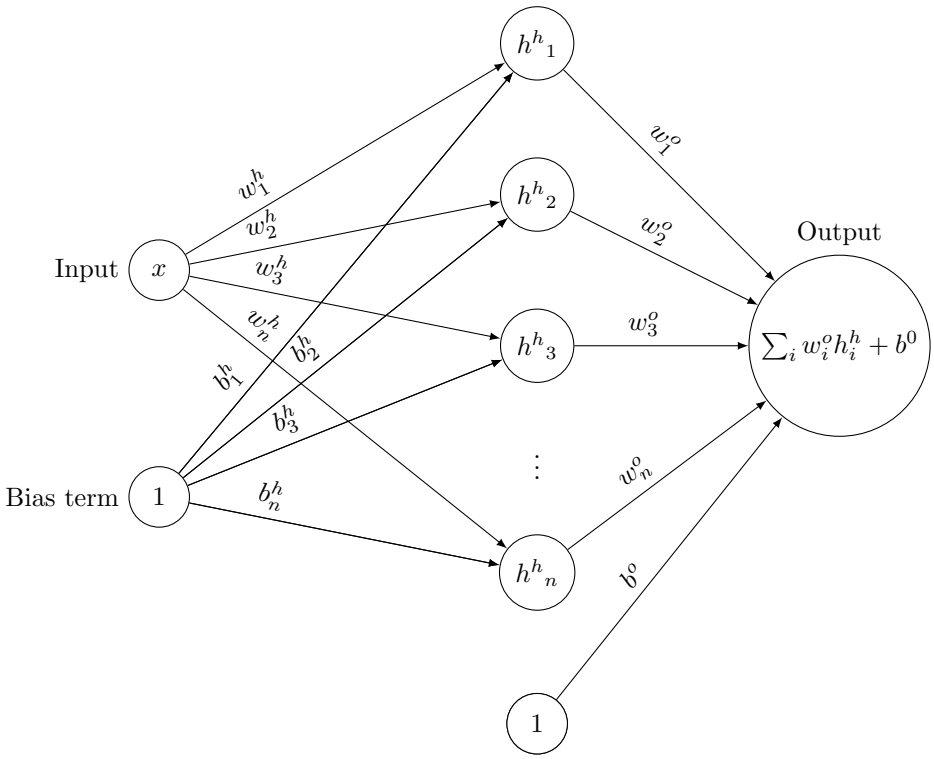
Sieci neuronowe są uniwersalnymi aproksymatorami

Rafał Skrzypiec

Co jeszcze?

- cytowania
- definicje i twierdzenia z polskiej literatury
- podpisy obrazków po polsku?



1 Uniwersalne twierdzenie aproksymacyjne

Według uniwersalnego twierdzenia aproksymacyjnego jednokierunkowa sieć neuronowa z jedną warstwą ukrytą i skończoną ale wystarczająco dużą liczbą neuronów, może przybliżyć z dowolną dokładnością każdą funkcję.

W 1989 roku Cybenko [cytowanie] udowodnił uniwersalne twierdzenie aproksymacyjne dla jednokierunkowej sieci neuronowej z sigmoidalną funkcją aktywacji. Jeszcze w tym samym roku, po pracy Cybenki ukazała się praca Hornika, Stinchcombe'a and White'a, którzy udowodnili prawdziwość powyższego twierdzenia dla dowolnej funkcji aktywacji.

Funkcje sigmoidalne to rodzina funkcji szeroko stosowanych w jednokierunkowych sieciach neuronowych, szczególnie tych stworzonych do celów regresji. W tej części zaprezentuję dowód uniwersalnego twierdzenia aproksymacyjnego podany przez Cybenkę w 1989 roku, następnie zademonstruję dowód wizualny posługując się sigmoidą jako funkcją aktywacji.

