Введение: neural network

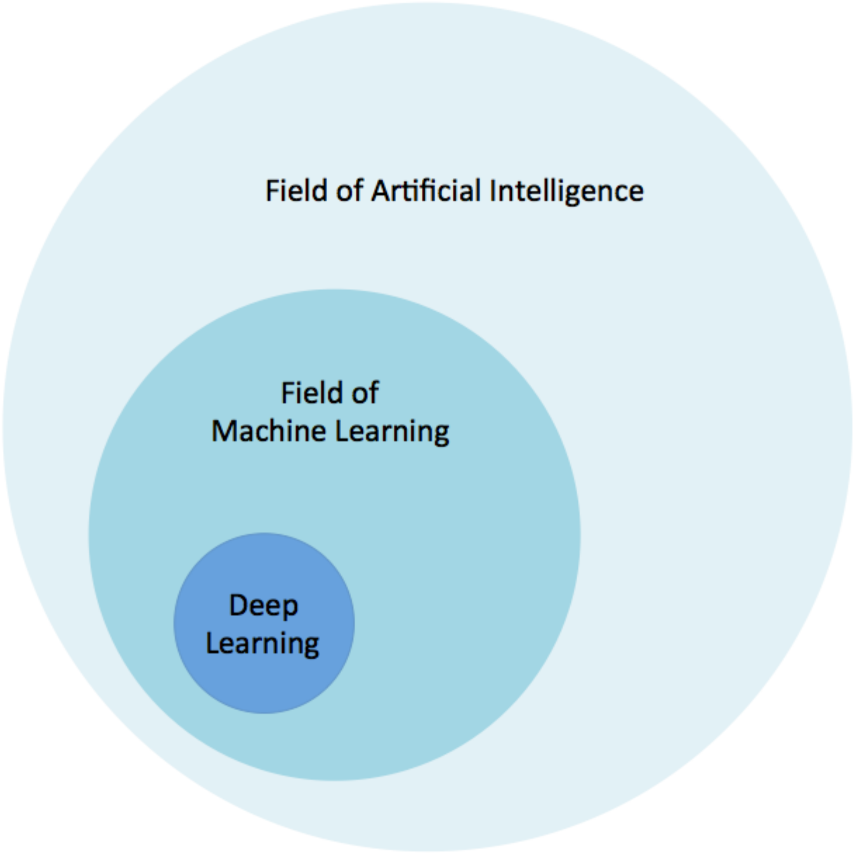
Марина Горлова

В чем разница?

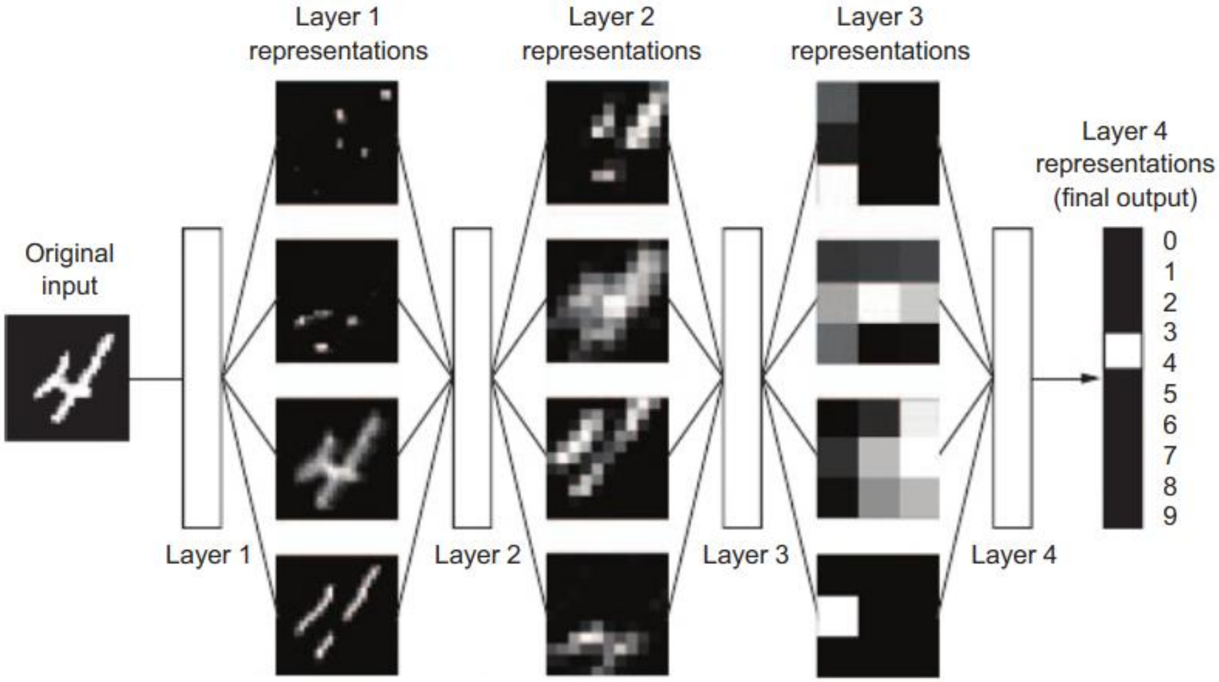
1. Может ли машина выполнять задачи, которые традиционно считаются человеческими?

2. Может ли машина выявлять правила и скрытые закономерности?

3. Может ли машина изучить и обобщить данные?



NN - система многоступенчатой фильтрации информации



Краткая история ML

1. 1950: Вероятностные методы (наивный байес, логистическая регрессия)
2. 1980: Алгоритм обратного распространения

распознавание почтовых индексов LeNet

1. 1990: Ядерные методы (SVM)
2. 2000: Деревья решений, random forests, gradient boosting machines
3. 2010: Нейронные сети

прогресс на датасете ImageNet

Преимущества deep learning

1. Простота подготовки данных:

    не требует предварительного тяжелого feature engineering

2. Масштабируемость:

    параллелизм на GPU

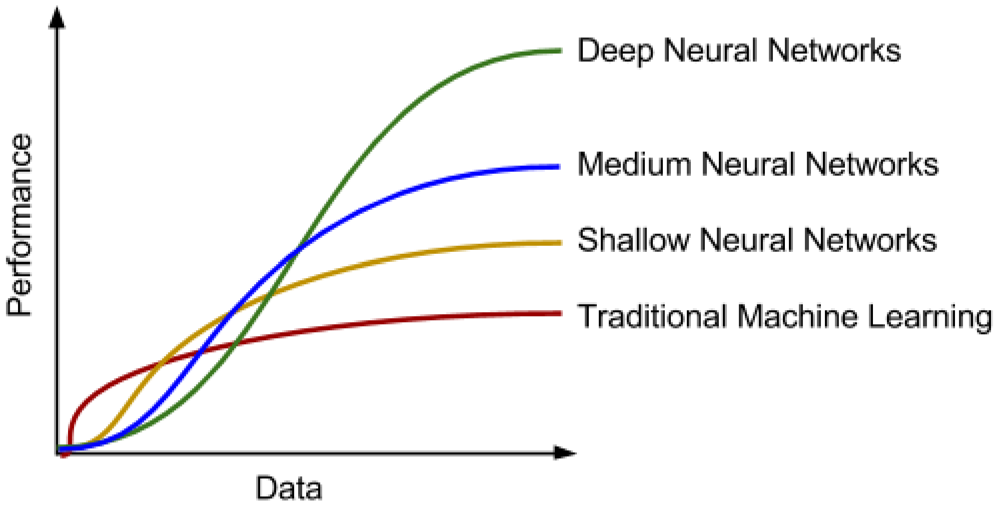
3. Возможность online использования

    сеть можно дообучать

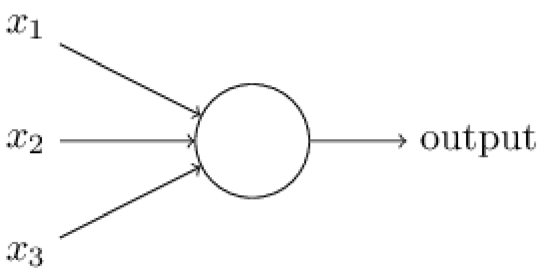
Где применяют нейронные сети?

