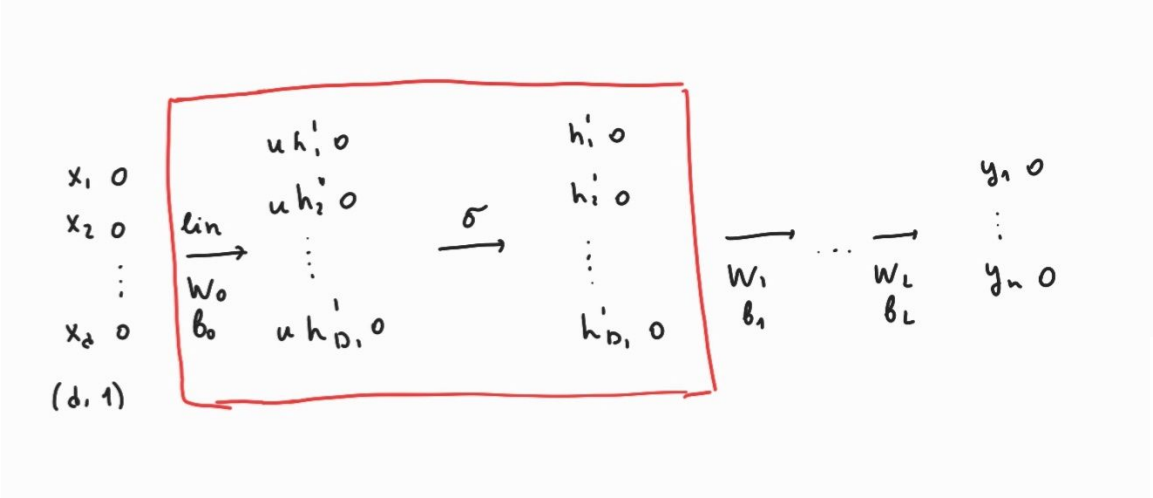
Решение пробной задачи ИАД, 2022

Работу выполнил Данилов Максим, гр. Б05-005

Основные проблемы

- Понять, в чем недостаток основных теорем об аппроксимациях: теорем Колмогорова и Цыбенко
- Понять, почему предпочтительнее не увеличивать число нейронов в скрытом слое сети, а число слоев сети, с точки зрения аппроксимации функции
- На практике проверить, действительно ли увеличение числа слоев работает лучше, чем увеличение числа нейронов в одном слое

Модель простой нейронной сети



$$\text{lin}(x, W_i, b_i) = W^T x + b_i$$

$\sigma(x)$ - функция активации

$$h^1 = \sigma(lin(x, W_0, b_0))$$

$$h^i = \sigma(lin(h^{i-1}, W_{i-1}, b_{i-1})) \text{ для } i = 2...L$$

В качестве σ будем рассматривать:

$$\sigma(x) = \text{sigmoid}(x) = \left(\frac{1}{1 + e^{-x_i}}\right)_{i=1...d}^T$$

$$\sigma(x) = \text{ReLU}(x) = (\max\{0, x_i\})_{i=1...d}^T$$

На изображении представлена сеть, содержащая L-1 скрытый слой, имеющая размерность входного вектора - d и выходного - n

Основные теоремы аппроксимации

Теорема(Колмогоров, 1956): пусть непрерывная функция $f(x)$ задана на $[0,1]^d$, $x = (x_1, ..., x_d)^T$. Тогда существуют непрерывные функции σ_i, g_{ij}, где $i = 1...2d+1$, $j = 1...d$, такие что:

$$f(x) = \sum_{i=1}^{2d+1} \sigma_i \left(\sum_{j=1}^{d} g_{ij}(x_j) \right)$$

Кроме того, g_{ij} не зависят от f.

Торема(Цыбенко, 1989): пусть σ - любая непрерывная сигмоидная функция. Тогда для любой непрерывной функции $f(x)$, заданной на $[0,1]^d$ и любого $\varepsilon > 0$ существуют $N > 0$, векторы $\{w^i\}_{i=1...N}$ $(w^i \in \mathbb{R}^d)$, $\theta = (\theta_1, ...\theta_N)^T \in \mathbb{R}^N$, $\alpha = (\alpha_1, ..., \alpha_N)^T \in \mathbb{R}^N$ такие, что:

$$|\sum_{i=1}^{N} \alpha_i \sigma(w^{iT} x + \theta_i)| - f(x)| < \varepsilon$$

Заметим, что $\sum_{i=1}^{N} \alpha_i \sigma(w^{iT} x + \theta_i) = W_1^T \sigma(W_0^T x + \theta) = \text{lin}(\sigma(\text{lin}(x, W_0, \theta)), W_1, 0)$, где $W_0 = (w^1, ..., w^N)$, $W_1 = \alpha$, это как следует из опредления - неронная сеть с одним скрытым слоем.