1.1 Przybliżenie przez kombinację liniową funkcji sigmoidalnych

Niech I_n oznacza n-wymiarową jednostkową kostkę , $[0,1]^n$. $C(I_n)$ to przestrzeń ciągłych funkjci na I_n. Dodatkowo, niech $M(I_n)$ oznacza przestrzeń skończonych, regularnych miar borelowskich na n-wymiarowej kostce jednostkowej I_n.

Definicja 1.1. Miara μ jest regularna jeśli dla każdego mierzelnego zbioru A, $\mu(A)$ równa się supremum miar zamkniętych podzbiorów A i infimum otwartych nadzbiorów A. [Probability measures on metric spaces K.R. Parthasarathy]

Definicja 1.2. Funkcja $\sigma : \mathbb{R} \to \mathbb{R}$ jest funkcją sigmoidalną jeśli

$$\sigma(x) \to \begin{cases} 1 & \text{as} \quad x \to +\infty \\ 0 & \text{as} \quad x \to -\infty \end{cases}$$

Definicja 1.3. Funkcja σ jest funkcją dyskryminaczyjną jeśli dla miary $\mu \in M(I_n)$ zachodzi

$$\int_{I_n} \sigma\left(w^{\mathsf{T}} x + b_0\right) d\mu(x) = 0 \tag{1}$$

dla każdego $w \in \mathbb{R}$ i $b_0 \in \mathbb{R}$ co implikuje, że $\mu = 0$.

Twierdzenie 1.1. *Każda ograniczona, mierzalna funkcja sigmoidalna σ jest funkcją dyskryminacyjną. W szczególności każda ciągła funkcja sigmoidalna jest dyskryminacyjna. [cytowanie Cybenko]*

Dowód uniwesalnego twierdzenia aproksymacyjnego przy wykorzystaniu funkcji sigmoidalnych wymaga wprowadzenia kilku przydatnych definicji i twierdzeń. Pierwsze z nich to twierdzenie Hahna-Banacha, które formułuje możliwość rozszerzenia każdego ograniczonego funkcjonału liniowego z podprzestrzeni unormowanej na całą podprzestrzeń, przy zachowaniu jego właściwości.

Twierdzenie 1.2 (Twierdzenie Hahna-Banacha)**.** *Niech X to rzeczywista przestrzeń wektorowa, p to funkcja rzeczywista zdefiniowana na X spełniająca*

$$p\left(\alpha x + (1-\alpha)y\right) \leqslant \alpha p(x) + (1-\alpha)p(y) \qquad \forall \alpha \in [0,1], x, y \in X$$

Przypuśmy, że λ to funkcjonał liniowy zdefiniowany na zbiorze $Y \subset X$, który spełnia

$$\lambda(x) \leqslant p(x) \qquad \forall x \in Y.$$

Wtedy istnieje funkcjonał liniowy Λ zdefiniowany na X spełniający

$$\Lambda(x) \leqslant p(x) \qquad \forall x \in X,$$

tak, że

$$\Lambda(x) = \lambda(x) \qquad \forall x \in Y.$$

Reed & Simon (1980), Methods of Modern Mathematical Physics. Functional Analysis

Definicja 1.4. Przestrzeń $\mathcal{L}(\mathcal{H}, \mathbb{C})$ nazywana jest przestrzenią dualną przestrzeni Hilberta $\mathcal{H}$ i oznaczamy ją przez $\mathcal{H}^*$. Elementy $\mathcal{H}^*$ nazywane są ciągłymi funkcjonałami liniowymi.

Reed & Simon (1980), Methods of Modern Mathematical Physics. Functional Analysis

Twierdzenie Riesza opisuje przestrzeń $\mathcal{H}^*$.

