Activation functions

Rafał Skrzypiec

1 Funkcje aktywacji

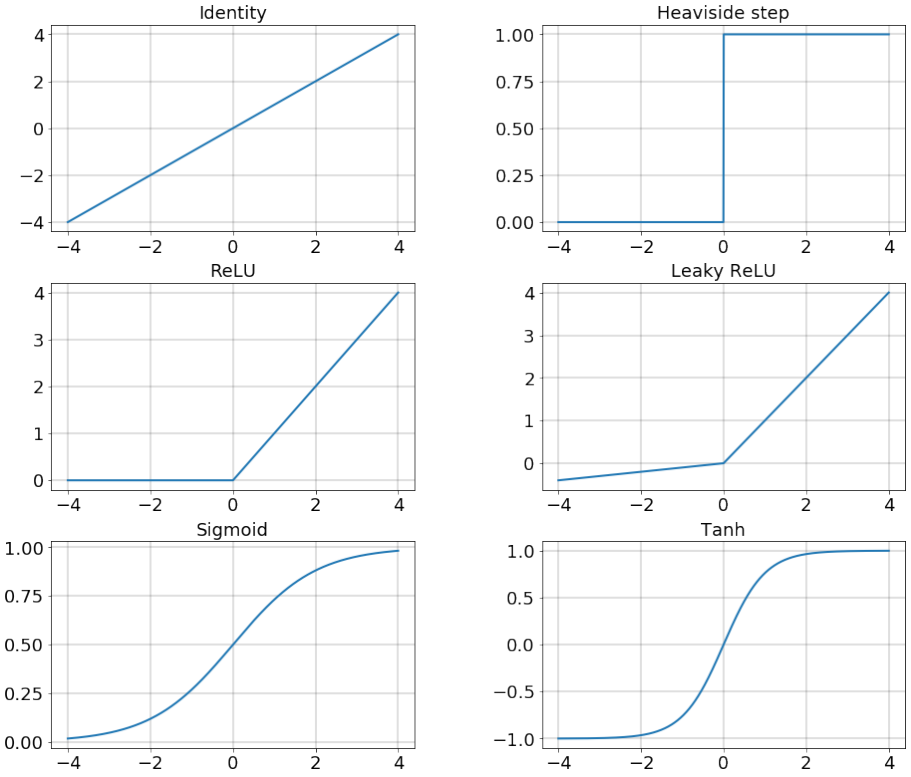
Funckja aktywacji to funkcja, która działa na każdy neuron w sieci neuronowej, jako argument przyjmuje sumę iloczynów wartości neuronów z warstwy poprzedzającej i odpowiadających im wag. Każda z warstw sieci neuronowej może mieć zdefiniowaną inną funckję aktywacji.

Perceptron, który był inspiracją powstania sieci neuronowcyh został skonstruowany jako uproszczony model biologicznego neuronu. W neurobiologii, neuron jest komórką, która odbiera, przetwarza i przesyła informacje wykorzystując elektryczne i chemiczne sygnały. Neurony połączone są ze sobą przez synapsy, jeden neuron może otrzymywać informacje od wielu komórek nerwowych. Jeśli suma sygnałów elektrycznych z wejściowych synaps przekroczy pewien próg, wtedy neuron transmituje dalej sygnał elektryczny. Perceptron naśladował ten mechanizm stosując przedstawioną w lewym górnym rogu na Rys. 1 funckję Heaviside'a jako funkcję aktywacji. Funkcja przyjmuję wartość jeden jeśli suma wartości wejściowych jest większa od zera, w innym przypadku funkcja przyjmuje wartość zero i neuron nie propaguje sygnału. Perceptron jest najprostszym przykładem sieci neuronowej.

Wyniki badań przeprowadzonych przez [publikacja, publikacja] pokazały, że wśród pożądanych cech funkcji aktywacji znajdują się atrybuty, których funkcja Heaviside'a nie posiada, z tego powodu nie jest w praktyce często stosowana.