1. Распознавание изображений, речи, рукописного текста на уровне человека
2. Машинный перевод
3. Перевод текста в речь
4. Электронные помощники
5. Автономное управление автомобилем
6. Таргетированная реклама
7. Персонализированный поиск
8. Семантический поиск
9. Супер игрок в Го

Deep learning это не серебряная пуля!



Что такое однослойный перцептрон?



$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{ threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{ threshold} \end{cases}$$

Состоится ли пикник в эти выходные?

x1: погода (0, 1) w1 = 6 threshold = 7

x2: идут ли друзья (0,1) w2 = 2

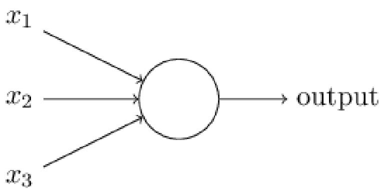
x3: далеко ли ехать (0, 1) w3 = 3

Sigmoid neuron vs. perceptron

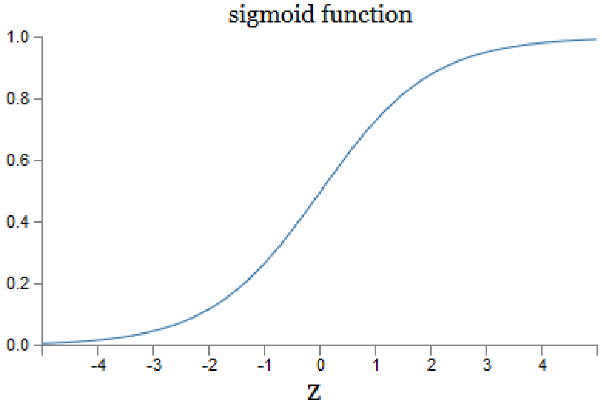
Функция активации:

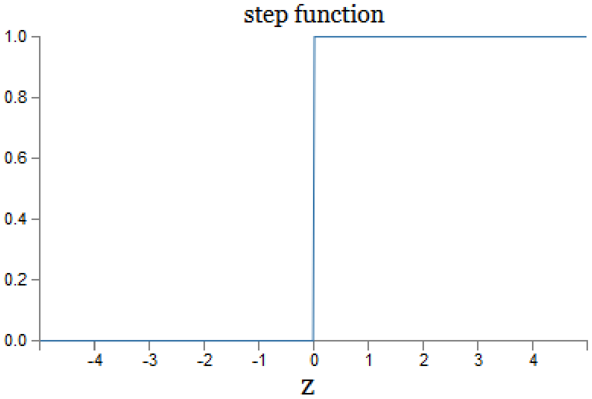
$$\sigma(z) \equiv \frac{1}{1+e^{-z}}$$

$$z \equiv w \cdot x + b$$

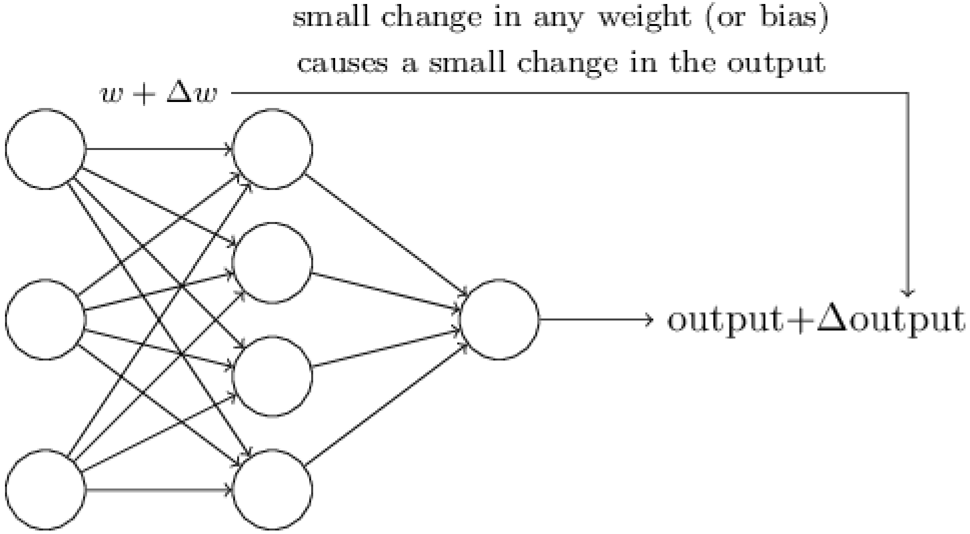


$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{ threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{ threshold} \end{cases}$$

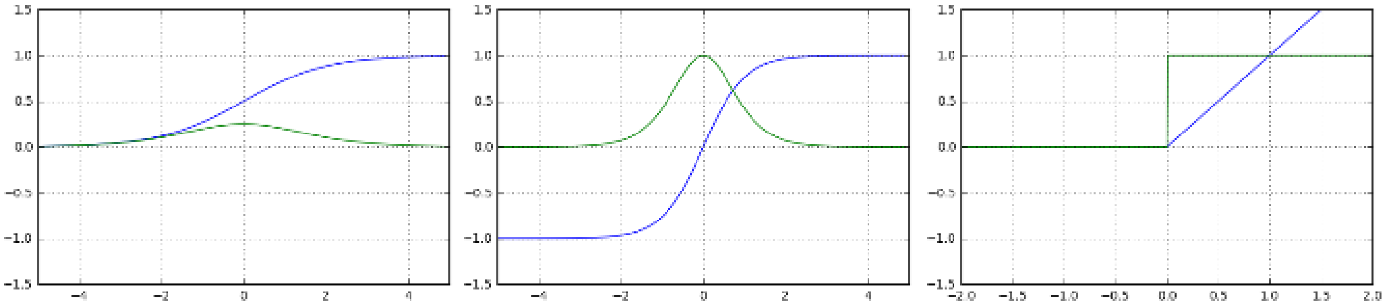




Чем сигмоид лучше перцептрона?