Twierdzenie 1.3 (Twierdzenie Riesza (znaleźć polskie źródło))**.** *Dla każdego $T \in \mathcal{H}^*$, istnieje unikalne $y_T \in \mathcal{H}$ takie, że*

$$T(x) = \langle y_T, x \rangle \qquad \forall x \in \mathcal{H}$$

Ponadto

$$||y_T||_{\mathcal{H}} = ||T||_{\mathcal{H}^*}$$

Reed & Simon (1980), Methods of Modern Mathematical Physics. Functional Analysis

Twierdzenie 1.4. *Niech σ będzie ciągłą funkcją dyskryminacyjną, wtedy skończona suma*

$$G(x) = \sum_{i=1}^{N} w_i^o \sigma\left({w_i^h}^{\mathsf{T}} x + b_i^h\right) \tag{2}$$

jest gęsta w $C(I_n)$. Innymi słowy, dla danej funkcji $f \in C(I_n)$ i $\epsilon > 0$, istnieje suma

$G(x)$ mająca powyższą postać, dla której

$$|G(x) - f(x)| < \epsilon \qquad \forall x \in I_n$$

Dowód. Niech $S \subset C(I_n)$ będzie zbiorem funkcji w postaci $G(x)$ lub w innych słowach - zbiorem sieci neuronowych. Z pewnością S jest podprzestrzenią liniową $C(I_n)$. Jeśli S jest gęsty, domknięcie S jest całą przestrzenią $C(I_n)$.

Przyjmijmy, że domknięcie S nie jest całą przestrzenią $C(I_n)$. Wtedy domknięcie $S - S'$ jest domkniętą podprzestrzenią $C(I_n)$. Przez twierdzenie Hahna-Banacha, istnieje ograniczony funkcjonał liniowy na $C(I_n)$, nazwijmy go L, z własnością, że $L \neq 0$ ale $L(S') = L(S) = 0$.

Przez twierdzenie Riesza, ograniczony funkcjonał liniowy L ma postać

$$L(h) = \int_{I_n} h(x) d\mu(x)$$

dla $\mu \in M(I_n)$, dla każdego $h \in C(I_n)$. W szczególności, odkąd $\sigma(w^\mathsf{T} x + b)$ $\in S'$ dla każdego w i b, musi zachodzić

$$\int_{I_n} \sigma\left(w^\mathsf{T} x + b\right) d\mu(x) = 0$$

Jednakże, założyliśmy, że σ jest funkcją dyskryminacyjną, ten warunek implikuje, że $\mu = 0$ co jest sprzeczne z naszym założeniem. Stąd, podprzestrzeń S jest gęsta w $C(I_n)$.

Pokazuje to, że suma

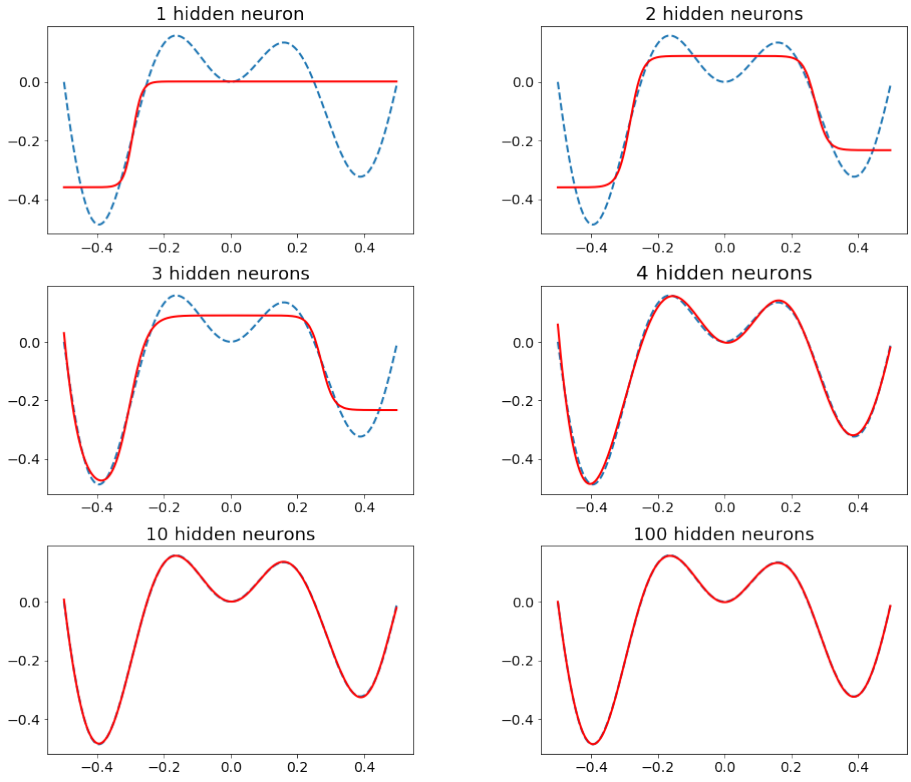
$$G\left(x\right) = \sum_{i=1}^{N} w_i^o \sigma\left({w_i^h}^\mathsf{T} x + b_i^h\right)$$

