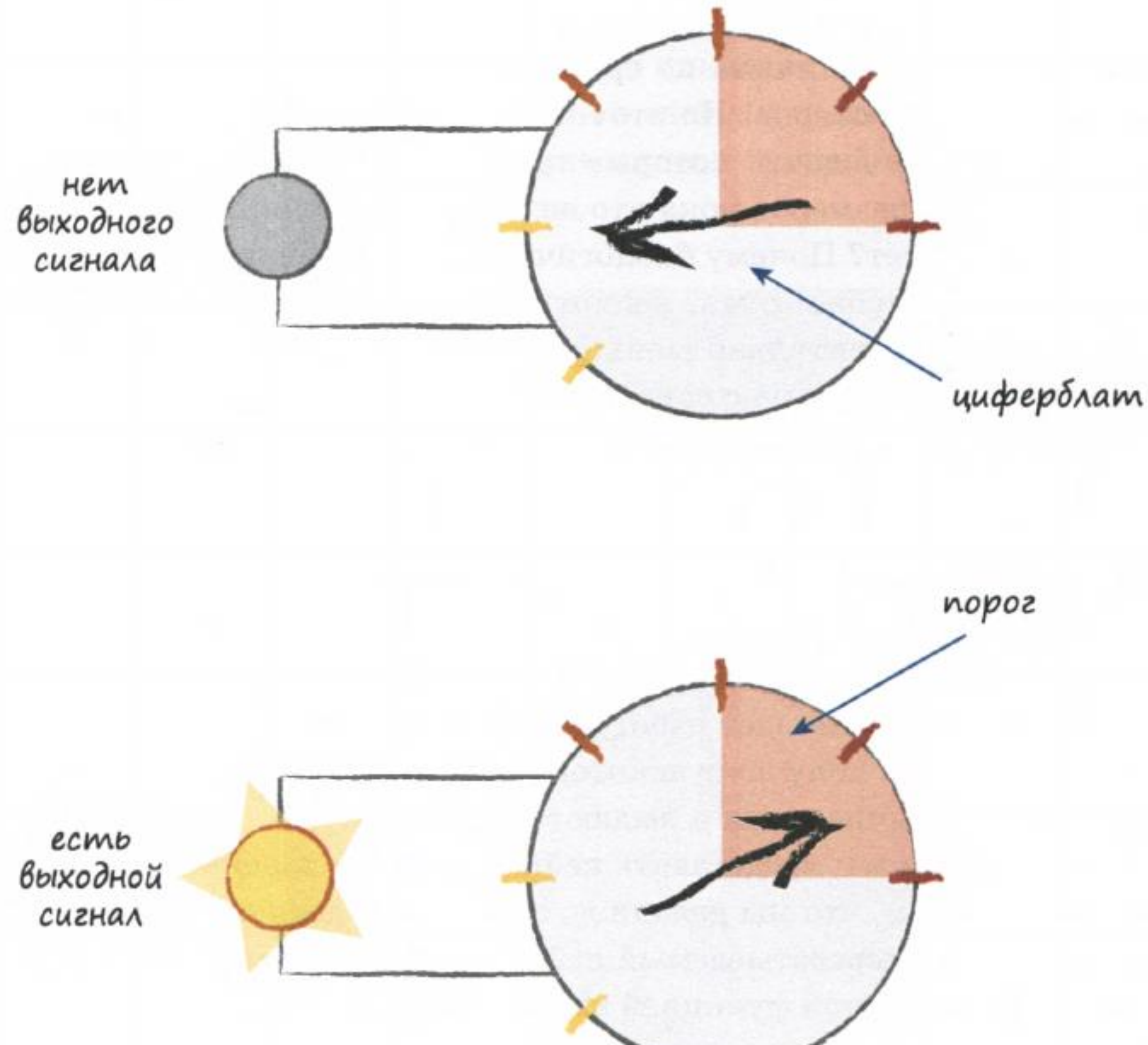
Модель нейрона



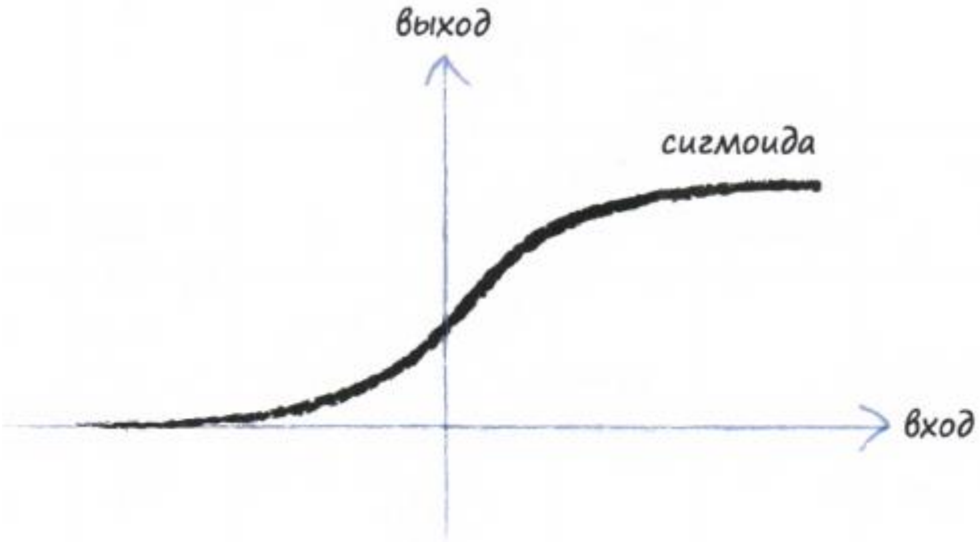
Output of neuron = Y= f(w1. X1 + w2.X2 + b)