Другие функции активации



$$\mathrm{sigm}(x) = \frac{1}{1+e^{-x}}$$

$$\mathrm{sigm}'(x) = \mathrm{sigm}(x)(1-\mathrm{sigm}(x))$$

$$\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$$

$$\tanh'(x) = 1-\tanh(x)^2$$

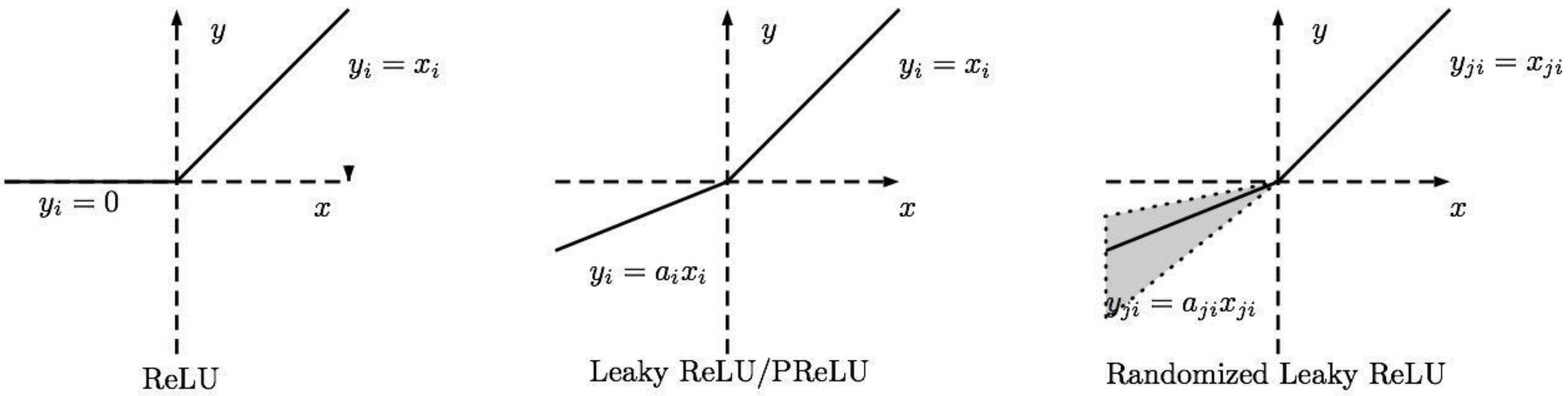
$$\mathrm{relu}(x) = \max(0,x)$$

$$\mathrm{relu}'(x) = 1_{x>0}$$

синий - функция активации

зеленый - ее производная

Модификации ReLU



Softmax function

$$softmax(\mathbf{x}) = \frac{1}{\sum_{i=1}^{n} e^{x_i}} \cdot \begin{bmatrix} e^{x_1} \\ e^{x_2} \\ \vdots \\ e^{x_n} \end{bmatrix}$$

$$\frac{\partial softmax(\mathbf{x})_i}{\partial x_j} = \begin{cases} softmax(\mathbf{x})_i \cdot (1 - softmax(\mathbf{x})_i) & i = j \\ -softmax(\mathbf{x})_i \cdot softmax(\mathbf{x})_j & i \neq j \end{cases}$$

- vector of values in (0, 1) that add up to 1
- $p(Y = c | X = \mathbf{x}) = \mathrm{softmax}(\mathbf{z}((\mathbf{x}))_c$

Softmax function

Многослойная нейронная сеть

входные данные $X \in \mathbb{R}^{n_x \times m}$

целевые значения $Y \in \mathbb{R}^{n_y \times m}$

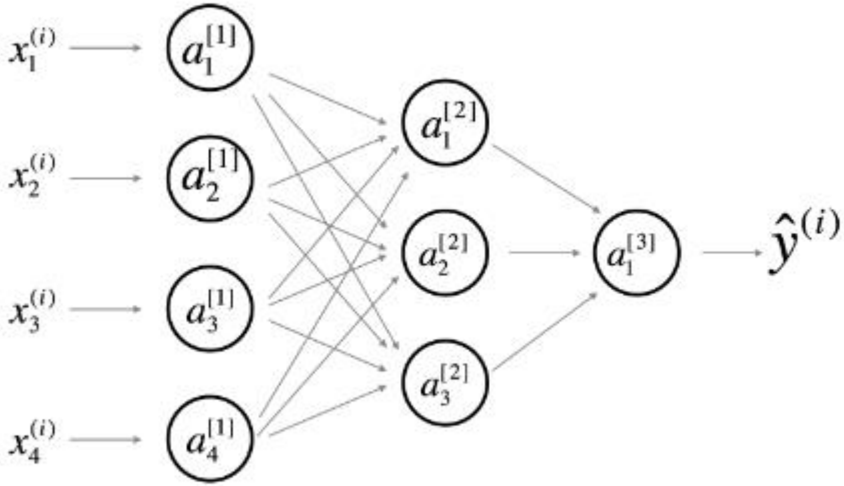
веса на слое [l] $W^{[l]}$

значения на скрытых слоях

$$a = g^{[l]}(W_x x^{(i)} + b_1) = g^{[l]}(z_1)$$

где $g^{[l]}$ - функция активации

предсказанные значения $\hat{y}^{(i)}$

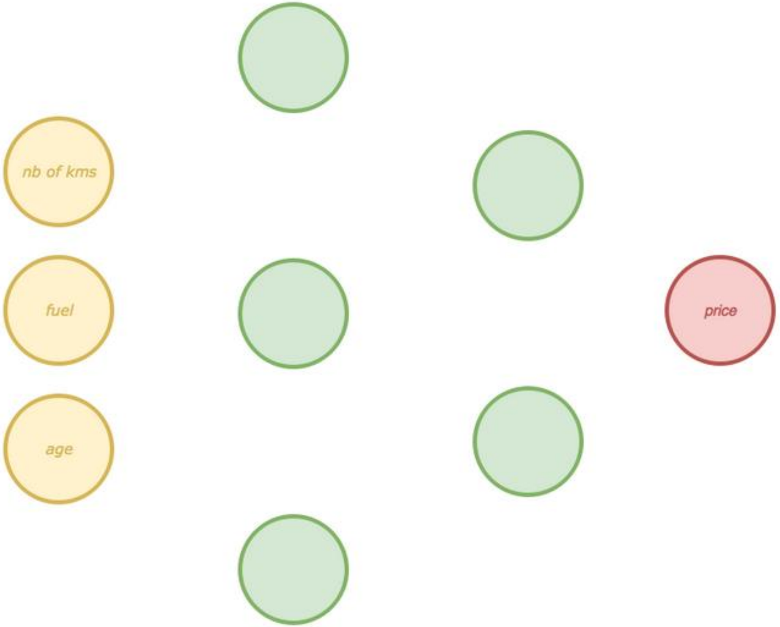


Задача: предсказать стоимость машины

- Number of kilometers: quantitative, number between 0 and 350k.
- Type of fuel: binary data diesel/gasoline.
- Age: quantitative, number between 0 and 40.
- Price: quantitative, number between 0 and 40k.

Number of kms	Fuel	Age	Price
38 000	Gasoline	3	17 000

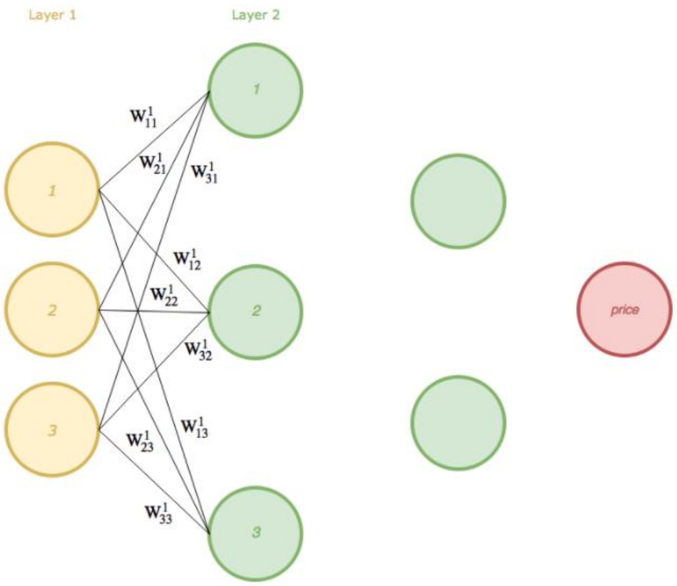
Задача: предсказать стоимость машины



Number of kms	Fuel	Age	Price
1.4	-1	0.4	0.45
0.4	-1	0.1	0.52
5.4	-1	4	0.25
1.5	-1	1	0.31
...	...	...	...