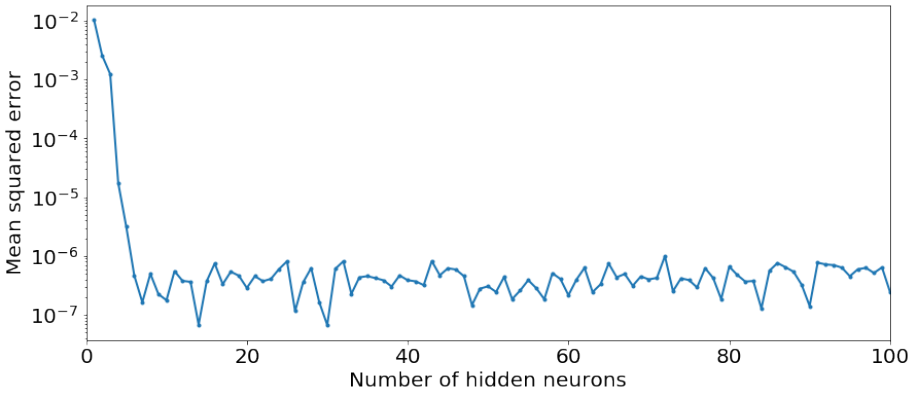
jest gęsta w $C(I_n)$ pod warunkiem, że σ jest ciągła i dyskryminacyjna.

Z twierdzenia wynika, że każda sieć neuronowa o wystarczająco dużej liczbie neuronów w jednej warstwie ukrytej i sigmoidalną funkcją aktywacyjną może z dowolną dokładnością przybliżyć przebieg każdej funkcji.

□

2 Dowód wizualny





2.1 Dane

Zbiór danych treningowych zawiera m jednowymiarowych próbek zadanych przez wektory $X \in \mathbb{R}^{1\times m}$ i odpowiadające im wyniki $Y \in \mathbb{R}^{1\times m}$.

2.2 Parametry

Sieć ma dwie warstwy: 1) ukryta, zawierająca L neuronów i 2) wyjściowa, składająca się z 1 neuronu. Warstwy są zdefiniowane przez:

1. parametry wartstwy ukrytej, które odwzorowują 1-wymiarowe wektory wejściowe w aktywacje L neuronów: macierz wag $W^h \in \mathbb{R}^{L\times 1}$ i wektor parametru bias $b^h \in \mathbb{R}^{L\times 1}$,

2. parametry wartstwy wyjściowe, które odwzorowują L-wymiarowy wektor aktywacji neuronów ukrytych w jeden neuron wartstwy wyjściowej: macierz wag $W^o \in 1 \times L$ i wektor bias $b^o \in \mathbb{R}^{1\times 1}$.

