Machine Learning Course - CS-433

Neural Nets – Some Popular Activation Functions

Dec 8, 2016

©Ruediger Urbanke 2016

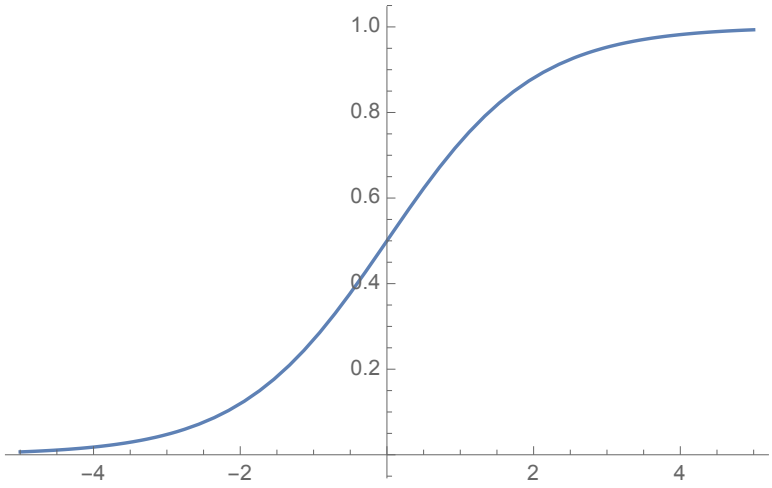

Figure 1: The sigmoid function $\phi(x)$.

Motivation

There is a variety of poplular activation functions that are being used. Let us list here some of them and briefly discuss their merits.

Sigmoid

We start with the sigmoid $\phi(x)$, which we have encountered already several times. Just to summarize, it is defined by

$$\phi(x) = \frac{1}{1 + e^{-x}},$$

and a plot is shown in Figure 1. Note that the sigmoid is always positive (not really an issue) and that it is bounded. Further, for $|x|$ large, $\phi'(x) \sim 0$. This can cause the gradient to become very small (which is known as the "vanishing gradient problem"), sometimes making learning slow.

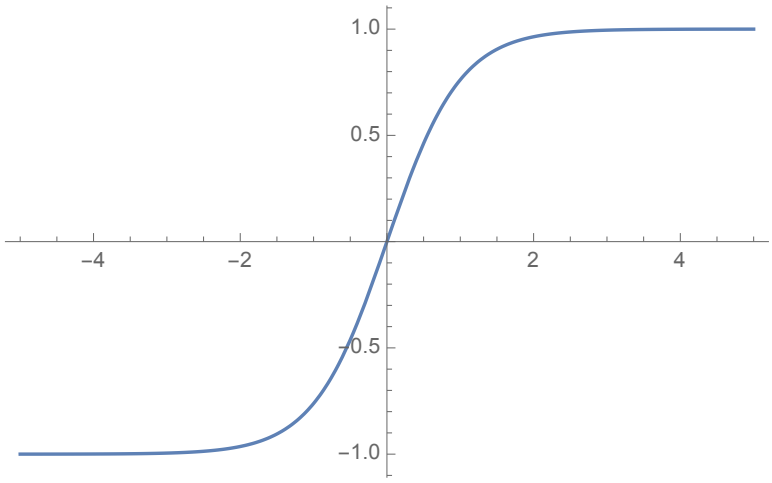

Figure 2: $\tanh(x)$.

Tanh

Very much related to the sigmoid is $\tanh(x)$. It is defined by

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \phi(2x) - \frac{1}{2},$$

and a plot is shown in Figure 2. Note that $\tanh(x)$ is "balanced" (positive and negative) and that it is bounded. But it has the same problem as the sigmoid function, namely for $|x|$ large, $\tanh'(x) \sim 0$. As mentioned, this can cause the gradient to become very small, sometimes making learning slow.

Rectified linear Unit – ReLU

Very popular is the rectified linear unit (ReLU) $(x)_+$ that we have also seen already. To recall, it is defined as

$$(x)_+ = \max\{0, x\},$$

and a plot is shown in Figure 3. Note that the ReLU is

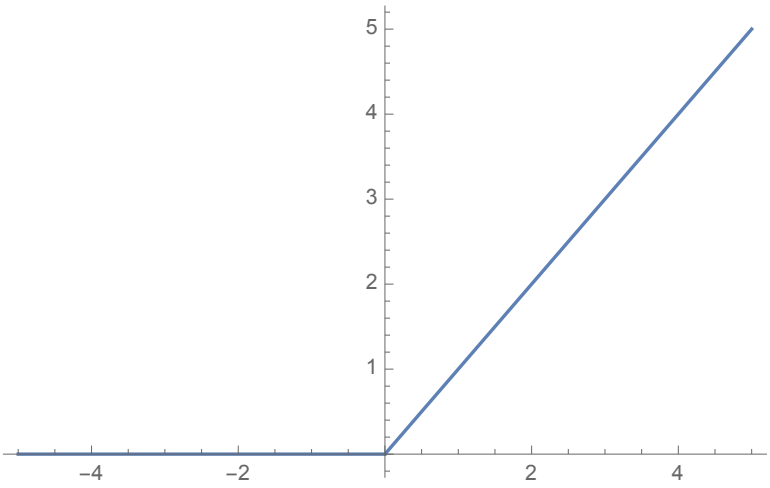

Figure 3: The ReLU $(x)_+$.