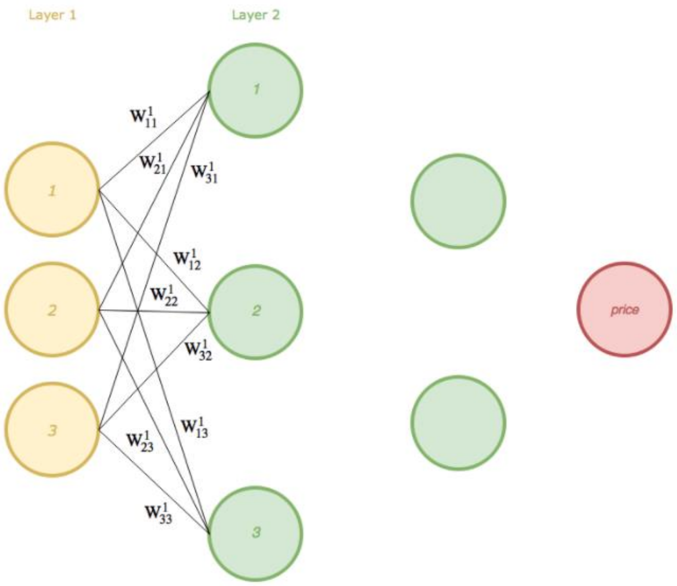
$$X = \begin{bmatrix} 1.4 & -1 & 0.4 & \\ 0.4 & -1 & 0.1 & \\ 5.4 & -1 & 4 & \\ 1.5 & -1 & 1 & \\ \dots & \dots & \dots & \dots \end{bmatrix} \quad y = \begin{bmatrix} 0.45 \\ 0.52 \\ 0.25 \\ 0.31 \\ \dots \end{bmatrix}$$

Forward propagation



$$W^1 = \begin{bmatrix} W_{11}^1 & W_{12}^1 & W_{13}^1 \\ W_{21}^1 & W_{22}^1 & W_{23}^1 \\ W_{31}^1 & W_{32}^1 & W_{33}^1 \end{bmatrix}$$

Forward propagation

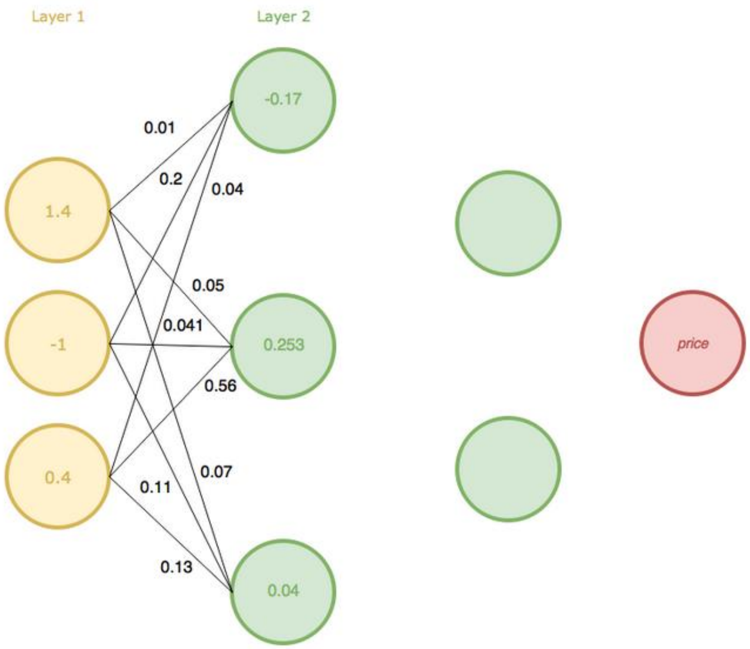


$$X = \begin{bmatrix} 1.4 & -1 & 0.4 \end{bmatrix}$$

$$W^1 = \begin{bmatrix} 0.01 & 0.05 & 0.07 \\ 0.20 & 0.041 & 0.11 \\ 0.04 & 0.56 & 0.13 \end{bmatrix}$$

$$Z^{(2)} = X.W^1$$

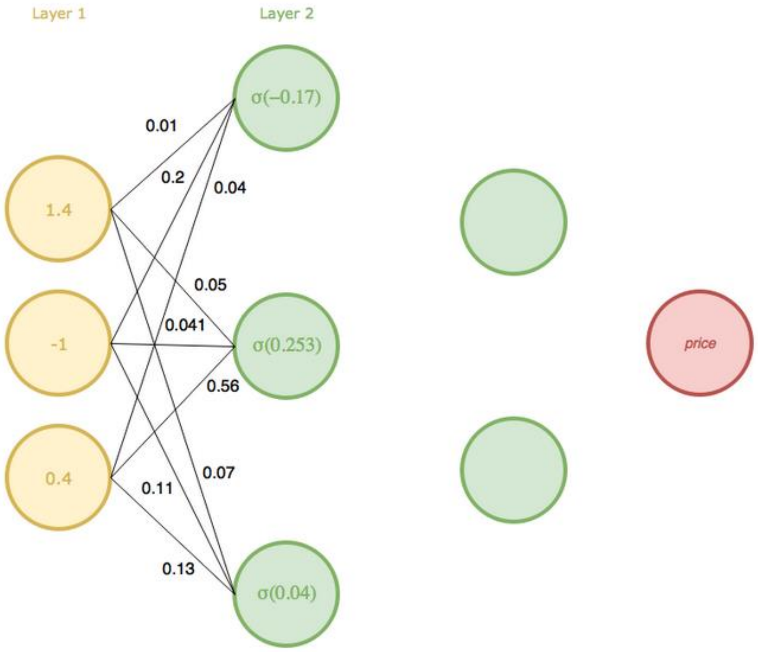
Forward propagation



$$Z^{(2)} = \begin{bmatrix} 1.4 & -1 & 0.4 \end{bmatrix} . \begin{bmatrix} 0.01 & 0.05 & 0.07 \\ 0.20 & 0.041 & 0.11 \\ 0.04 & 0.56 & 0.13 \end{bmatrix}$$

$$Z^{(2)} = \begin{bmatrix} -0.17 & 0.253 & 0.04 \end{bmatrix}$$

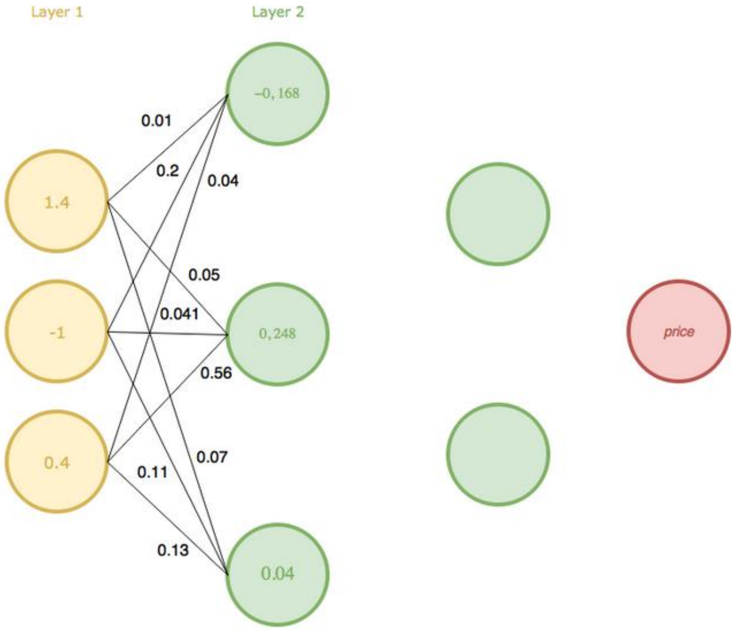
Forward propagation



$$a^{(2)} = \sigma(Z^{(2)})$$

$$\sigma(x) = tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Forward propagation



$$\sigma(-0.17) = tanh(-0.17) = \frac{e^{-0.17} - e^{0.17}}{e^{-0.17} + e^{0.17}} = -0.168$$

$$\sigma(0.253) = tanh(0.253) = \frac{e^{0.253} - e^{-0.253}}{e^{0.253} + e^{-0.253}} = 0,248$$

