Activation functions

Rafał Skrzypiec

1 Activation functions

Activation function is a function that acts on every unit in neural network and gives the output value for neuron's input or set of inputs.

The perceptron, which is a progenitor and inspiration of artificial neural network was constructed as a simplified model of a biological neuron. In neuroscience, each neuron is connected with other neurons via connections that conduct electrical impulses. If the sum of the input signals surpasses a certain threshold neuron transmits electrical signal. Translating it to perceptron idea, only if the sum of input values is greater than zero, signal is propagated through neuron. Thus perceptron uses activation function that is shown in left upper corner of Fig. 1 - Heaviside step function. The single-layered perceptron is the simplest example of feed-forward neural network.

However it is not commonly used activation function, it does not include highly desirable property. One of the most important properties of activation functions is nonlinearity, that is the attribute that gives neural network nonlinear capabilities [Lecun].

1.1 Sigmoid

A typical choice of activation function is sigmoid. It is nonlinear function that is monotonically increasing. Logistic sigmoid function is defined by equation

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

It is continuously differentiable what makes possible to use gradient-based optimization methods. Derivative with respect to x is given by simple relation

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

[zwiazek z tanh, wspomniec kiedy uzywa sie czego albo wykasowac obrazek]

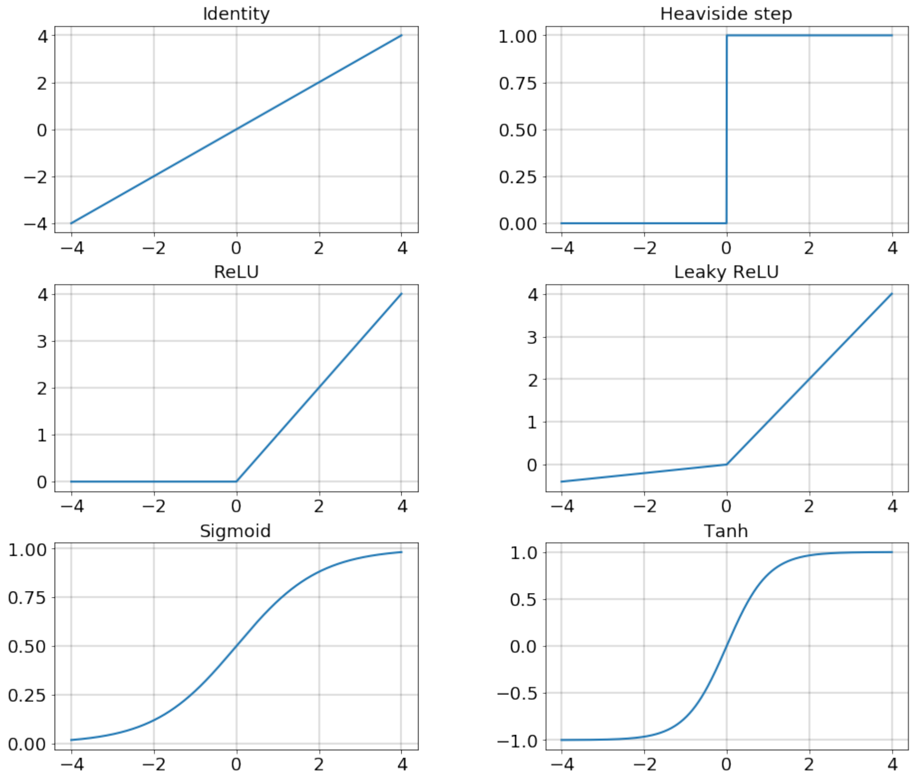

Figure 1: Set of several activation functions.

