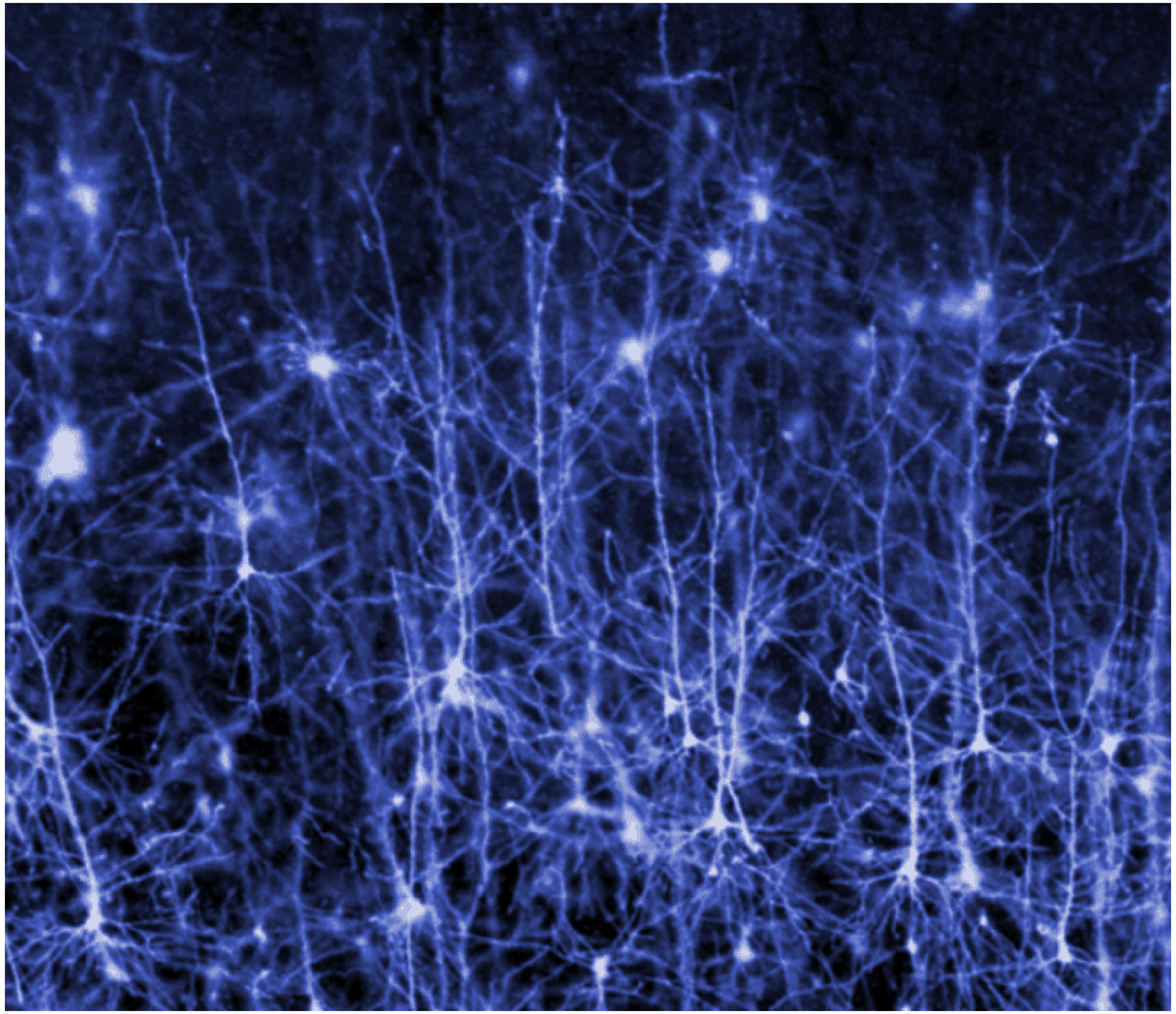
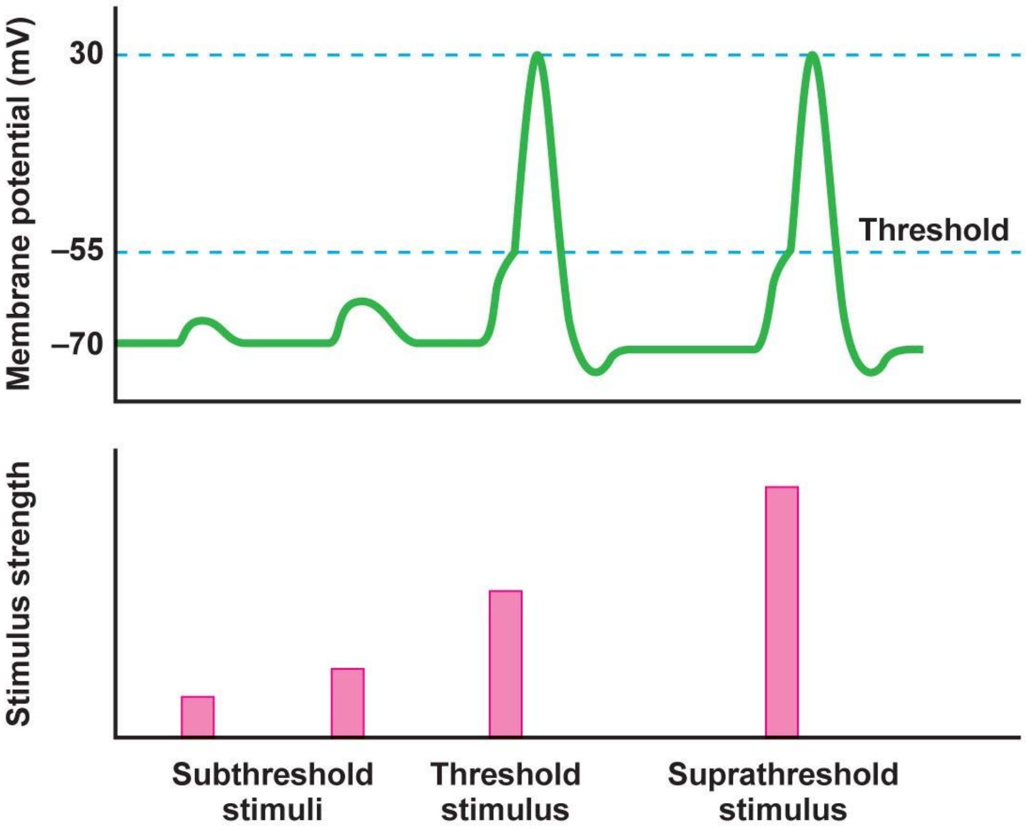
Нейронные сети