Koniecznym wymaganiem jest nieliniowość stosowanej funkcji, jest to cecha, która pozwala sieci neuronowej odwzorować nieliniowe zależności [LeCun, Cybenko?, Hornik]. Jedynym wyjątkiem od reguły jest stosowanie w problemach regresyjnych funkcji tożsamościowej w ostatniej warstwie wyjściowej. Dobrze gdy funkcja posiada ciągłą pochodną, pozwala to na stosowanie metod optymizacji opartych o obliczanie gradientu. Tu wyjątkiem jest stosowana poprawiona jednostka liniowa (ReLU), również przedstawiona na Rys. 1. Zakładając, że w zerze jej gradient równy jest zero możemy skorzystać z jej wielu zalet. Wśród nich wymienia się dokładniejsze odwzorowanie obserwowanego w neurobiologii zjawiska – tylko neurony, które otrzymały odpowiednio silny sygnał są aktywowane. Brak podatności na przeuczenie, podczas inicjalizacji sieci losowymi wagami, tylko około 50% ukrytych neuronów jest aktywowanych. Brak problemu znikającego gradientu uniemożliwiającego uczenie, w porównaniu do sigmoidy, u

której wysyca się on w obu kierunkach. Jest to również funkcja często wykorzystywana w metodach głębokiego uczenia. W warstwach splotowych sieci która służy do rozpoznawania obrazów wykorzystamy ReLU poszukując atrybutów, które nie zmieniają się podczas jej użycia.



Rysunek 1: Kilka przykładów często stosowanych funkcji aktywacji.

1.1 Funkcje sigmoidalne

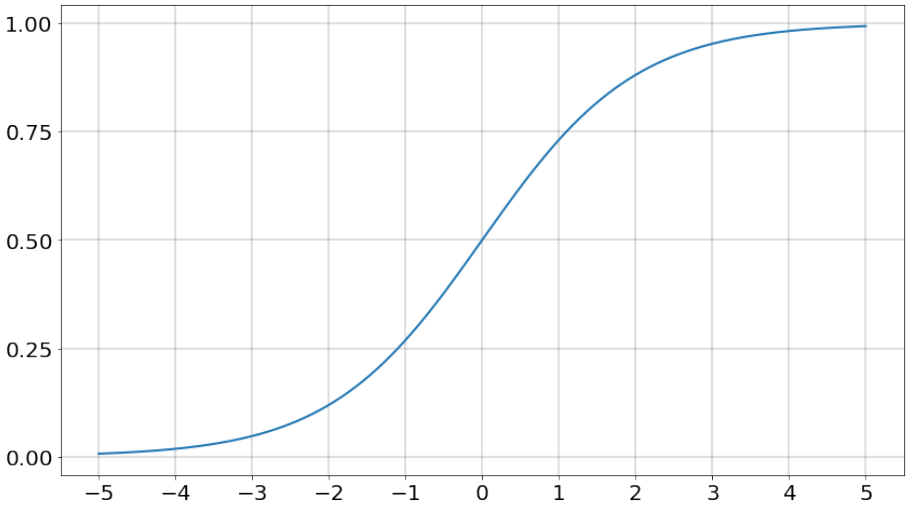
Częstym wyborem funkcji aktywacji są funkcje sigmiodalne. Jest to grupa monotonicznie rosnących funkcji, których zbiór wartości jest ograniczony przez asymptoty o skończonych wartościach, do których wartość funkcji dąży w $\pm\infty$ [lecun98]. Jednym z najczęściej wykorzystywanych przykładów funkcji sigmoidalnych jest sigmoida zdefiniowana równaniem

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{1}$$

Sigmoida jest różniczkowalna w każdym punkcie co pozwala używać podczas procesu uczenia metod optymizacji wykorzysujących gradient. Ponadto pochod-

na względem argumentu x wyraża się prostą relacją

$$\frac{d}{dx}\sigma(x) = \sigma(x)\left(1 - \sigma(x)\right). \tag{2}$$



Rysunek 2: Przykład funkcji sigmoidalnej - sigmoida, $\sigma(x) = \frac{1}{1+e^{-x}}$