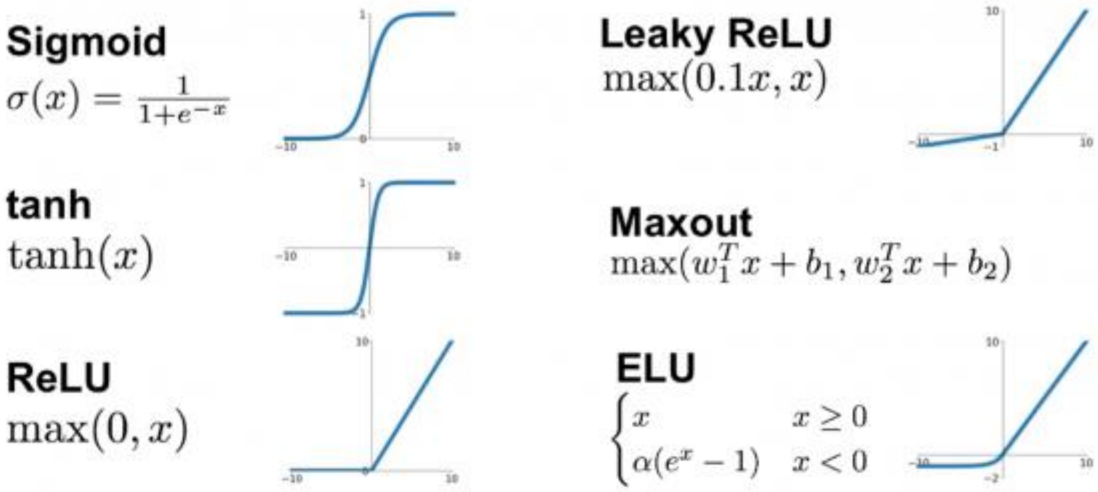
$$x_1 w_1 + x_2 w_2 + x_3 w_3 + b$$

Функция активации

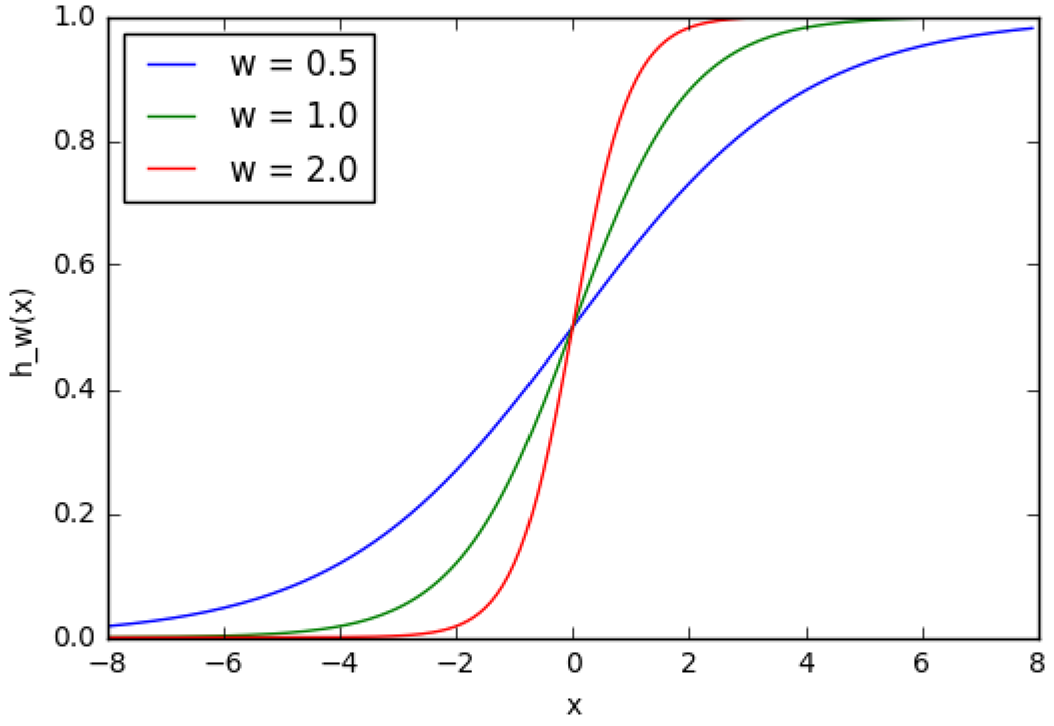


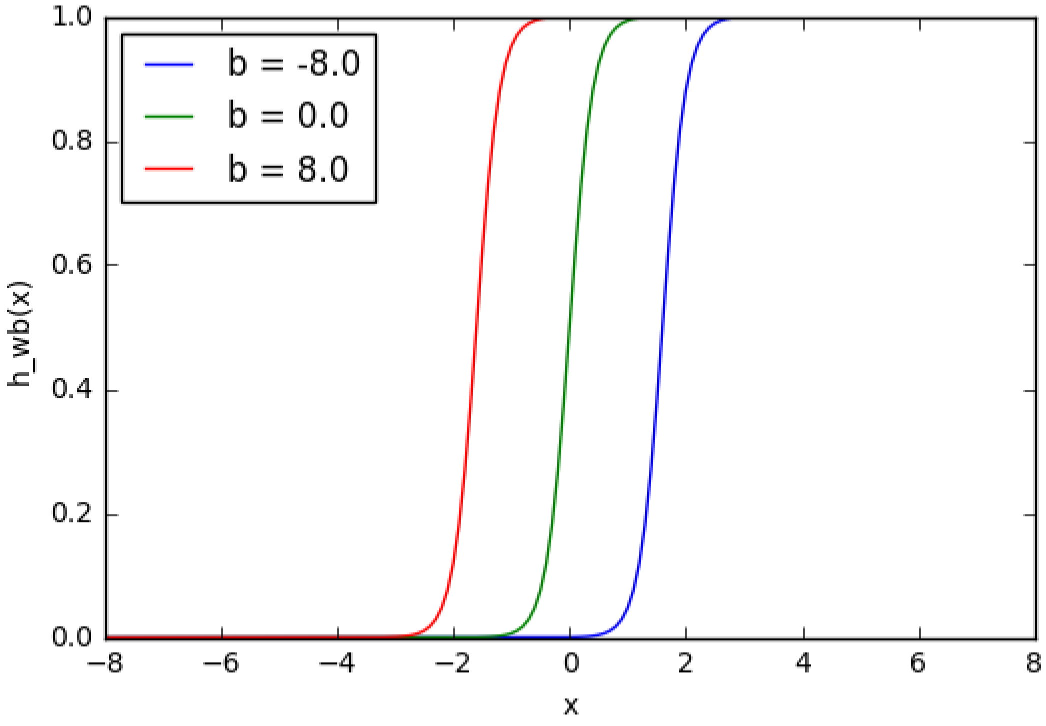

выход
ступенчатая функция
вход
порог


выход
сигмоида
вход


Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$
tanh
$\tanh(x)$
ReLU
$\max(0, x)$
Leaky ReLU
$\max(0.1x, x)$
Maxout
$\max(w_1^T x + b_1, w_2^T x + b_2)$
ELU
$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$

Поведение функции активации





- С помощью линейных операций и одной нелинейной *функции активации* σ можно приблизить любую непрерывную функцию с любой желаемой точностью.

Практические рекомендации:

- Двух-трёх слоёв теоретически достаточно.
- Глубокие сети — это встроенное обучение признаков.