Forward propagation

$$X = \begin{bmatrix} 1.4 & -1 & 0.4 \\ 0.4 & -1 & 0.1 \\ 5.4 & -1 & 4 \\ 1.5 & -1 & 1 \\ 1.8 & 1 & 1 \end{bmatrix}$$

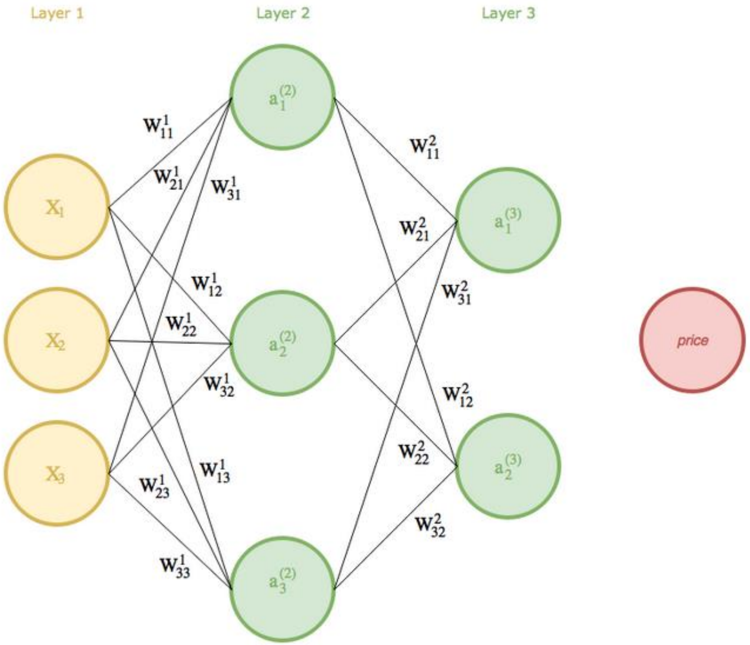
$$Z^{(2)} = \begin{bmatrix} 1.4 & -1 & 0.4 \\ 0.4 & -1 & 0.1 \\ 5.4 & -1 & 4 \\ 1.5 & -1 & 1 \\ 1.8 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.01 & 0.05 & 0.07 \\ 0.20 & 0.041 & 0.11 \\ 0.04 & 0.56 & 0.13 \end{bmatrix}$$

$$Z^{(2)} = \begin{bmatrix} -0.17 & 0.253 & 0.04 \\ -0.192 & 0.035 & -0.069 \\ 0.014 & 2.469 & 0.788 \\ -0.145 & 0.594 & 0.125 \\ 0.258 & 0.691 & 0.366 \end{bmatrix}$$

$$a^{(2)} = \sigma(Z^{(2)}) = tanh(Z^{(2)}) = \begin{bmatrix} tanh(-0.17) & tanh(0.253) & tanh(0.04) \\ tanh(-0.192) & tanh(0.035) & tanh(-0.069) \\ tanh(0.014) & tanh(2.469) & tanh(0.788) \\ tanh(-0.145) & tanh(0.594) & tanh(0.125) \\ tanh(0.258) & tanh(0.691) & tanh(0.366) \end{bmatrix}$$

$$a^{(2)} = \begin{bmatrix} -0.16838105 & 0.24773663 & 0.03997868 \\ -0.18967498 & 0.03498572 & -0.06889071 \\ 0.01399909 & 0.98576421 & 0.65727455 \\ -0.14399227 & 0.53276635 & 0.124353 \\ 0.25242392 & 0.59862403 & 0.35048801 \end{bmatrix}$$

Forward propagation



$$W^2 = \begin{bmatrix} W_{11}^2 & W_{12}^2 \\ W_{21}^2 & W_{22}^2 \\ W_{31}^2 & W_{32}^2 \end{bmatrix} \quad W^2 = \begin{bmatrix} 0.04 & 0.78 \\ 0.40 & 0.45 \\ 0.65 & 0.23 \end{bmatrix}$$

$$Z^{(3)} = a^{(2)}.W^2$$

$$a^{(3)} = tanh(Z^{(3)})$$

Forward propagation

$$Z^{(3)} = a^{(2)}.W^2$$

$$a^{(3)} = tanh(Z^{(3)})$$

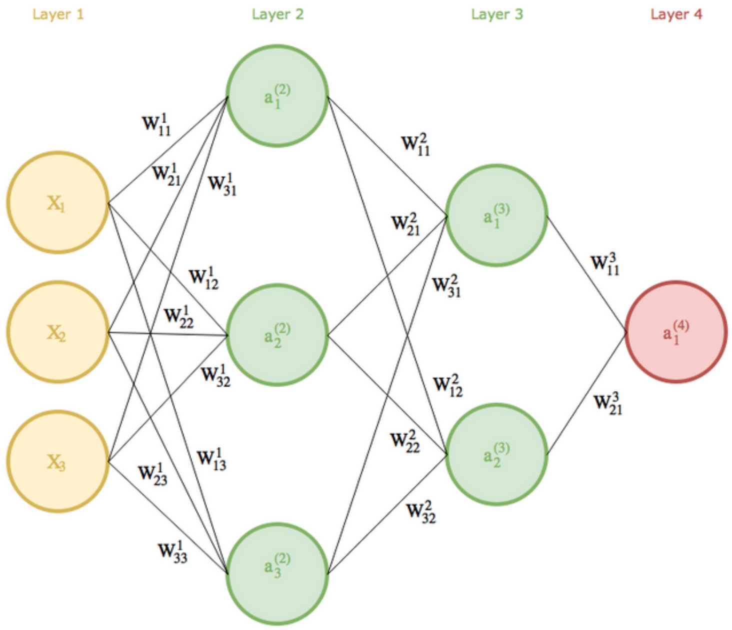
$$Z^{(3)} = \begin{bmatrix} -0.16838105 & 0.24773663 & 0.03997868 \\ -0.18967498 & 0.03498572 & -0.06889071 \\ 0.01399909 & 0.98576421 & 0.65727455 \\ -0.14399227 & 0.53276635 & 0.124353 \\ 0.25242392 & 0.59862403 & 0.35048801 \end{bmatrix} . \begin{bmatrix} 0.04 & 0.78 \\ 0.40 & 0.45 \\ 0.65 & 0.23 \end{bmatrix}$$

$$a^{(3)} = \begin{bmatrix} tanh(0.11834555) & tanh(-0.01066064) \\ tanh(-0.03837167) & tanh(-0.14804778) \\ tanh(0.8220941) & tanh(0.60568633) \\ tanh(0.2881763) & tanh(0.15603208) \\ tanh(0.47736378) & tanh(0.54688371) \end{bmatrix}$$

$$Z^{(3)} = \begin{bmatrix} 0.11834555 & -0.01066064 \\ -0.03837167 & -0.14804778 \\ 0.8220941 & 0.60568633 \\ 0.2881763 & 0.15603208 \\ 0.47736378 & 0.54688371 \end{bmatrix}$$

$$a^{(3)} = \begin{bmatrix} 0.11779613 & -0.01066023 \\ -0.03835285 & -0.14697553 \\ 0.67620804 & 0.54108347 \\ 0.28045542 & 0.15477804 \\ 0.44412987 & 0.49818098 \end{bmatrix}$$