Innym przykładem często wykorzystywanej w sztucznych sieciach neuronowych funkcji sigmoidalnej jest tangens hyperboliczny (prawy dolny róg Rys. 1). Wzór tej funkcji możemy wyrazić korzystając z definicji sigmoidy

$$\mathrm{tgh}(x) = 2\sigma(2x) - 1 \tag{3}$$

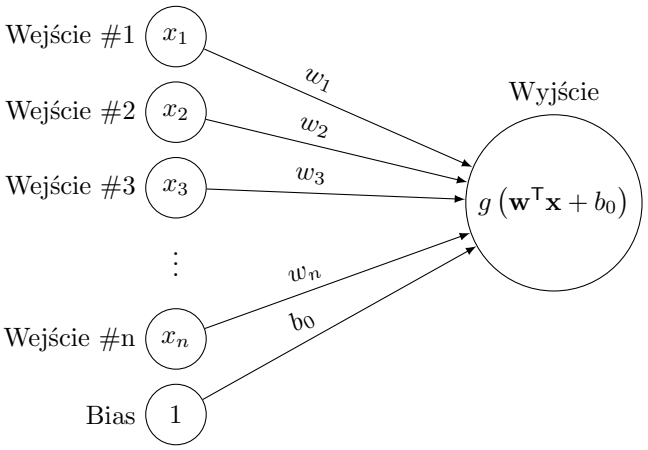
Jedną z zalet tej funkcji jest symetryczność względem początku układu współrzędnych.

1.2 Interpretacja probabilistyczna sigmoidy

Zastosowanie sigmoidy jako funkcji aktywacji naturalnie wynika z postaci prawdopodobieństwa a posteriori w Bayesowskim podejściu do problemu klasyfikacji dwóch klas. Rozważmy sztuczną sieć neuronową z jedną warstwą ukrytą oraz funkcję dyskryminacyjną $y(\mathbf{x})$ taką, że wektor $\mathbf{x}$ jest przypisany do klasy C_1 jeśli $y(\mathbf{x}) > 0$ i do klasy C_2 jeśli $y(\mathbf{x}) < 0$.
W najprostszej, liniowej formie funkcja może być zapisana jako:

$$y(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\mathbf{x} + b_0. \tag{4}$$

Wektor $\mathbf{w}$, to d-wymiarowy wektor wag, natomiast parametr b_0 to bias. Rozważmy funkcję $g(\cdot)$ nazywaną dalej funkcją aktywacji, która jako argument przyj-



Rysunek 3: Reprezentacja funkcji dyskryminacyjnej $y(x)$ w postaci diagramu sieci neuronowej, mającej n wejść, parametr bias i jedno wyjście.

muje jako argument sumę z równania (4):

$$y = g\left(\mathbf{w}^\mathsf{T}\mathbf{x} + b_0\right) \tag{5}$$

Załóżmy, że funkcja rozkładu prawdopodobieństwa danych pod warunkiem klasy C_k zadane jest przez wielowymiarowy rozkład normalny z równymi macierzami kowariancji $\Sigma_1 = \Sigma_2 = \Sigma$

$$p(x|C_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{x} - \mu_k\right)^\mathsf{T} \Sigma^{-1} \left(\mathbf{x} - \mu_k\right)\right]. \tag{6}$$

Prawdopodobieństwo a posteriori klasy C_1 można zapisać używając twierdzenia Bayesa:

$$\begin{aligned} p(C_1|\mathbf{x}) &= \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)} \\ &= \frac{1}{1 + \frac{p(\mathbf{x}|C_2)p(C_2)}{p(\mathbf{x}|C_1)p(C_1)}} \\ &= \frac{1}{1 + \exp(-a)}, \end{aligned} \tag{7}$$

gdzie

$$\begin{aligned} a &= \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)} \\ &= \left(\mu_1 - \mu_2\right)^\mathsf{T} \Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_1^\mathsf{T}\mu_1 + \frac{1}{2}\mu_2^\mathsf{T}\Sigma^{-1}\mu_2 + \ln \frac{p(C_1)}{p(C_2)}, \end{aligned} \tag{8}$$

pamiętając o tym, że macierz kowariancji jest symetryczna otrzymujemy

$$\mathbf{x} = \Sigma^{-1}\left(\mu_1 - \mu_2\right) \tag{9a}$$

$$b_0 = -\frac{1}{2}\mu_1^\mathsf{T}\mu_1 + \frac{1}{2}\mu_2^\mathsf{T}\Sigma^{-1}\mu_2 + \ln \frac{p(C_1)}{p(C_2)} \tag{9b}$$

Zatem widzimy, że użycie funkcji aktywacji w postaci sigmoidy pozwala nie tylko dokonać decyzji klasyfikacji ale również interpretować wynik funkcji dyskryminacyjnej jako prawdopodobieństwa a posteriori.