2.3 Propagacja sygnału

Wejście każdego neuronu w warstwie ukrytej jest iloczynem danych wejściowych i odpowiadającej im wagi plus parametr bias. Na przykład dla i-tego przykładu danych wejściowych, w l-tym neuronie mamy

$$a^{h}{}_{l}^{(i)} = W^{h}{}_{l} x^{(i)} + b^{h}{}_{l} \tag{3}$$

Funkcją aktywacyjną neuronów jest sigmoida $\sigma(a) = \frac{1}{1+e^{-a}}$, jako argument przyjmuje ona wejście neuronów:

$$h^{h}{}_{l}^{(i)} = \sigma(a^{h}{}_{l}^{(i)}) \tag{4}$$

Neuron warstwy wyjściowej zawiera sumę iloczynów aktywacji neuronów i odpowiadających im wag plus parametr bias. Dla i-tego przykładu mamy

$$\begin{aligned} a^{o(i)} &= \sum_l W^{o}{}_{l} h^{h}{}_{l}^{(i)} + b^o \\ &= \sum_l W^{o}{}_{l} \sigma(a^{h}{}_{l}^{(i)}) + b^o \\ &= \sum_l W^{o}{}_{l} \sigma(W^{h}{}_{l} x^{(i)} + b^{h}{}_{l}) + b^o \end{aligned} \tag{5}$$

Jako funkcja straty zostanie wykorzystany błąd średniokwadratowy

$$\begin{aligned} J^{(i)}(\Theta) &= \frac{1}{2}\left(y^{(i)} - a^{o(i)}\right)^2 \\ J(\Theta) &= \frac{1}{m}\sum_{i=1}^{m} J^{(i)}(\Theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(y^{(i)} - a^{o(i)}\right)^2. \end{aligned} \tag{6}$$

2.4 Propagacja wsteczna

Użycie reguły łańcuchowej umożliwia obliczenie gradientu funkcji straty względem parametrów sieci neuronowej.

Na początku policzmy gradient względem wyniku wartstwy wyjściowej.

$$\frac{\partial J}{\partial a^{o(i)}} = \frac{1}{m}\left(y^{(i)} - a^{o(i)}\right), \tag{7}$$

następnie policzmy gradient wyjścia neuronów ukrytych:

$$\frac{\partial J}{\partial h^{h}_{l}{}^{(i)}} = \frac{\partial J}{\partial a^{o(i)}}\frac{\partial a^{o(i)}}{\partial h^{h}_{l}{}^{(i)}} = \frac{\partial J}{\partial a^{o(i)}} W^{o}{}_{l}, \tag{8}$$

co umożliwia obliczenie gradientu względem wejścia neruonów ukrytych:

$$\frac{\partial J}{\partial a^{h}_{l}{}^{(i)}} = \frac{\partial J}{\partial h^{h}_{l}{}^{(i)}}\frac{\partial h^{h}_{l}{}^{(i)}}{\partial a^{h}_{l}{}^{(i)}} = \frac{\partial J}{\partial h^{h}_{l}{}^{(i)}} h^{h}_{l}{}^{(i)}(1 - h^{h}_{l}{}^{(i)}) \tag{9}$$

gdzie została wykorzystana relacja

$$\frac{\partial \sigma(x)}{\partial x} = \sigma(x)(1 - \sigma(x)).$$

Ostatecznie możemy policzyć gradienty względem parametrów sieci, np. dla warstwy wejściowej:

$$\frac{\partial J}{\partial W^{o}{}_{l}} = \sum_{i} \frac{\partial J}{\partial a^{o(i)}}\frac{\partial a^{o(i)}}{\partial W^{o}{}_{l}} = \sum_{i} \frac{\partial J}{\partial a^{o(i)}} h^{h}_{l}{}^{(i)}, \tag{10}$$

$$\frac{\partial J}{\partial b^{o}} = \sum_{i} \frac{\partial J}{\partial a^{o(i)}}\frac{\partial a^{o(i)}}{\partial b^{o}} = \sum_{i} \frac{\partial J}{\partial a^{o(i)}}. \tag{11}$$