Нейронные сети

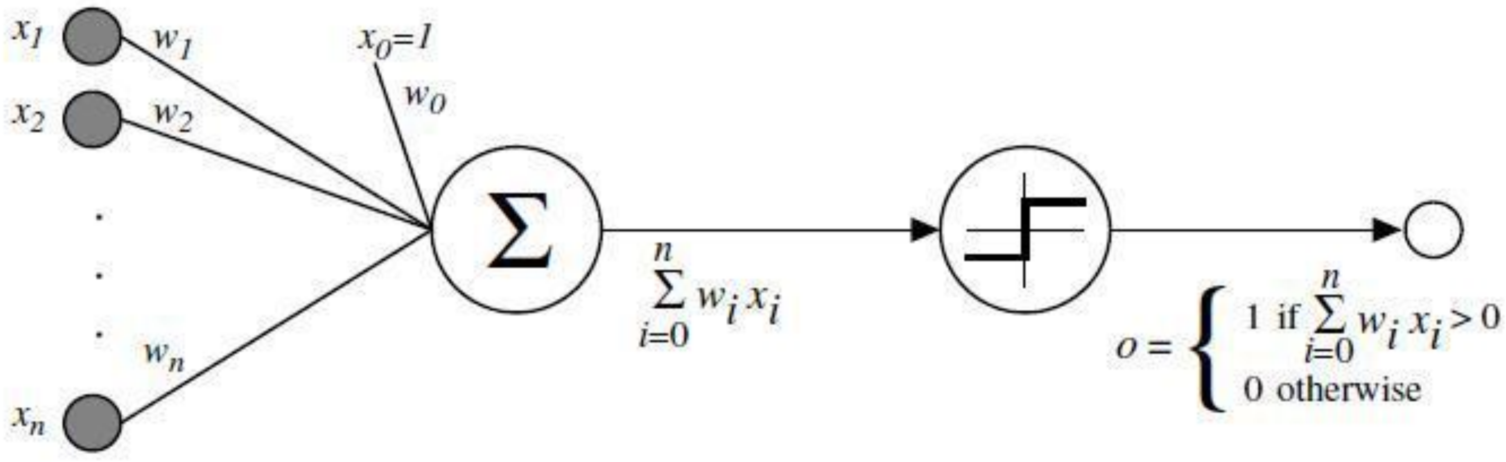


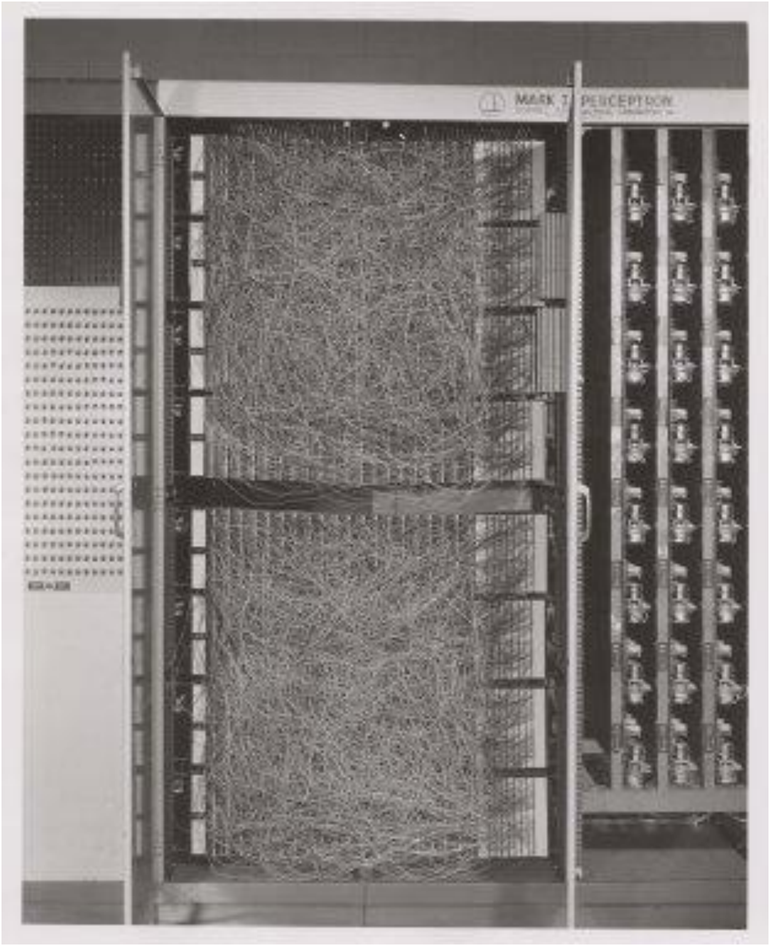


Принцип «все или ничего»