Функция $\sigma(z)$ — *сигмоида*, если $\lim\limits_{z\to-\infty}\sigma(z)=0$ и $\lim\limits_{z\to+\infty}\sigma(z)=1$.

Теорема Цыбенко (1989)

Если $\sigma(z)$ — непрерывная сигмоида, то для любой непрерывной на $[0,1]^n$ функции $f(x)$ существуют такие значения параметров $w_h\in\mathbb{R}^n$, $w_0\in\mathbb{R}$, $\alpha_h\in\mathbb{R}$, что двухслойная сеть

$$a(x)=\sum_{h=1}^{H}\alpha_h\sigma\bigl(\langle x,w_h\rangle+w_0\bigr)$$

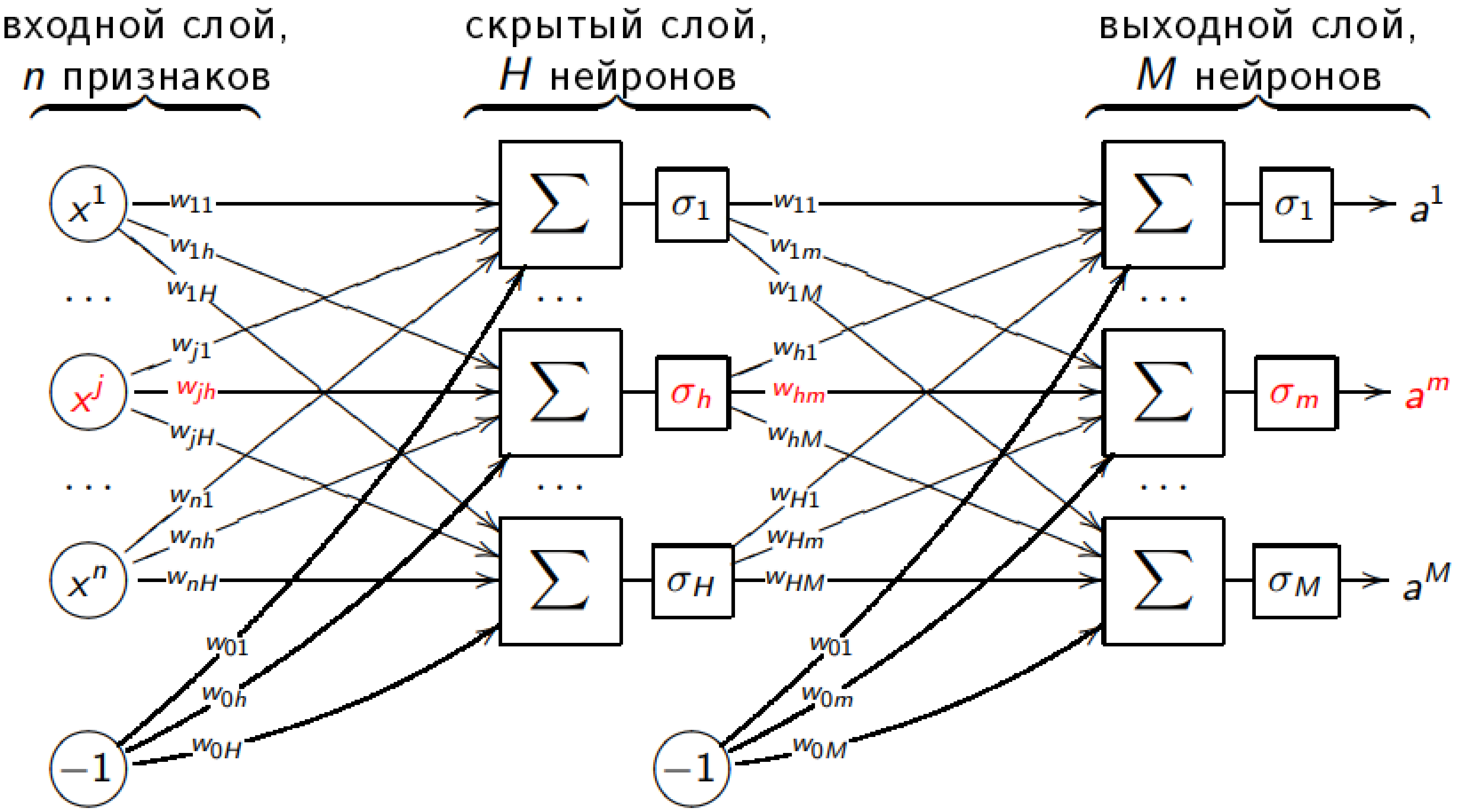
равномерно приближает $f(x)$ с любой точностью ε:

$$\bigl|a(x)-f(x)\bigr|<\varepsilon,\ \text{для всех } x\in[0,1]^n.$$

George Cybenko. Approximation by Superpositions of a Sigmoidal function. Mathematics of Control, Signals, and Systems. 1989.

