Activation functions

Rafaĺ Skrzypiec

1 Activation functions

There are many possible choices for the non-linear activation functions in a multilayered network... biology czemu non-linear? rne funkcje w zalenoci od zastosowa, ReLU, Sigmiod, TanH, Heavyside

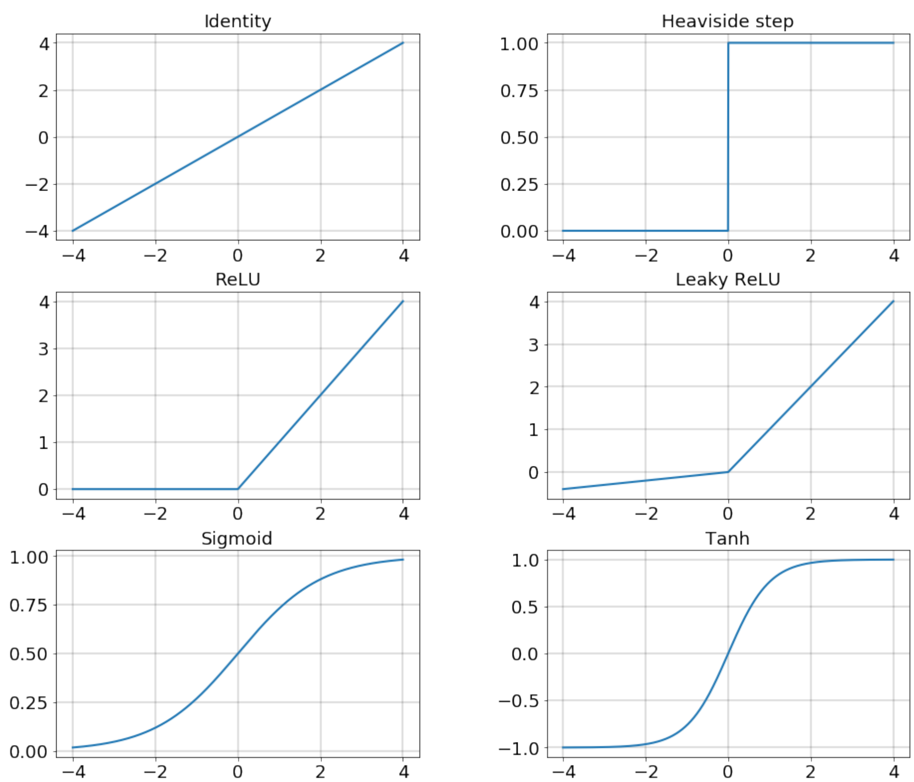

Figure 1: COO

1.1 Sigmoid

-nieliniowa -dobrze odwzorowuje liniowe zaleznosci dla niewielkich wag -moze byc funkcj schodkow dla duej wartosci sumy -prosta pochodna

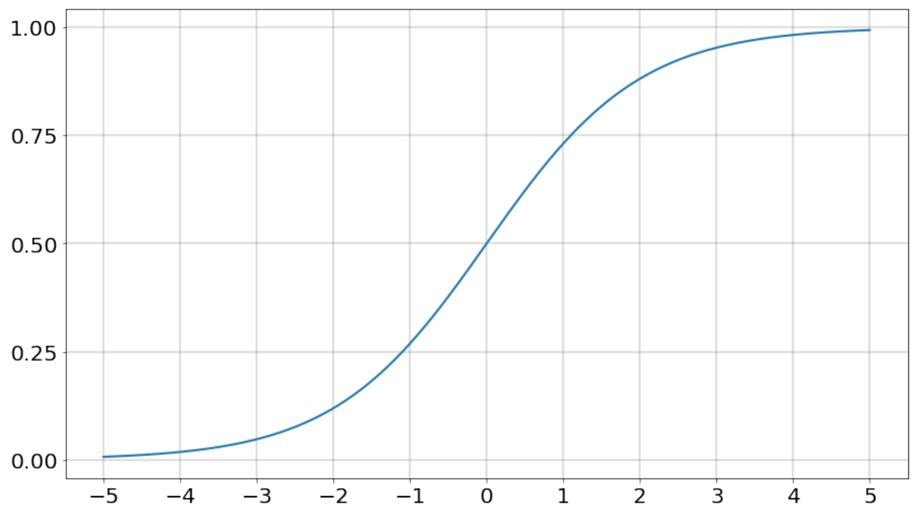

Figure 2: Example of sigmoidal function - sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$

narysowa jeszcze tanh, zwizek z sigmoid, jeden obrazek z kilkoma funkcjami aktywacji