Нейронные сети

Перцептрон (Frank Rosenblatt, 1957)

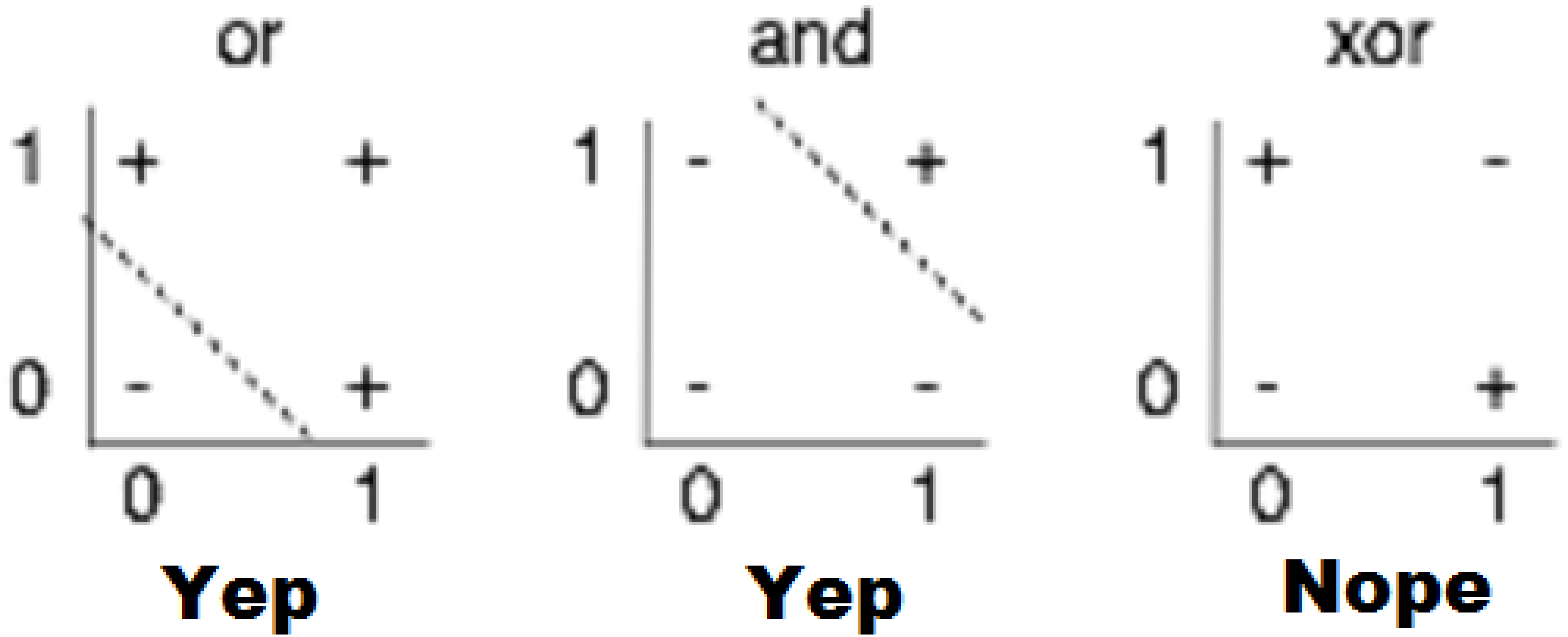




$$\operatorname{sign}(\mathbf{w}^\intercal \mathbf{x}_n) \neq y_n$$

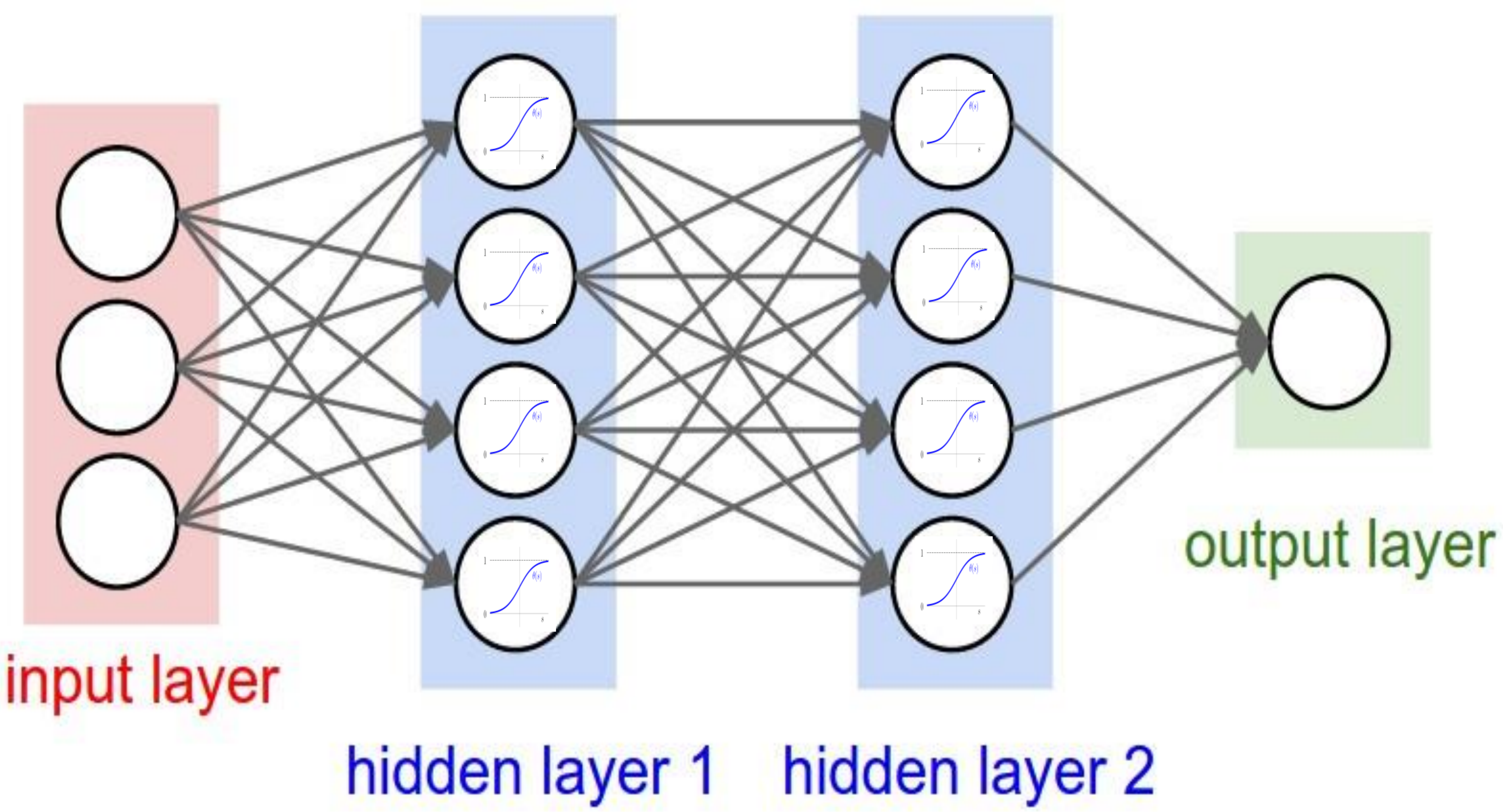
$$\mathbf{w} \leftarrow \mathbf{w} + y_n \mathbf{x}_n$$