Оптимизация градиентными методами, обратное распространение ошибки

Пусть для общности $Y = \mathbb{R}^M$, для простоты слоёв только два.



Вектор параметров модели $w \equiv \left(w_{jh}, w_{hm}\right) \in \mathbb{R}^{Hn+H+MH+M}$.

Минимизация средних потерь на обучающей выборке:

$$Q(w) := \frac{1}{\ell} \sum_{i=1}^{\ell} \mathscr{L}_i(w) \to \min_{w}.$$

Вход: выборка X^ℓ; темп обучения η; параметр λ;

Выход: вектор весов $w \equiv \bigl(w_{jh}, w_{hm}\bigr)$;

```
инициализировать веса w и текущую оценку Q(w);
повторять
    выбрать объект x_i из X^ℓ (например, случайно);
    вычислить потерю L_i := L_i(w);
    градиентный шаг: w := w − η L'_i(w);
    оценить значение функционала: Q := (1 − λ)Q + λ L_i;
пока значение Q и/или веса w не стабилизируются;
```

$$S = \sum_{i=1}^{n} x_i w_i, \ (1) \qquad Y = f(S), \ (2) \qquad f(x) = \frac{1}{1 + e^{-ax}}, \ (3) \qquad f'(x) = \alpha f(x)\left(1 - f(x)\right), \ (4)$$

$$E(w) = \frac{1}{2} \sum_{j=1}^{p} (y_i - d_i)^2, \ (5) \qquad \Delta w_{ij} = -\eta \cdot \frac{\partial E}{\partial w_{ij}}, \ (6) \qquad \frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} \cdot \frac{dy_i}{dS_j} \cdot \frac{\partial S_j}{\partial w_{ij}}, \ (7)$$

$$\frac{\partial S_j}{\partial w_{ij}} = x_i, \ (8) \qquad \frac{\partial E}{\partial y_j} = \sum_k \frac{\partial E}{\partial y_k} \cdot \frac{dy_k}{dS_k} \cdot \frac{\partial S_k}{\partial y_j} = \sum_k \frac{\partial E}{\partial y_k} \cdot \frac{dy_k}{dS_k} \cdot w_{jk}^{(n+1)}, \ (9)$$

$$\delta_j^{(n)} = \frac{\partial E}{\partial y_j} \cdot \frac{dy_j}{dS_j}, \ (10) \qquad \delta_j^{(n)} = \left[\sum_k \delta_k^{(n+1)} \cdot w_{jk}^{(n+1)}\right] \cdot \frac{dy_j}{dS_j}, \ (11) \qquad \delta_j^{(N)} = \left(y_i^{(N)} - d_i\right) \cdot \frac{dy_j}{dS_j}, \ (12)$$

$$\Delta w_{ij}^{(n)} = -\eta \cdot \delta_j^{(n)} \cdot x_i^n, \ (13)$$

Рассмотрим теперь полный алгоритм обучения нейросети:

1. подать на вход НС один из требуемых образов и определить значения выходов нейронов нейросети
2. рассчитать для выходного слоя НС по формуле (12) и рассчитать изменения весов выходного слоя N по формуле (13)
3. Рассчитать по формулам (11) и (13) соответственно и $\Delta w_{ij}^{(N)}$ для остальных слоев НС, *n = N-1..1*
4. Скорректировать все веса НС
$$w_{ij}^{(n)}(t) = w_{ij}^{(n)}(t-1) + \Delta w_{ij}^{(n)}(t), \quad (14)$$
5. Если ошибка существенна, то перейти на шаг 1

Выходные значения сети $a^m(x_i)$, $m = 1..M$ на объекте x_i:

$$a^m(x_i) = \sigma_m\Biggl(\sum_{h=0}^{H} w_{hm} u^h(x_i)\Biggr); \qquad u^h(x_i) = \sigma_h\Biggl(\sum_{j=0}^{J} w_{jh} f_j(x_i)\Biggr).$$

Без ограничения общности (только для примера) будем рассматривать среднеквадратичную функцию потерь:

$$\mathscr{L}_i(w) = \frac{1}{2}\sum_{m=1}^{M}\bigl(a^m(x_i) - y_i^m\bigr)^2.$$

Промежуточная задача: найти частные производные

$$\frac{\partial \mathscr{L}_i(w)}{\partial a^m}; \qquad \frac{\partial \mathscr{L}_i(w)}{\partial u^h}.$$

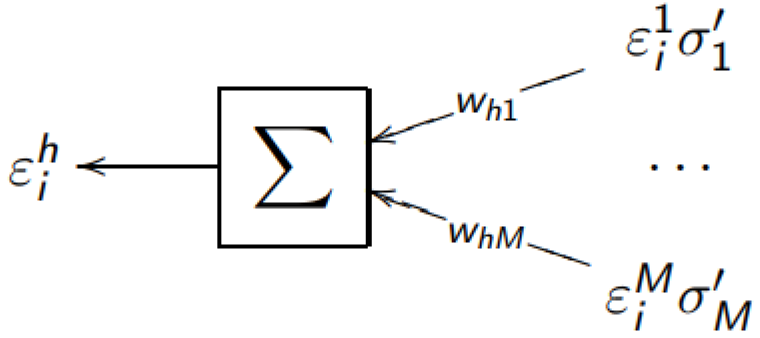
Промежуточная задача: частные производные

$$\frac{\partial \mathscr{L}_i(w)}{\partial a^m} = a^m(x_i) - y_i^m = \varepsilon_i^m$$