always positive and that it is unbounded. One nice property of the ReLU is that its derivative is 1 (and does not vanish) for positive values of x (it has 0 derivative for negative values of x though).

Leaky ReLU

In order to solve the 0-derivative problem of the ReLU (for negative values of x) one can add a very small slope α in the negative part. This gives rise to the leaky rectified linear unit (LReLU). It is defined as

$$f(x) = \max\{\alpha x, x\}$$

and a plot is shown in Figure 4. The constant α is of course a hyper-parameter that can be optimized.

Maxout

The maxout generalizes ReLU and LReLU. It is defined by

$$f(x) = \max\{\mathbf{x}^\top \mathbf{w}_1 + b_1, \cdots, \mathbf{x}^\top \mathbf{w}_k + b_k\}$$

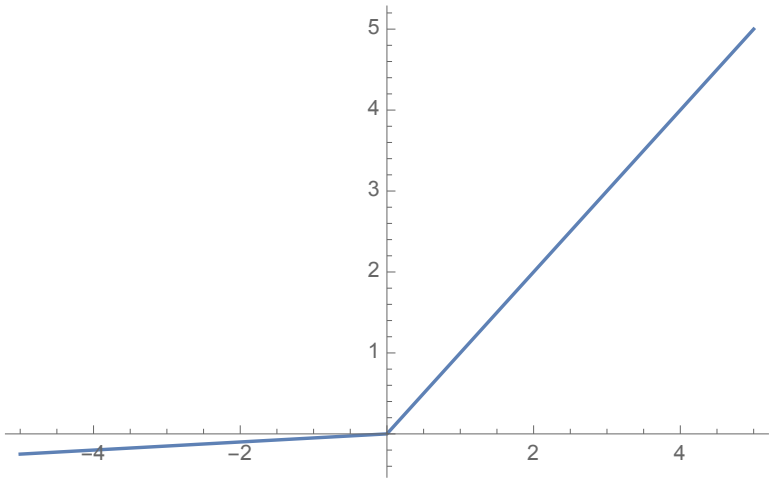

Figure 4: LReLU with $\alpha = 0.05$

and a plot is shown in Figure 5. The constants in this function are of course parameters that can be chosen for the particular application. Note that this activation function is quite different from the previous cases. In the previous cases we computed a weighted sum and then applied the activation function to it, whereas here we compute two or more different weighted sums and then choose the maximum.

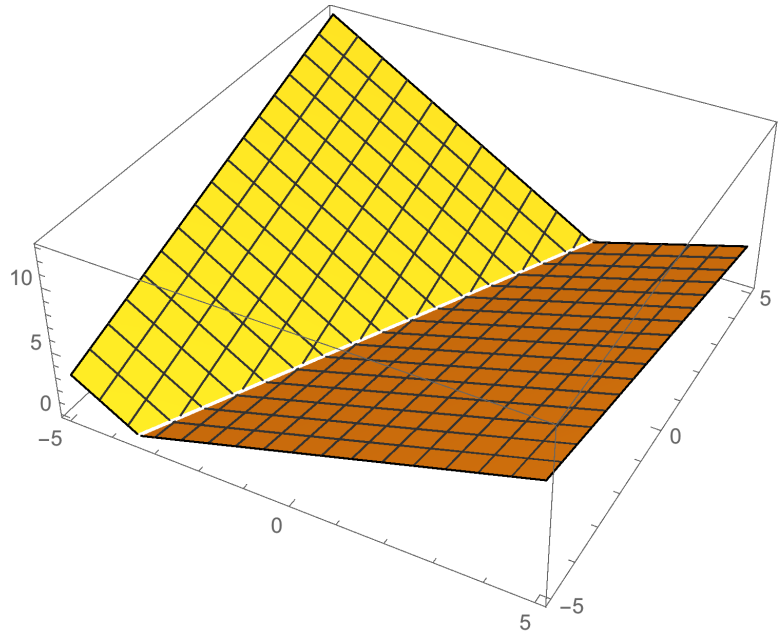

Figure 5: Maxout function with two terms, $\max\{x_1 - 0.5x_2 + 1, -2x_1 + x_2 - 2\}$.