Главный недостаток
Перцептронных нейронных сетей



Marvin Minsky and Seymour Papert, 1969

Нейронные сети с нелинейными преобразованиями

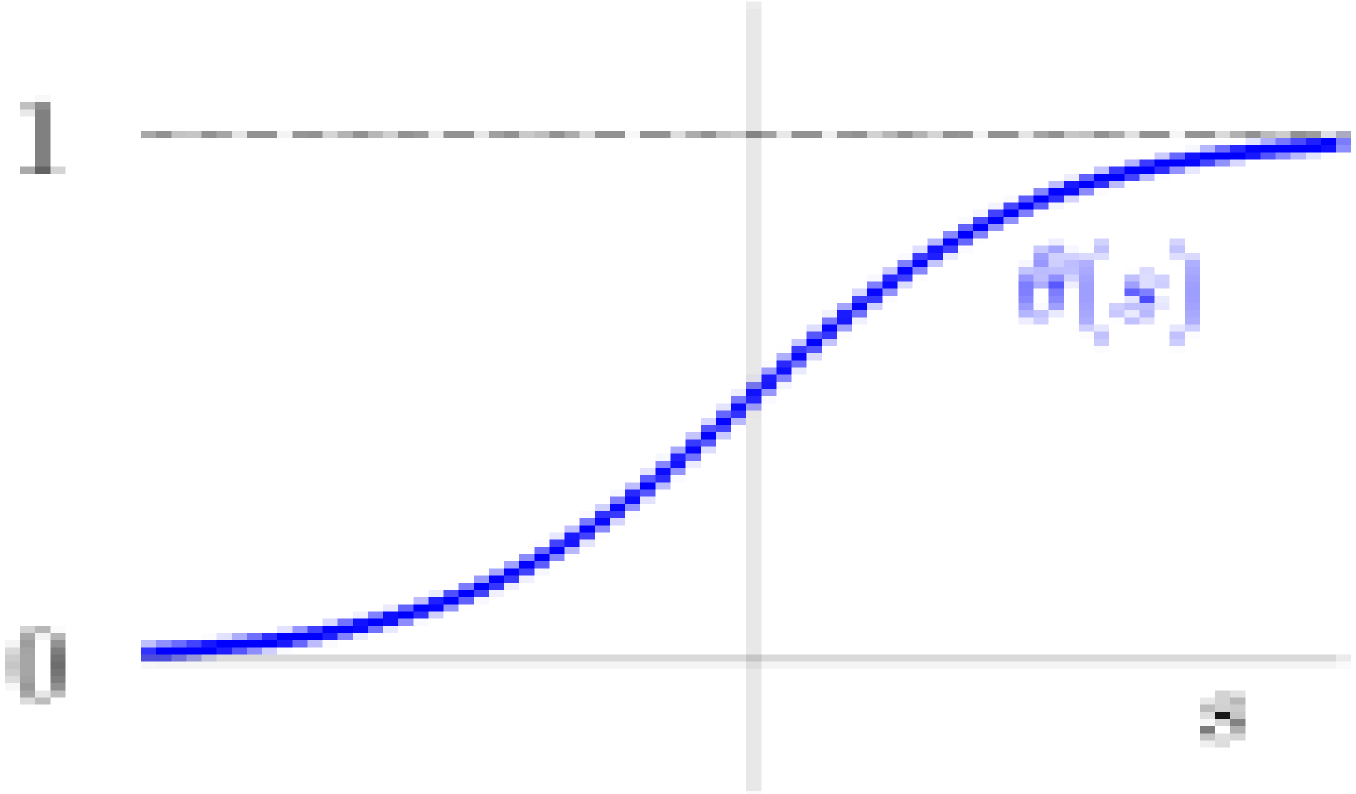


Логистическая функция

$$P(y \mid \mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{for } y = +1; \\ 1 - f(\mathbf{x}) & \text{for } y = -1. \end{cases}$$