— это ошибка на выходном слое (для квадратичных потерь);

$$\frac{\partial \mathscr{L}_i(w)}{\partial u^h} = \sum_{m=1}^{M} \frac{\partial \mathscr{L}_i(w)}{\partial a^m} \sigma'_m(\cdot) w_{hm} = \sum_{m=1}^{M} \varepsilon_i^m \sigma'_m w_{hm} = \varepsilon_i^h$$

— назовём это *ошибкой на скрытом слое*. Похоже, что ε_i^h вычисляется по ε_i^m, если запустить сеть «задом наперёд»:



Теперь, имея частные производные $\mathscr{L}_i(w)$ по a^m и u^h, легко выписать градиент $\mathscr{L}_i(w)$ по весам w:

$$\frac{\partial \mathscr{L}_i(w)}{\partial w_{hm}} = \frac{\partial \mathscr{L}_i(w)}{\partial a^m} \frac{\partial a^m}{\partial w_{hm}} = \varepsilon_i^m \sigma'_m u^h(x_i), \quad m = 1..M, \quad h = 0..H;$$

$$\frac{\partial \mathscr{L}_i(w)}{\partial w_{jh}} = \frac{\partial \mathscr{L}_i(w)}{\partial u^h} \frac{\partial u^h}{\partial w_{jh}} = \varepsilon_i^h \sigma'_h f_j(x_i), \quad h = 1..H, \quad j = 0..n;$$

Алгоритм обратного распространения ошибки BackProp:

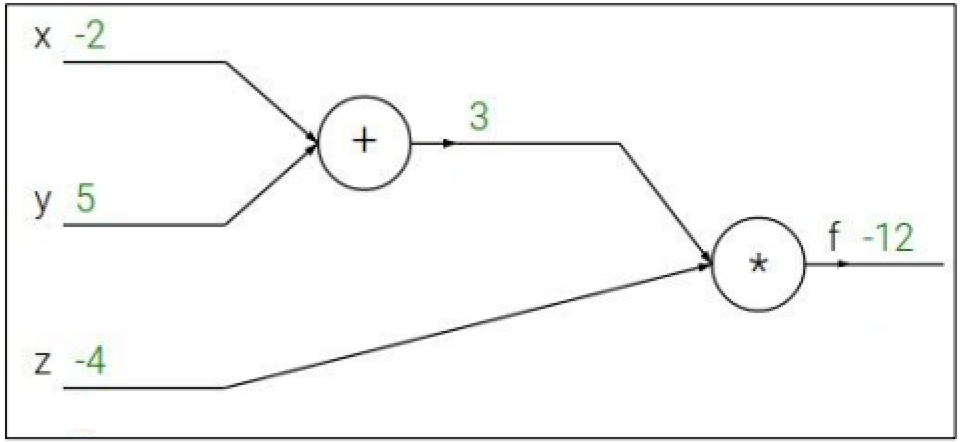
Вход: $X^\ell = (x_i, y_i)_{i=1}^\ell \subset \mathbb{R}^n \times \mathbb{R}^M$; параметры H, λ, η;
Выход: синаптические веса w_{jh}, w_{hm};

1: инициализировать веса w_{jh}, w_{hm};
2: **повторять**
3: выбрать объект x_i из X^ℓ (например, случайно);
4: прямой ход:
$u_i^h := \sigma_h\bigl(\sum_{j=0}^J w_{jh} x_i^j\bigr)$, $h = 1..H$;
$a_i^m := \sigma_m\bigl(\sum_{h=0}^H w_{hm} u_i^h\bigr)$, $m = 1..M$;
$\varepsilon_i^m := \frac{\partial \mathscr{L}_i(w)}{\partial a^m}$, $m = 1..M$;
5: $Q := (1-\lambda)Q + \lambda \mathscr{L}_i(w)$;
6: обратный ход:
$\varepsilon_i^h := \sum_{m=1}^M \varepsilon_i^m \sigma'_m w_{hm}$, $h = 1..H$;
7: градиентный шаг:
$w_{hm} := w_{hm} - \eta \varepsilon_i^m \sigma'_m u_i^h$, $h = 0..H$, $m = 1..M$;
$w_{jh} := w_{jh} - \eta \varepsilon_i^h \sigma'_h x_i^j$, $j = 0..n$, $h = 1..H$;
8: **пока** Q не стабилизируется;