Недостатки

- Теорема Колмогорова дает оценку количества нейронов в скрытом слое, однако не дает способ получения б и g, и, в общем случае, их сложно найти и они имеют сложный вид
- Теорема Цыбенко фиксирует все функции и, по сути, задает обычную нейронную сеть с одним скрытым слоем. Однако, теорема не говорит, сколько нейронов нужно.

Таким образом, теоремы утверждают, что нейронной сетью с одним открытым слоем можно сколь угодно приблизить любую непрерывную функцию. Однако не дают точного способа. Поэтому возникает вопрос, можно ли обойтись только такими сетями.

Нижние оценки

Зафиксируем $\sigma(x) = \mathrm{ReLU}(x)$.

Лемма(Zhang, 2018): Пусть нейронная сеть с L скрытыми слоями и N нейронами задает функцию $f : \mathbb{R}^d \longrightarrow \mathbb{R}^D$. Тогда область определения функции можно разбить на K выпуклых множеств $\{S^i\} \in \mathbb{R}^d$, в которых сужение $f_{S^i} : S^i \longrightarrow \mathbb{R}^D$ является линейной функции, причем:

$$K \leq \left(e\frac{N}{dL} + e\right)^{dL}$$

$m-$сильно выпуклой назовем функцию f, такую что для любых x, y

$$f(y) \geq f(y) + \nabla f(x)(y - x) + \frac{m}{2}||y - x||^2$$

Теорема: рассмотрим функцию $f : \mathbb{R} \longrightarrow \mathbb{R}$. Тогда пусть $\widetilde{f}$ задается сетью с одним скрытым слоем и N нейронами. Тогда $||f - \widetilde{f}|| = \sup_{|x|<1}|f(x) - \widetilde{f}| \geq O\left(\frac{1}{N^2}\right)$. Причем константа зависит только от d и множества, на котором определена f.

Доказательство теоремы

По лемме, можно разбить область определения S на $K \leq (eN + e)$ выпуклых множеств $\{S^i\}$, на каждом из которых $\widetilde{f}_{S^i}$ - линейная. Тогда найдется k, что $A = S^k$ и $m(A) \geq \frac{m(S)}{K} \geq \frac{m(S)}{eN+e}$. Из определения $||.||$ имеем: $||f - \widetilde{f}|| \geq ||f_A - \widetilde{f}_A|| = \varepsilon$.

Рассмотрим $g = f_A - \widetilde{f}_A$. Рассмотрим $z = \text{argmin}_{x \in A}(g(x))$. Так как g - разность $m-$сильно выпуклой и линейной функции, то g - m-сильно выпуклая функция. Кроме того, $g'(z) = 0$, значит, для любого y имеем: $\Delta g \geq \frac{m}{2}\Delta z^2$. На A $g(x) \in [-\varepsilon, \varepsilon]$. Тогда имеем, что A лежит в $B_r(z)$, где $r = \sqrt{\frac{4\varepsilon}{m}}$. Тогда $m(B_r(z)) = 2r \geq m(A) \geq \frac{m(S)}{eN+e}$, т.е. $\varepsilon \geq \frac{1}{16}\frac{m(S)}{(eN+e)^2} = O(\frac{1}{N^2})$.

Многослойные сети

Теорема (Ханин, 2017): пусть $f : [0,1]^d \longrightarrow \mathbb{R}$ - выпуклая функция и Липшицева с константой L. Тогда существует константа $C > 0$, такая что для любой m существует нейронная сеть N глубины m и со скрытыми слоями размера $d + 1$, такая что:

$$||f - f_N|| \leq CLd^{\frac{3}{2}}m^{-\frac{2}{d}}$$

Недостаток этих теорем

Теорема Ханина, как и теорема Цыбенко, говорит только существовании нейронной сети с заданной точностью, однако, она позволяет оценить глубину сети.

Глубокие сети обучать сложнее, для быстрой сходимости нужно использовать дополнительные приемы, в своей реализации я сделал батч-нормализацию после каждой функции активации.