$$\sigma(s) = \frac{1}{1 + e^{-s}}$$

$$\sigma(-s) = 1 - \sigma(s)$$



Логистическая регрессия

$$P(y \mid \mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{for } y = +1; \\ 1 - f(\mathbf{x}) & \text{for } y = -1. \end{cases}$$

$$\sigma(-s) = 1 - \sigma(s)$$

$$P(y \mid \mathbf{x}) = \sigma(y\ \mathbf{w}^{\mathsf{T}}\mathbf{x})$$

Правдоподобие: $$\prod_{i=1}^{N} P(y_i | \mathrm{x}_i) = \prod_{i=1}^{N} \sigma\left(y_i \mathrm{w}^{\mathrm{T}} \mathrm{x}_i\right)$$

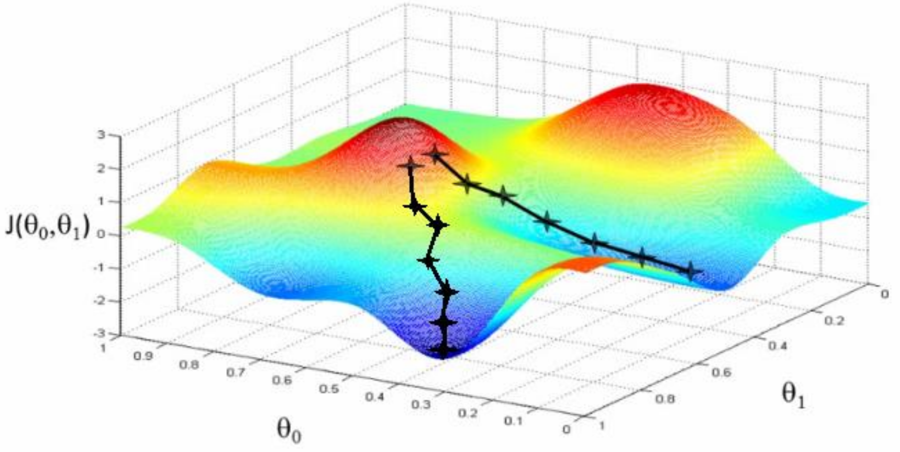
Функция потерь логистической регрессии

$$L(\mathrm{w}) = -\frac{1}{N}\ln\left(\prod_{i=1}^{N}\sigma\left(y_i\mathrm{w}^{\mathrm{T}}\mathrm{x}_i\right)\right) = \frac{1}{N}\sum_{i=1}^{N}\ln\left(\frac{1}{\sigma(y_i\mathrm{w}^{\mathrm{T}}\mathrm{x}_i)}\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\ln(1 + e^{-y_i\mathrm{w}^{\mathrm{T}}\mathrm{x}_i})$$

Градиентный спуск (Gradient Descent)

$$w(t+1) = w(t) - \eta \frac{\partial C(w)}{\partial w}$$



Стохастический градиентный спуск (SGD): *C(w)* считается по малой части обучающей выборки (вплоть до одного примера).