Backpropagation: a simple example

$$f(x, y, z) = (x + y)z$$

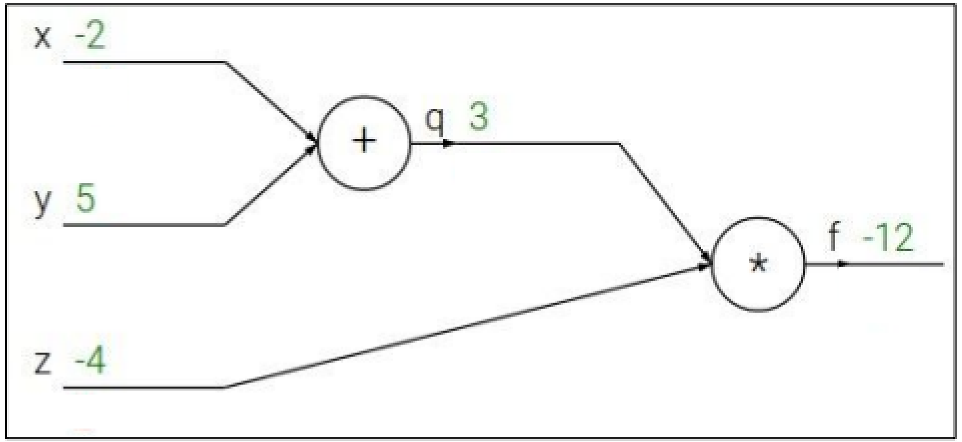
e.g. x = -2, y = 5, z = -4



Backpropagation: a simple example

$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4



$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

Backpropagation: a simple example

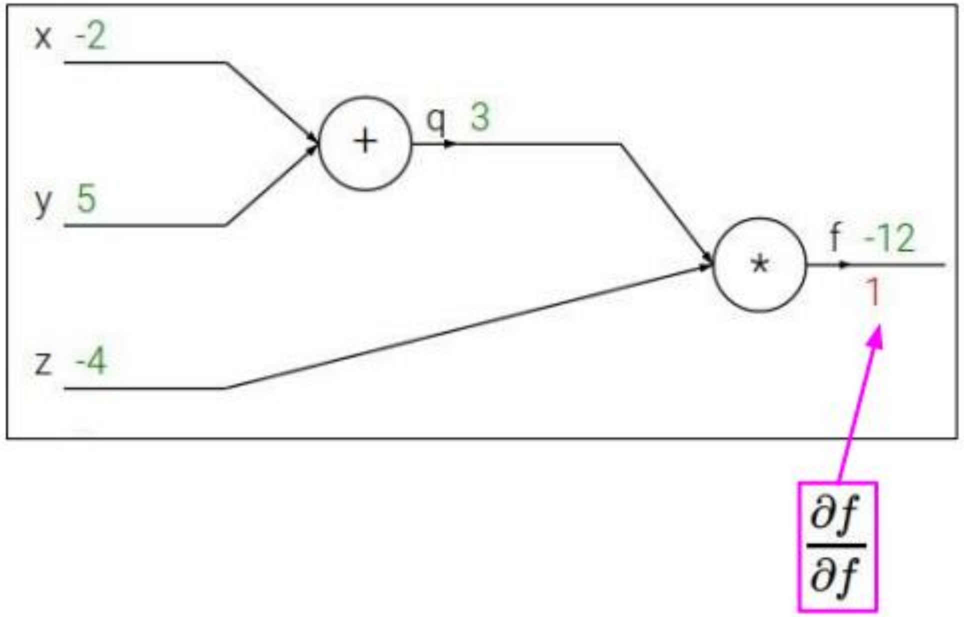
$$f(x,y,z) = (x+y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Backpropagation: a simple example

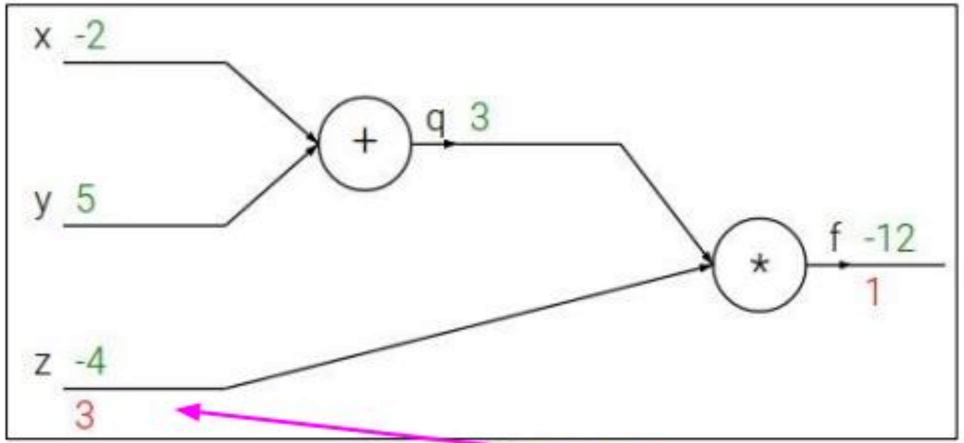
$$f(x,y,z) = (x+y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial z}$$

Backpropagation: a simple example

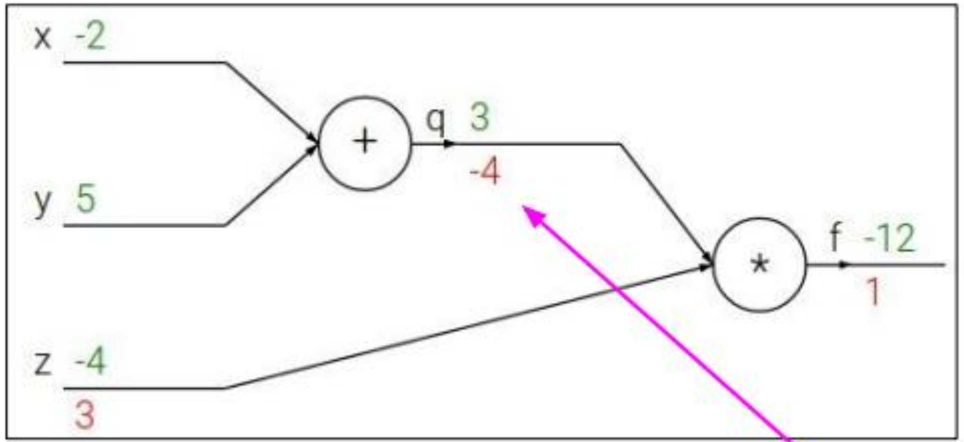
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

$$\boxed{q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1}$$

$$\boxed{f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q}$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\boxed{\frac{\partial f}{\partial q}}$$