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:

$$y(x) = w^{\mathsf{T}}x + b_0. \tag{3}$$

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

$$y = g\left(w^{\mathsf{T}}x + b_0\right) \tag{4}$$

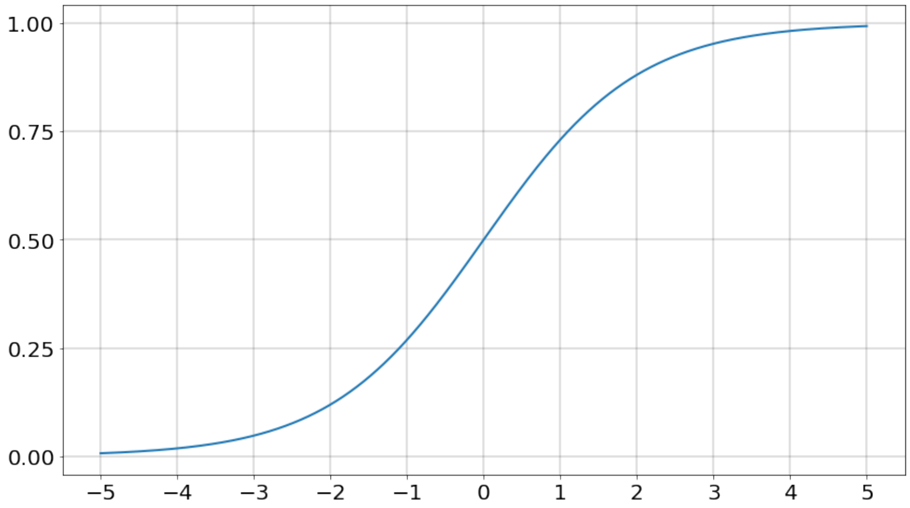

Figure 2: Example of sigmoidal function - sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$

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:

$$p(x|C_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_k)^{\mathsf{T}} \Sigma^{-1} (x-\mu_k)\right]. \tag{5}$$

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

$$\begin{aligned} p(C_1|x) &= \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1)+p(x|C_2)p(C_2)} \\ &= \frac{1}{1+\frac{p(x|C_2)p(C_2)}{p(x|C_1)p(C_1)}} \\ &= \frac{1}{1+\exp(-a)}, \end{aligned} \tag{6}$$

where

$$\begin{aligned} a &= \ln\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} \\ &= (\mu_1-\mu_2)^{\mathsf{T}} \Sigma^{-1} x - \frac{1}{2}\mu_1^{\mathsf{T}}\mu_1 + \frac{1}{2}\mu_2^{\mathsf{T}}\Sigma^{-1}\mu_2 + \ln\frac{p(C_1)}{p(C_2)}, \end{aligned} \tag{7}$$

hence

$$w = \Sigma^{-1}(\mu_1-\mu_2) \tag{8a}$$

$$b_0 = -\frac{1}{2}\mu_1^{\mathsf{T}}\mu_1 + \frac{1}{2}\mu_2^{\mathsf{T}}\Sigma^{-1}\mu_2 + \ln\frac{p(C_1)}{p(C_2)} \tag{8b}$$

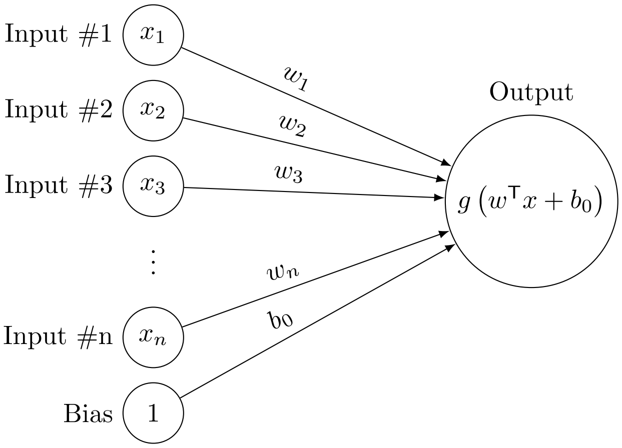

Figure 3: Representation of discriminant function $y(x)$ as a neural network diagram, having n inputs, bias term and one output.

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.