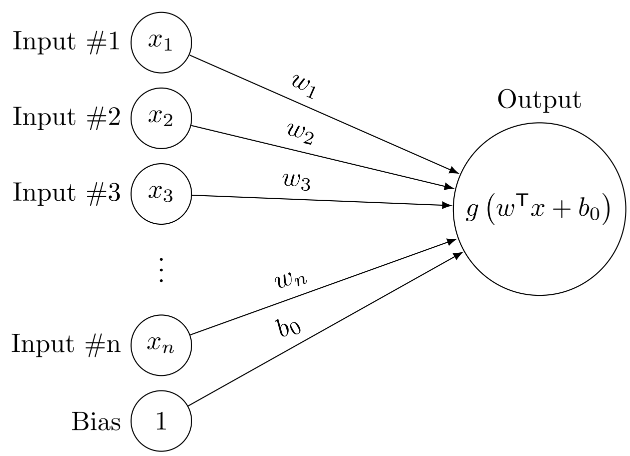

Figure 3: Representation of discriminant function $y(x)$ as a neural network diagram, having n inputs, bias term and one output.

1.2 Probabilistic interpretation of sigmoid

The application of sigmoid as an activation function arises naturally as the form of the posterior probability distribution in Bayesian treatment of two-class classification problem. Let us consider a single-layer network and the concept of a discriminant function $y(x)$, such that the vector x is assigned to class C_1 if $y(x) > 0$ and to class C_2 if $y(x) < 0$.

In the simplest, linear form, discriminant function can be written as:

$$y(x) = w^{\mathsf{T}}x + b_0. \tag{1}$$

We should refer to d-dimensional vector w as the weight vector and the parameter b_0 as the bias. There are several ways to generalize such functions, here we consider a function $g(\cdot)$ called activation function that acts on a aforementioned linear sum, and gives a discrimination function of the form

$$y = g\left(w^{\mathsf{T}}x + b_0\right) \tag{2}$$

Assumption that probability distribution functions of data given the class C_k are given by Gaussian distributions with equal covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$ gives:

$$p(x|C_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x - \mu_k)^{\mathsf{T}} \Sigma^{-1} (x - \mu_k)\right]. \tag{3}$$

The posterior probability of class C_1 can be written using Bayes' theorem:

$$\begin{aligned} p(C_1|x) &= \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1)+p(x|C_2)p(C_2)} \\ &= \frac{1}{1+\frac{p(x|C_2)p(C_2)}{p(x|C_1)p(C_1)}} \\ &= \frac{1}{1+\exp(-a)}, \end{aligned} \tag{4}$$

where

$$\begin{aligned} a &= \ln \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} \\ &= (\mu_1-\mu_2)^{\mathsf{T}}\,\Sigma^{-1}x - \frac{1}{2}\mu_1^{\mathsf{T}}\mu_1 + \frac{1}{2}\mu_2^{\mathsf{T}}\Sigma^{-1}\mu_2 + \ln\frac{p(C_1)}{p(C_2)}, \end{aligned} \tag{5}$$

hence

$$w = \Sigma^{-1}(\mu_1-\mu_2) \tag{6a}$$

$$b_0 = -\frac{1}{2}\mu_1^{\mathsf{T}}\mu_1 + \frac{1}{2}\mu_2^{\mathsf{T}}\Sigma^{-1}\mu_2 + \ln\frac{p(C_1)}{p(C_2)} \tag{6b}$$

Thus the network output is given by a sigmoid activation function acting on a weighted linear combination of inputs. This reasoning can be generalized to multi-layered network. Then outputs of each hidden neuron with logistic sigmoid activation function can be interpreted as probabilities of the presence of corresponding attribute conditioned by the inputs.