Backpropagation: a simple example

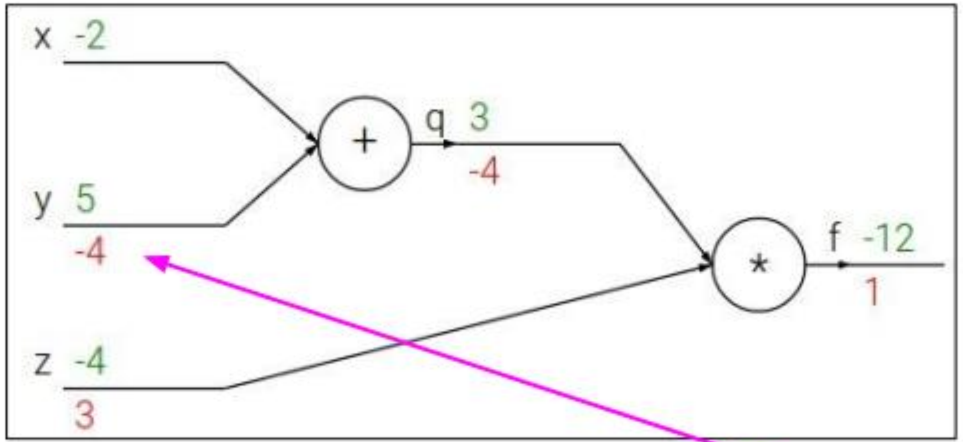
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial y}$$

Chain rule:

$$\boxed{\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y}}$$

Backpropagation: a simple example

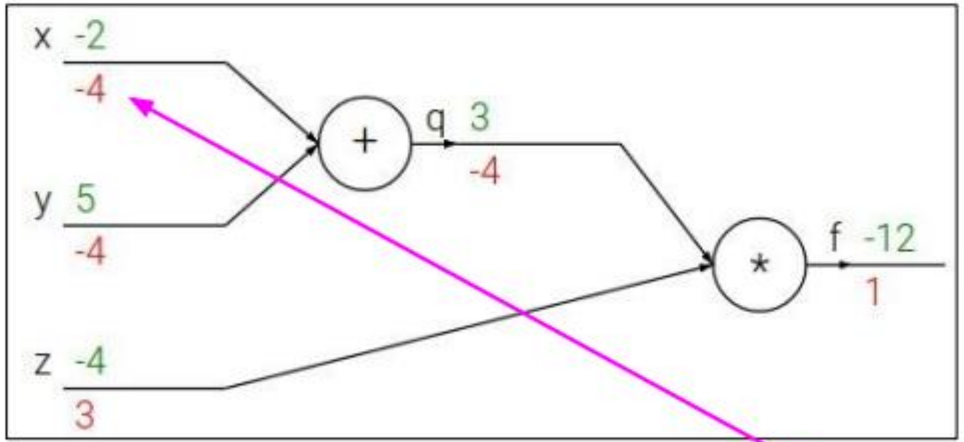
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

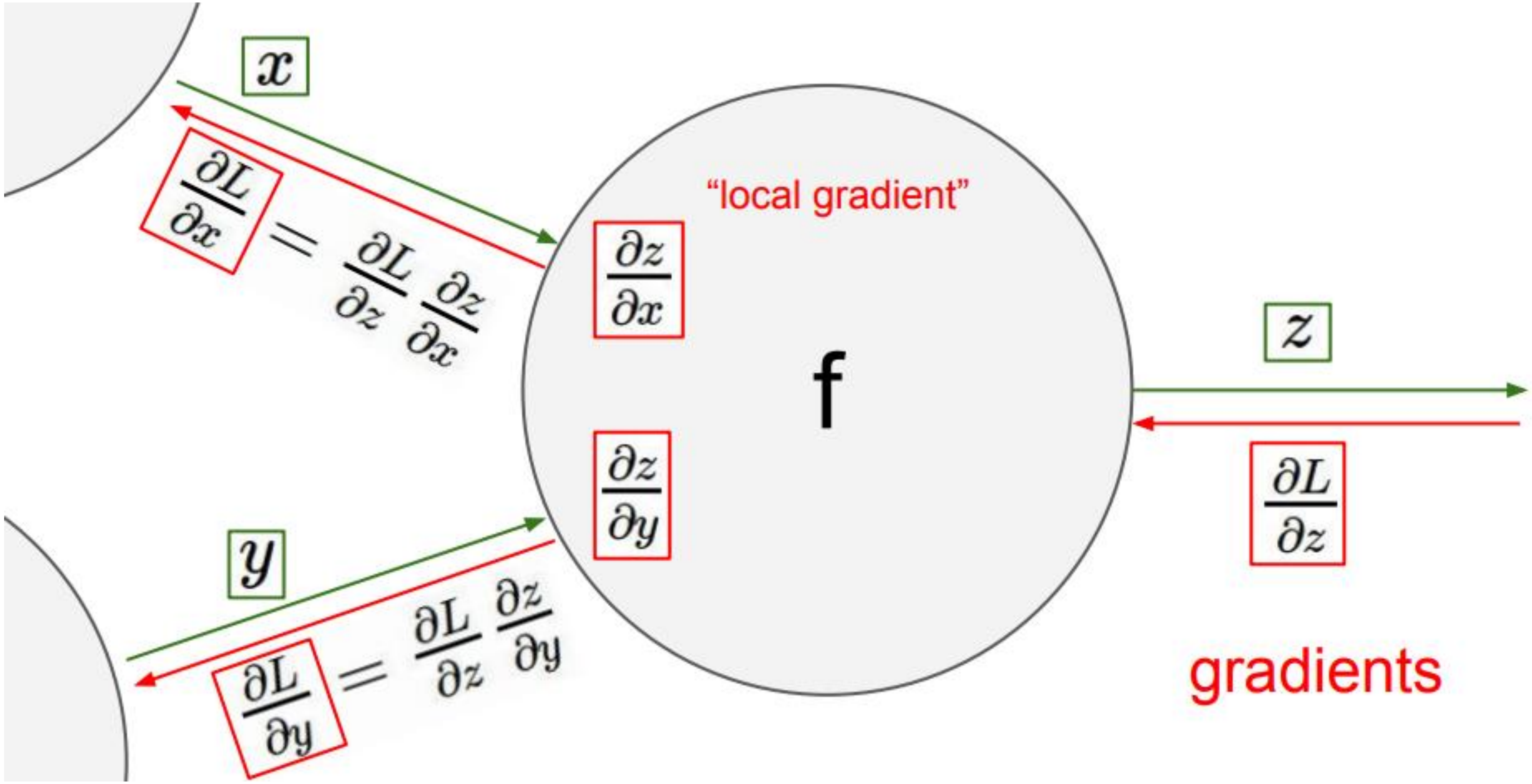
Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$$\frac{\partial f}{\partial x}$$