Forward propagation



$$W^3 = \begin{bmatrix} W_{11}^3 \\ W_{21}^3 \end{bmatrix} \qquad W^3 = \begin{bmatrix} 0.04 \\ 0.41 \end{bmatrix}$$

$$Z^{(4)} = a^{(3)} . W^3$$

$$a^{(4)} = tanh(Z^{(4)})$$

Forward propagation

$$Z^{(4)} = a^{(3)}.W^3$$

$$a^{(4)} = tanh(Z^{(4)})$$

$$Z^{(4)} = \begin{bmatrix} 0.11779613 & -0.01066023 \\ -0.03835285 & -0.14697553 \\ 0.67620804 & 0.54108347 \\ 0.28045542 & 0.15477804 \\ 0.44412987 & 0.49818098 \end{bmatrix} \cdot \begin{bmatrix} 0.04 \\ 0.41 \end{bmatrix}$$

$$a^{(4)} = tanh(Z^{(4)}) = \begin{bmatrix} tanh(0.00034115) \\ tanh(-0.06179408) \\ tanh(0.24889254) \\ tanh(0.07467721) \\ tanh(0.22201939) \end{bmatrix}$$

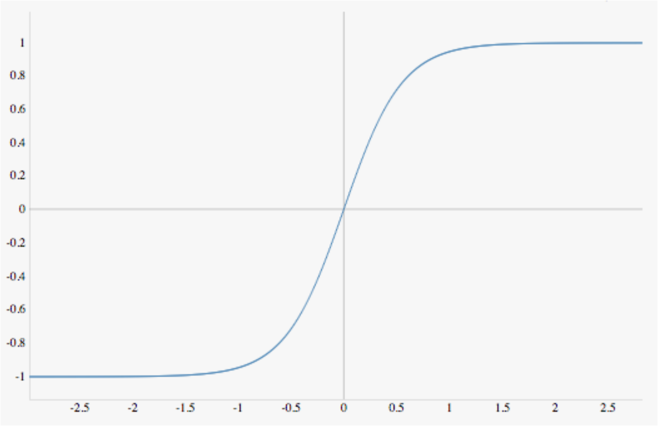
$$Z^{(4)} = \begin{bmatrix} 0.00034115 \\ -0.06179408 \\ 0.24889254 \\ 0.07467721 \\ 0.22201939 \end{bmatrix}$$

$$a^{(4)} = \begin{bmatrix} 0.000341156 \\ -0.0617156 \\ 0.243877 \\ 0.0745387 \\ 0.218442 \end{bmatrix}$$

Forward propagation

Оставим только пробег и поймем нужен ли нам b

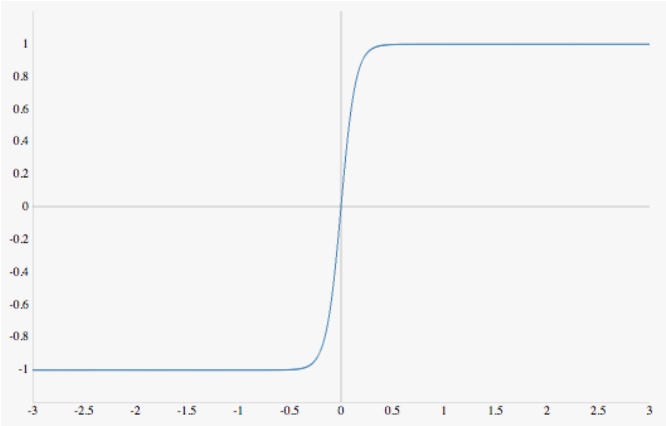
$$a_1^{(2)} = tanh(Z_1^{(2)}) = tanh(X_1 \times W_{11}^1).$$



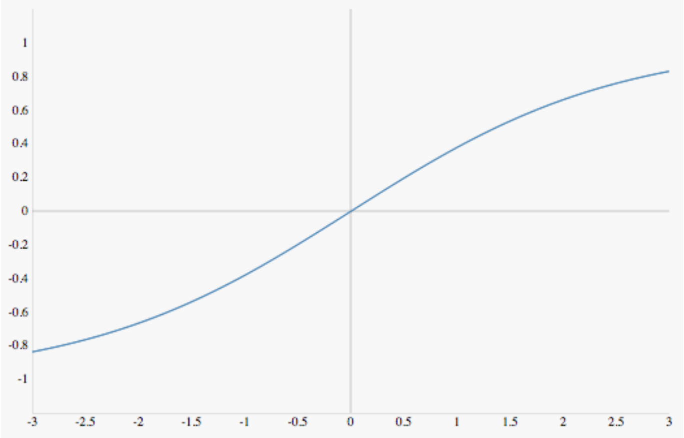
1.8

Forward propagation

Оставим только пробег и поймем нужен ли нам b



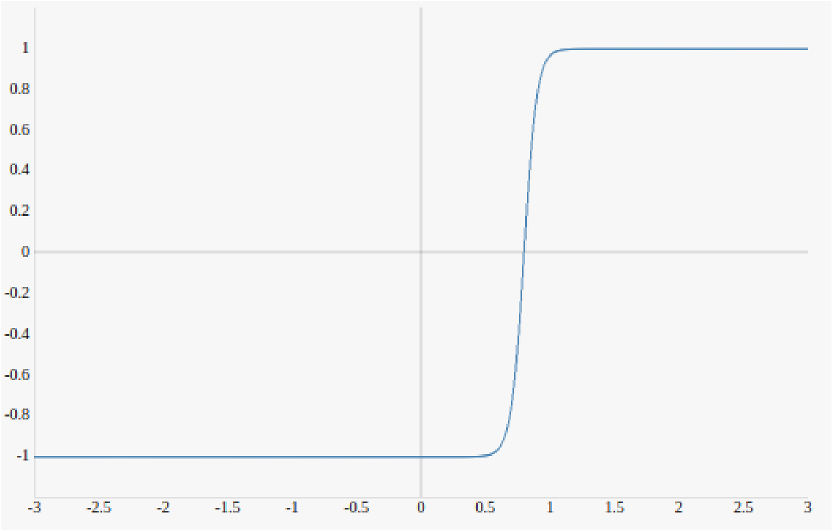
7



0.4

Forward propagation

Резкое изменение цены происходит в 50000км (0.7)



$$a_1^{(2)} = tanh(X_1 \times W_{11}^1 + X_2 \times W_{21}^1 + X_3 \times W_{31}^1 + b)$$

Forward propagation

Будем добавлять приставлением единичного вектора к данным

$$X = \begin{bmatrix} 1.4 & -1 & 0.4 & 1 \\ 0.4 & -1 & 0.1 & 1 \\ 5.4 & -1 & 4 & 1 \\ 1.5 & -1 & 1 & 1 \\ 1.8 & 1 & 1 & 1 \end{bmatrix} \qquad W^1 = \begin{bmatrix} 0.01 & 0.05 & 0.07 \\ 0.20 & 0.041 & 0.11 \\ 0.04 & 0.56 & 0.13 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$$

Forward propagation

$$a_1^{(2)} = \tanh(Z_1^{(2)}) = \tanh(X_1 \times W_{11}^1 + X_2 \times W_{21}^1 + X_3 \times W_{31}^1 + b)$$

$$a_2^{(2)} = \tanh(Z_2^{(2)}) = \tanh(X_1 \times W_{12}^1 + X_2 \times W_{22}^1 + X_3 \times W_{32}^1 + b)$$

$$a_3^{(2)} = \tanh(Z_3^{(2)}) = \tanh(X_1 \times W_{13}^1 + X_2 \times W_{23}^1 + X_3 \times W_{33}^1 + b)$$