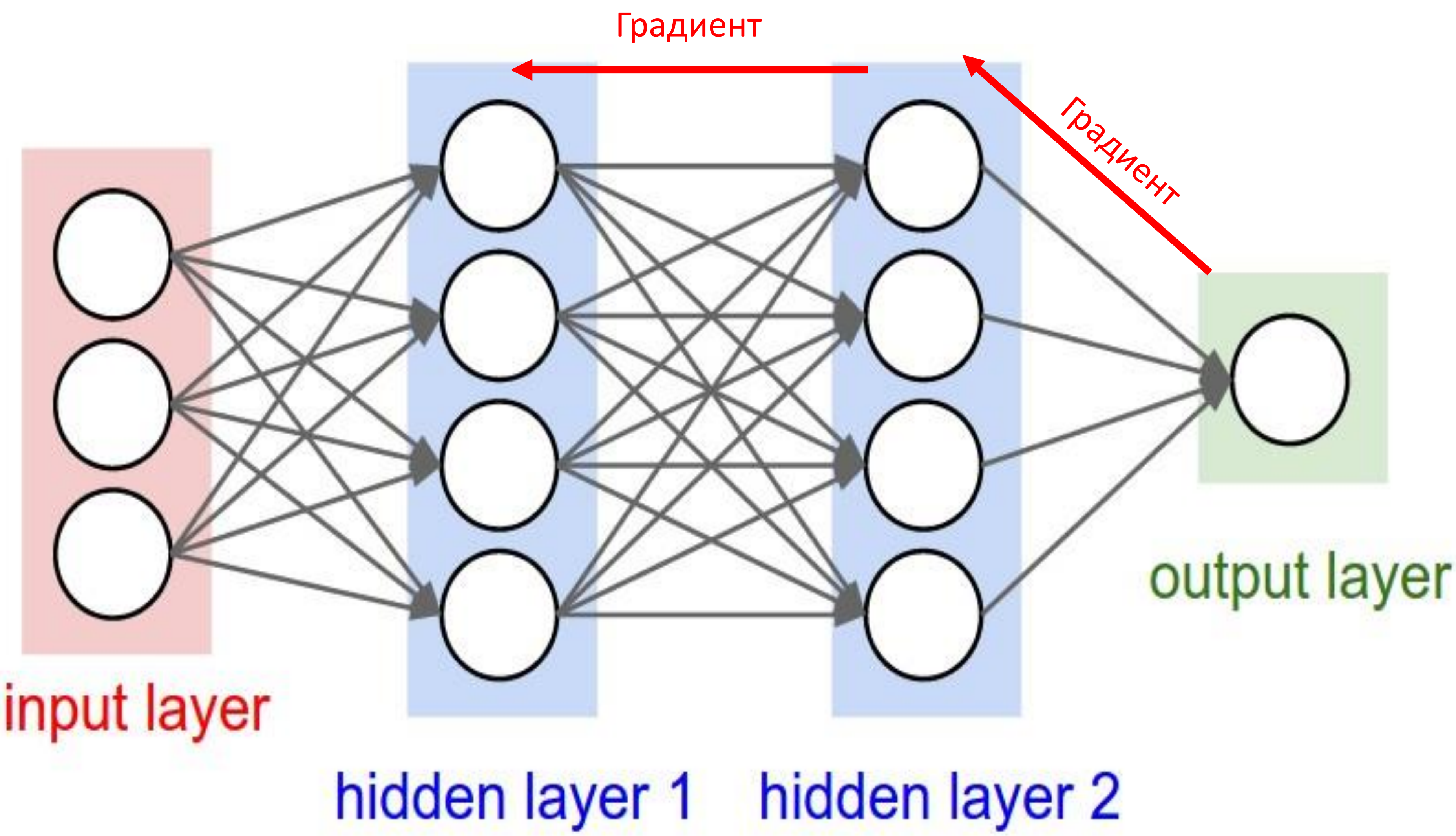
Градиентный спуск

$$\mathrm{w} \leftarrow \mathrm{w} - \eta \left(-\frac{1}{N} \sum_{i=1}^{N} \frac{y_i \mathrm{x}_i}{1 + e^{y_i \mathrm{w}^{\mathrm{T}} \mathrm{x}_i}} \right)$$

Стохастический градиентный спуск

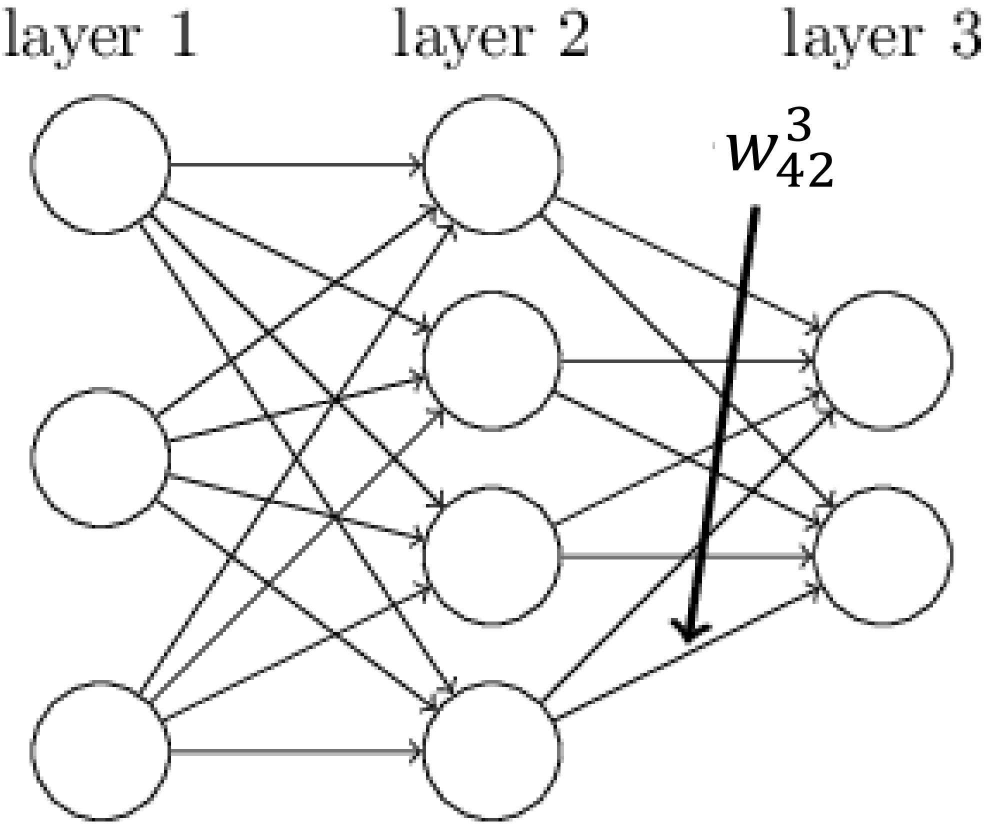
$$\mathrm{w} \leftarrow \mathrm{w} - \eta \frac{\partial \mathrm{C}\left(\mathrm{w}^{\mathrm{T}} \mathrm{x}_i, y_i\right)}{\partial \mathrm{w}}$$