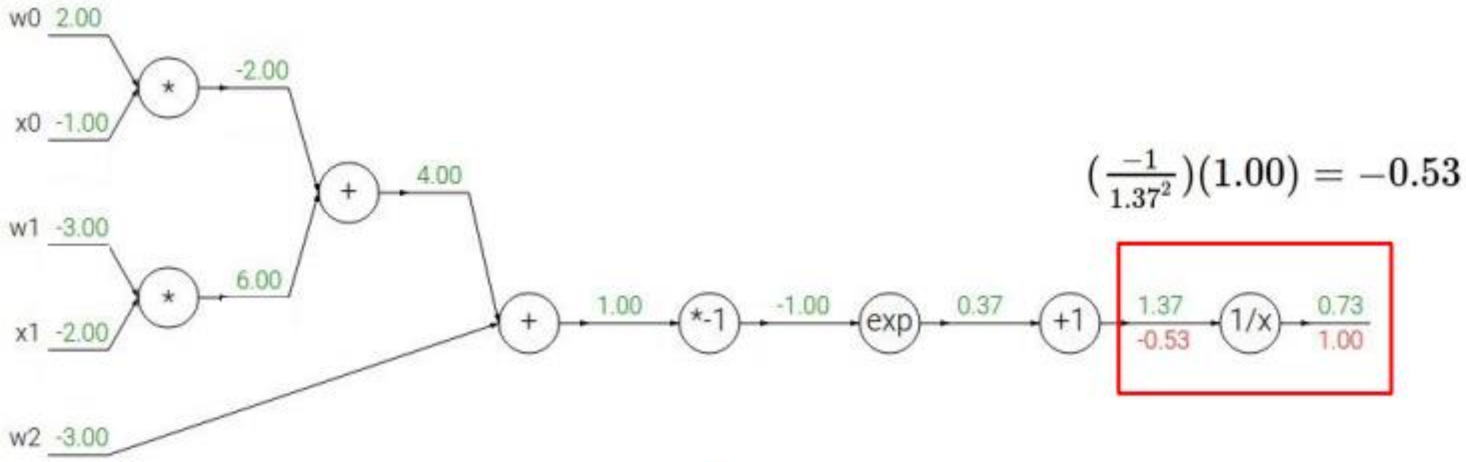
Chain rule:

$$\boxed{\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial x}}$$



x
$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial x}$
"local gradient"
$\frac{\partial z}{\partial x}$
f
$\frac{\partial z}{\partial y}$
z
$\frac{\partial L}{\partial z}$
y
$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial y}$
gradients

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



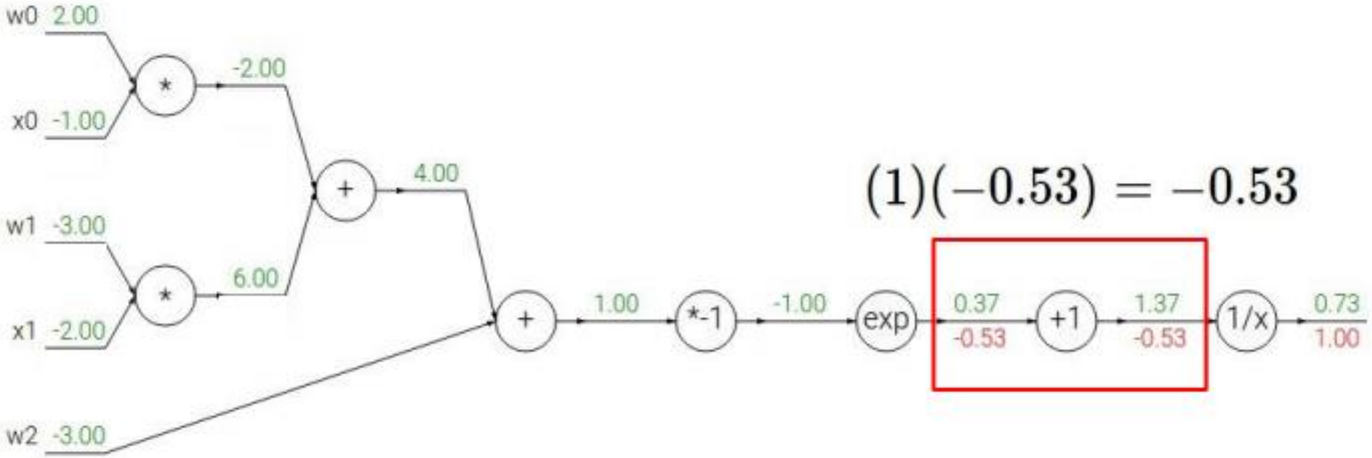
$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$$(1)(-0.53) = -0.53$$

$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$\boxed{f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1}$$