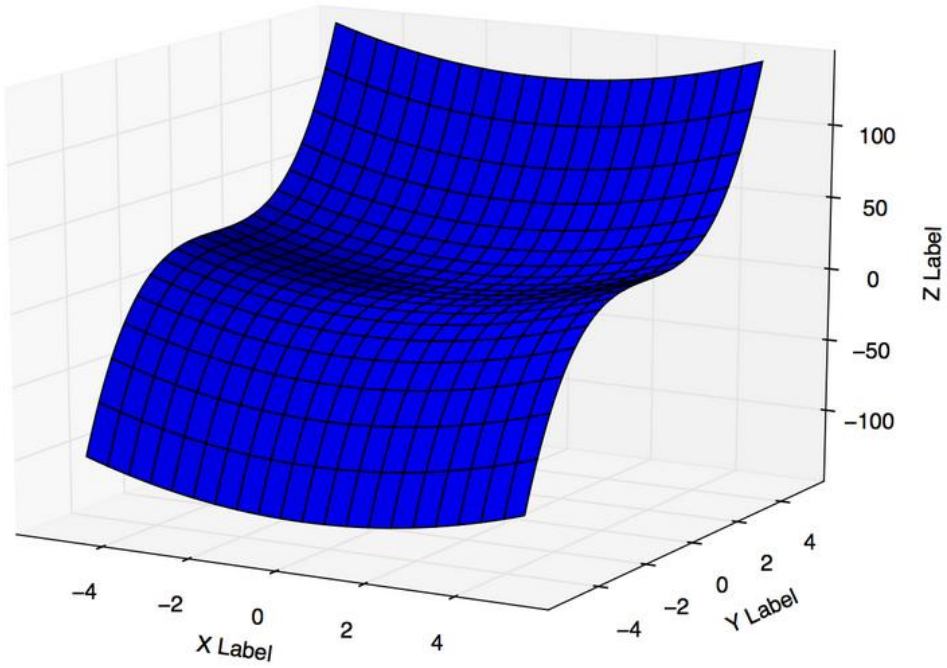
Gradient descent

$$J(W) = \sum_{1}^{n} \frac{1}{2}(y - \hat{y})^2$$

$$J(W) = \tfrac{1}{2}(0.45 - 0.2023543)^2 = 0,031$$

$$J(W) = \sum_{1}^{n} \frac{1}{2}(y - \tanh(\tanh(\tanh(X.W_1).W_2).W_3))^2$$

Gradient descent



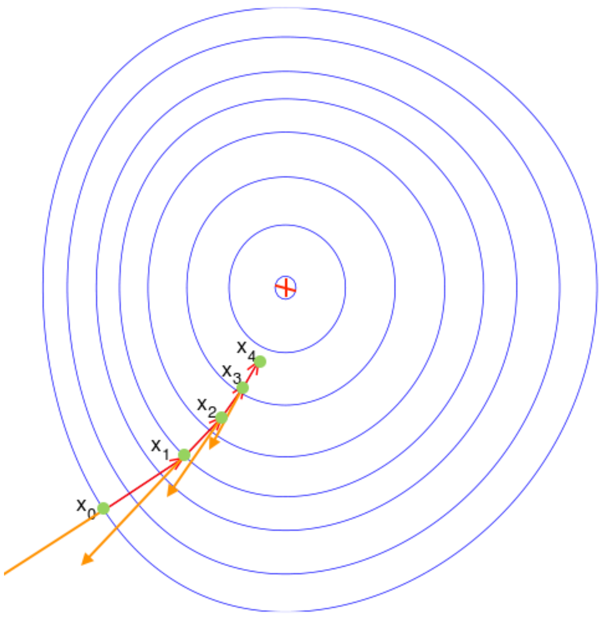
$$z = x^2 + y^3$$

$$F = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right]$$

$$F = [2x, 3y^2]$$

$x = 4$ and $y = 2$

Gradient descent



Backward propagation

$$J(W) = \sum_{1}^{n} \frac{1}{2}(y - \hat{y})^2$$

$$\nabla(J(W)) = \left[\frac{\partial J(W)}{\partial W_1}, \frac{\partial J(W)}{\partial W_2}, \frac{\partial J(W)}{\partial W_3}\right]$$

$$W_1 = W_1 - \alpha \frac{1}{n} \frac{\partial J(W)}{\partial W_1}$$

$$J(W) = \sum_{1}^{n} \frac{1}{2}(y - \tanh(\tanh(\tanh(X.W_1).W_2).W_3))^2$$

Backward propagation: chain rule

$$(f \circ g)' = (f' \circ g).g'$$

For instance if we take the function $f(x) = (2x^2 + 8)^3$ we see a composition. The result of the first function $g(x) = 2x^2 + 8$ is used by the second function $f(g(x)) = (g(x))^3$. The derivative of $g(x)$ is $g'(x) = 4x$ and the derivative of $f(g(x))$ is $f'(g(x)) = 3g(x)^2$. We apply the above formula:

$$f'(x) = f'(g(x)).g'(x) = 3(2x^2 + 8)^2 \cdot 4x$$

Backward propagation: chain rule

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

Then we compute the derivative:

$$\frac{\partial f}{\partial g} = 3g(x)^2$$

$$\frac{\partial g}{\partial x} = 4x$$

$$\frac{\partial f}{\partial x} = 3(2x^2 + 8)^2 \cdot 4x$$

Backward propagation: chain rule

The chain rule can also be written in the following way:

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial x}$$

Meaning that if y depends on x and z depends on y, z also depends on x. If we use our previous example. We want $\frac{\partial f}{\partial x}$. We know that f depends on g and g depends on x because $f(g(x)) = g(x)^3$ and $g(x) = 2x^2 + 8$ so we can write:

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

Backward propagation

Our goal is therefore to find the gradients of our function $J(W)$ and use them to update the weights of our network. Our cost function computes three inputs that are the network weights. We have to find the partial derivatives (gradients) with regards to its weights.

$$\nabla(J(W)) = \left[\frac{\partial J(W)}{\partial W_1}, \frac{\partial J(W)}{\partial W_2}, \frac{\partial J(W)}{\partial W_3}\right]$$

Then we will update the weights using the gradients, for instance W_1 will be updated using the following rule:

$$W_1 = W_1 - \alpha \frac{1}{n} \frac{\partial J(W)}{\partial W_1}$$

Backward propagation

We have:

$$J(W) = \frac{1}{2}(y - \hat{y})^2$$

We can see a first composition. $J(W) = \frac{1}{2}(g(x))^2$ where $g(x) = y - \hat{y}$ so we have:

$$\frac{\partial J(W)}{\partial W_3} = \frac{\partial J(W)}{\partial g} \cdot \frac{\partial g}{\partial W_3}$$

Backward propagation

$$\frac{\partial J(W)}{\partial W_3} = (y - \hat{y}) \cdot -\frac{\partial \hat{y}}{\partial W_3}$$

As we said before, $\hat{y}$ is our predictions, meaning the output of our network, meaning $a^{(4)}$. We know that $\hat{y} = a^{(4)} = tanh(Z^{(4)})$ so our $\hat{y}$ depends on $Z^{(4)}$ and $Z^{(4)}$ depends on W_3. So we can write:

$$\frac{\partial J(W)}{\partial W_3} = (y - \hat{y}) \cdot -\frac{\partial \hat{y}}{\partial Z^{(4)}} \cdot \frac{\partial Z^{(4)}}{\partial W_3}$$

Backward propagation

We can find $\frac{\partial \hat{y}}{\partial Z^{(4)}}$ directly, we have: $\hat{y} = tanh(Z^{(4)})$ so our derivative with regards to $Z^{(4)}$ is the following:

$$\frac{\partial \hat{y}}{\partial Z^{(4)}} = tanh'(Z^{(4)}) = 1 - \tanh(Z^{(4)})^2$$