Back-propagation



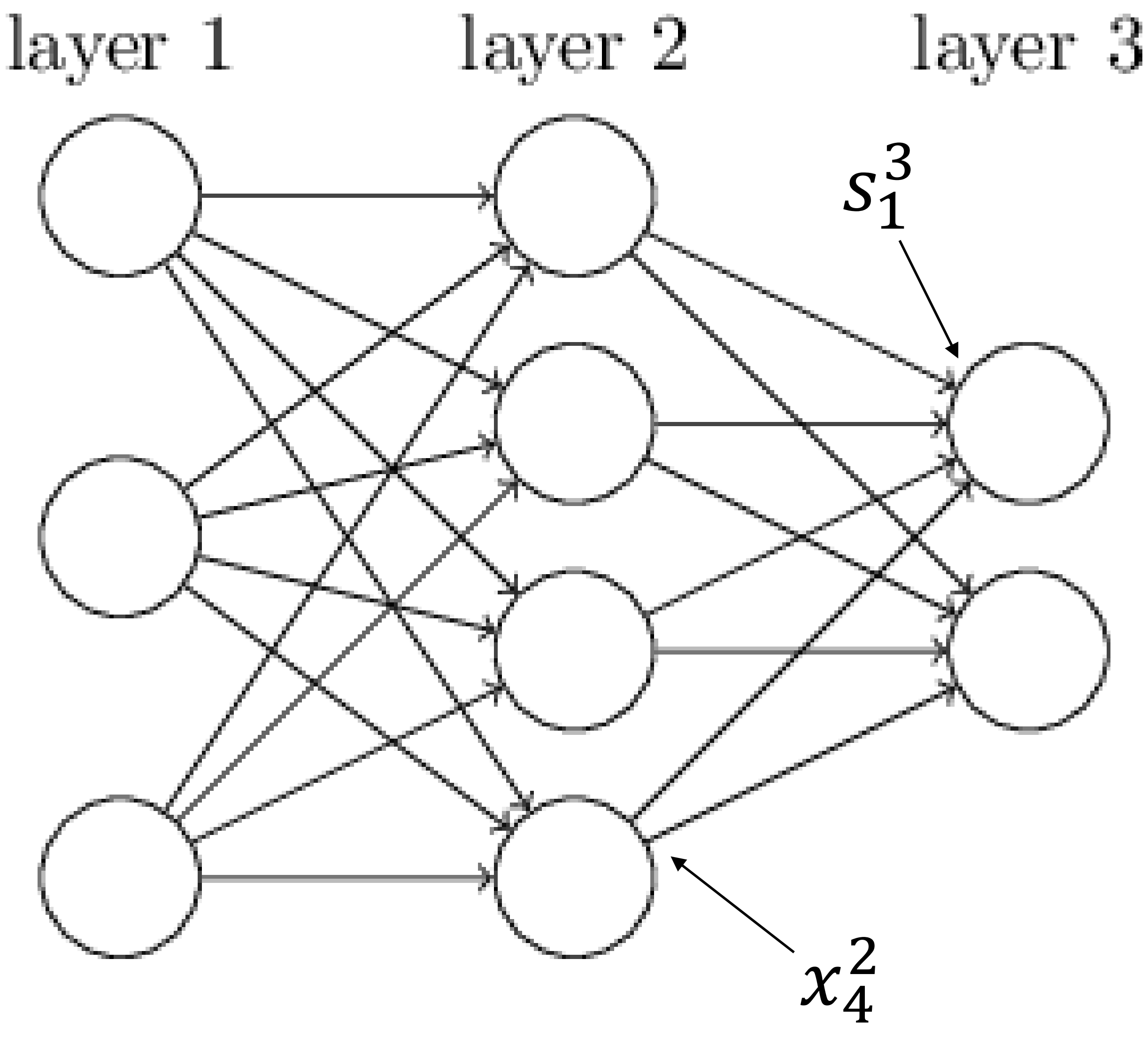
David Rumelhart, Geoffrey Hinton and Rondald Williams, 1986

Back-propagation



w^l_{jk}- это коэффициент (вес) между j-ым нейроном слоя $l-1$, и k-ым нейроном слоя l

Back-propagation



s_j^l - поступающий «заряд» (activation) в j-ом нейроне слоя l на j-ом нейроне слоя l

$$s_j^l = \sum w_{ij}^l x_i^{l-1}$$

x_j^l - «заряд» после функции активации (activation function) в j-ом нейроне слоя l на j-ом нейроне слоя l

$$x_j^l = \sigma(s_j^l) = \sigma\left(\sum w_{ij}^l x_i^{l-1}\right)$$

$$x_j^1 = x_j$$

Back-propagation

Задача – оценить $\nabla C_{w_{ij}^l}(\mathrm{w})$: $\frac{\partial C(\mathrm{w})}{\partial w_{ij}^l} = \frac{\partial C(\mathrm{w})}{\partial s_j^l} \times \frac{\partial s_j^l}{\partial w_{ij}^l}$

$\frac{\partial s_j^l}{\partial w_{ij}^l} = x_i^{l-1}$ - посчитано $\frac{\partial C(\mathrm{w})}{\partial s_j^l} = \delta_j^l$ - посчитаем…

Back-propagation

Последний слой:

$$\delta_i^L = \frac{\partial C(\mathrm{w})}{\partial s_j^L}, \qquad C(\mathrm{w}) = f(\mathrm{x}^L), \qquad \mathrm{x}_i^L = \sigma(s_i^L)$$

Предыдущие слои:

$$\delta_i^{l-1} = \frac{\partial C(\mathrm{w})}{\partial s_j^{l-1}} = \sum \frac{\partial C(\mathrm{w})}{\partial s_j^l} \times \frac{\partial s_j^l}{\partial x_i^{l-1}} \times \frac{\partial x_i^{l-1}}{\partial s_i^{l-1}} = \sum \delta_j^l \times w_{ij}^l \times \sigma'\left(s_i^{l-1}\right)$$

Back-propagation

Алгоритм:

1. Инициализировать веса случайно*
2. Forward: Посчитать все x и s
3. Backward: Посчитать все δ
4. $w_{ij}^{l} \leftarrow w_{ij}^{l} - \eta x_{i}^{l-1} \delta_{j}^{l}$
5. -> 2.