We can replace its value in our initial formula:

$$\frac{\partial J(W)}{\partial W_3} = (y - \hat{y}) \cdot -(1 - \tanh(Z^{(4)})^2) \cdot \frac{\partial Z^{(4)}}{\partial W_3}$$

Backward propagation

Now we have one final term to compute $\frac{\partial Z^{(4)}}{\partial W_3}$ and we know that $Z^{(4)} = a^{(3)} \cdot W^3$. Finally $Z^{(4)}$ depends on W_3 directly so no more chain rule needed for this first gradient. We keep $a^{(3)}$ as a constant and W^3 becomes one because we are differentiating with regard to W^3. We have:

$$\frac{\partial Z^{(4)}}{\partial W_3} = a^{(3)} \times 1 = a^{(3)}$$

We can replace it in our initial formula:

$$\frac{\partial J(W)}{\partial W_3} = (y - \hat{y}) \cdot -(1 - \tanh(Z^{(4)})^2) \cdot a^{(3)}$$

Backward propagation

We will introduce $\delta^{(4)}$ equals to:

$$\delta^{(4)} = (y - \hat{y}) \cdot -(1 - \tanh(Z^{(4)})^2)$$

So our previous gradient is in fact:

$$\frac{\partial J(W)}{\partial W_3} = \delta^{(4)} \cdot a^{(3)}$$

Backward propagation

Using the exact same steps as for W_3 we will arrive at:

$$\frac{\partial J(W)}{\partial W_2} = (y - \hat{y}) \cdot -(1 - \tanh(Z^{(4)})^2) \cdot \frac{\partial Z^{(4)}}{\partial W_2}$$

You can see above that we have the same term as before, that's why we introduced $\delta^{(4)}$, we can replace it in our formula:

$$\frac{\partial J(W)}{\partial W_2} = \delta^{(4)} \cdot \frac{\partial Z^{(4)}}{\partial W_2}$$

Backward propagation

Before we were searching the derivative of $Z^{(4)}$ with regards to W_3 so the derivative was $a^{(3)} \times 1$, as a reminder $Z^{(4)} = a^{(3)} \cdot W_3$ but this time we are searching the derivative with regards to W_2. W_3 does not depend on W_2 so it becomes a constant, meanwhile $a^{(3)}$ depends on W_2 so we have to find its derivative with regards to W_2. This gives us:

$$\frac{\partial Z^{(4)}}{\partial W_2} = W_3 \cdot \frac{\partial a^{(3)}}{\partial W_2}$$

We can replace it in our original formula:

$$\frac{\partial J(W)}{\partial W_2} = \delta^{(4)} \cdot W_3 \cdot \frac{\partial a^{(3)}}{\partial W_2}$$

Backward propagation

Now we have to compute $\frac{\partial a^{(3)}}{\partial W_2}$, we know that $a^{(3)}$ depends on $z^{(3)}$ (because $a^{(3)} = tanh(z^{(3)})$) which itself depends on W_2 (because $z^{(3)} = a^{(2)} \cdot W_2$). Using the chain rule we can write:

$$\frac{\partial a^{(3)}}{\partial W_2} = \frac{\partial a^{(3)}}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial W_2}$$

We can replace it in our original formula:

$$\frac{\partial J(W)}{\partial W_2} = \delta^{(4)} \cdot W_3 \cdot \frac{\partial a^{(3)}}{\partial z^{(3)}} \cdot \frac{\partial z^{(3)}}{\partial W_2}$$

Backward propagation

We can replace it in our original formula:

$$\frac{\partial J(W)}{\partial W_2} = \delta^{(4)} \cdot W_3 \cdot 1 - \tanh(Z^{(3)})^2 \cdot a^{(2)}$$

We found the second gradient of our function $J(W)$. As you can see, the more you go toward the beginning of the network, the more the differentiation will be long. That's why we introduced the $\delta^{(l)}$ terms where l is the layer number. So that we don't have to differentiate again the first part of the function but directly use $\delta^{(l)}$. For the second gradient we introduce:

$$\delta^{(3)} = \delta^{(4)} \cdot W_3 \cdot 1 - \tanh(Z^{(3)})^2$$

Backward propagation

Using the exact same steps as for W_2 we will arrive at:

$$\frac{\partial J(W)}{\partial W_1} = \delta^{(4)} \cdot W_3 \cdot 1 - \tanh(Z^{(3)})^2 \cdot \frac{\partial z^{(3)}}{\partial W_1}$$

As we introduced $\delta^{(3)}$ we can use it:

$$\frac{\partial J(W)}{\partial W_1} = \delta^{(3)} \cdot \frac{\partial z^{(3)}}{\partial W_1}$$

$$\dots$$

Backward propagation

Before we were searching the derivative of $z^{(3)}$ with regards to W_2 and so the derivative was equal to $a^{(2)}$. As a reminder $z^{(3)} = a^{(2)} \cdot W_2$. This time we are searching the derivative of $z^{(3)}$ with regards to W_1 and so W_2 is only a constant, we have:

$$\frac{\partial z^{(3)}}{\partial W_1} = W_2 \cdot \frac{\partial a^{(2)}}{\partial W_1}$$

We can replace it in our original formula:

$$\frac{\partial J(W)}{\partial W_1} = \delta^{(3)} \cdot W_2 \cdot \frac{\partial a^{(2)}}{\partial W_1}$$

Backward propagation

$$\frac{\partial a^{(2)}}{\partial W_1} = \frac{\partial a^{(2)}}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial W_1}$$

Where:

$$\frac{\partial a^{(2)}}{\partial z^{(2)}} = tanh'(z^{(2)}) = 1 - \tanh(Z^{(2)})^2$$

We replace it in our original formula:

$$\frac{\partial J(W)}{\partial W_1} = \delta^{(3)} \cdot W_2 \cdot 1 - \tanh(Z^{(2)})^2 \cdot \frac{\partial z^{(2)}}{\partial W_1}$$

Backward propagation

We have one last term to differentiate, if you remember $z^{(2)} = X \cdot W_1$ as we differentiate with regards to W_1 we have:

$$\frac{\partial z^{(2)}}{\partial W_1} = X \cdot W_1 = X$$

So our last gradient is:

$$\frac{\partial J(W)}{\partial W_1} = \delta^{(3)} \cdot W_2 \cdot 1 - \tanh(Z^{(2)})^2 \cdot X$$

Backward propagation

We also introduce the term $\delta^{(2)}$, we have:

$$\delta^{(2)} = \delta^{(3)} \cdot W_2 \cdot 1 - \tanh(Z^{(2)})^2$$

And:

$$\frac{\partial J(W)}{\partial W_1} = \delta^{(2)} \cdot X$$

Backward propagation

$$\frac{\partial J(W)}{\partial W_1} = \delta^{(2)} \cdot X$$

$$\delta^{(3)} = \delta^{(4)} \cdot W_3 \cdot 1 - \tanh(Z^{(3)})^2$$

$$\delta^{(2)} = \delta^{(3)} \cdot W_2 \cdot 1 - \tanh(Z^{(2)})^2$$

$$\frac{\partial J(W)}{\partial W_3} = \delta^{(4)} \cdot a^{(3)}$$

$$\frac{\partial J(W)}{\partial W_2} = \delta^{(3)} \cdot a^{(2)}$$

$$\delta^{(4)} = (y - \hat{y}) \cdot -(1 - \tanh(Z^{(4)})^2)$$

Нормализация и инициализация параметров

- Input data should be normalized to have approx. same range:
  - standardization or quantile normalization
- Initializing W^h and W^o:
  - Zero is a saddle point: no gradient, no learning
  - Constant init: hidden units collapse by symmetry
  - Solution: random init, ex: $w \sim \mathcal{N}(0, 0.01)$

- Biases can (should) be initialized to zero