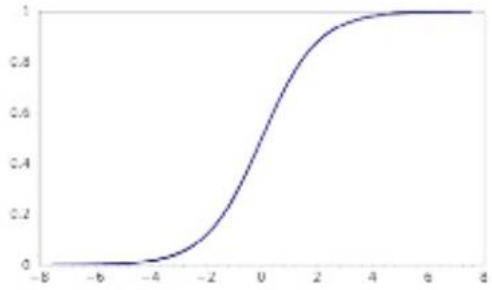
Функции активации:

Сигмоид: $\sigma(s) = \frac{1}{1+e^{-s}}$

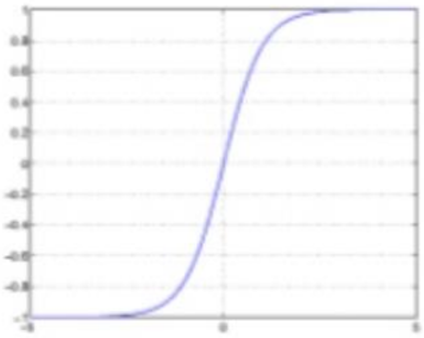
Tanh: $\sigma(s) = \frac{e^{s}-e^{-s}}{e^{s}+e^{-s}}$

Rectified Linear Unit (ReLU):

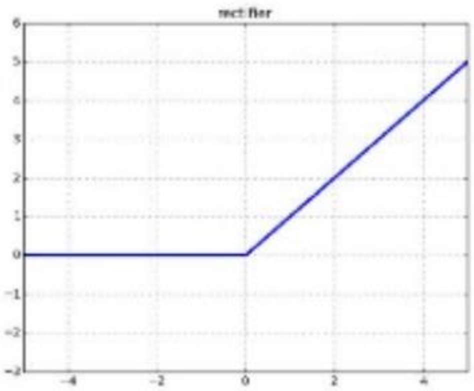
$$\sigma(s) = \max(0, s)$$



Sigmoid

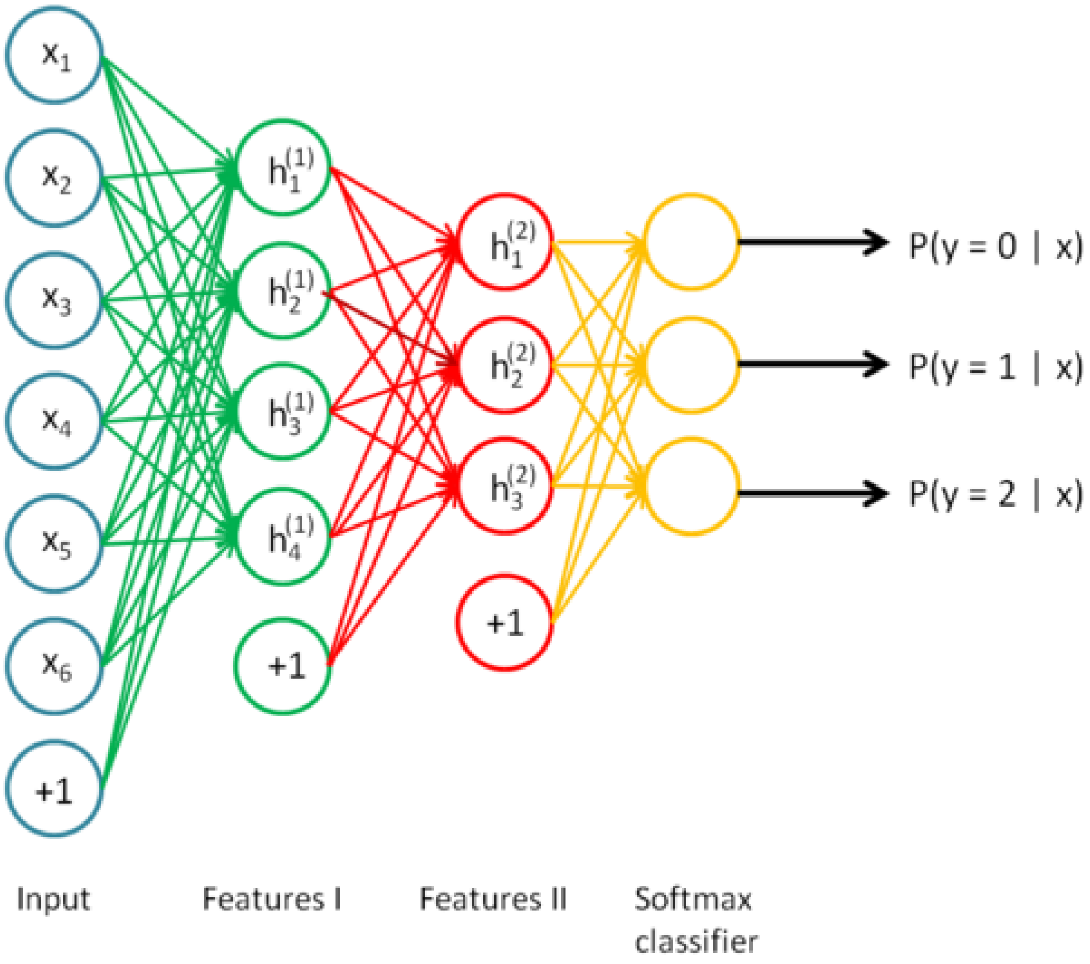


Tanh



ReLU

Функция потерь для мультиклассификации Softmax и Cross-Entropy



$$x_j^L = \sigma^L(s_j^L) = \frac{e^{s_j^L}}{\sum e^{s_i^L}}$$

$$C(\mathrm{w}) = -\sum o_i \log(x_i^L)$$

$$o_i = \begin{cases} 1, \text{если } y = i \\ 0, \text{если } y \neq i \end{cases}$$