Deep Learning

Machine Learning - Prof. Dr. Stephan Günnemann

Leonardo Freiherr von Lerchenfeld

December 20, 2017

Contents

1 Activation functions

1.1 Problem 1

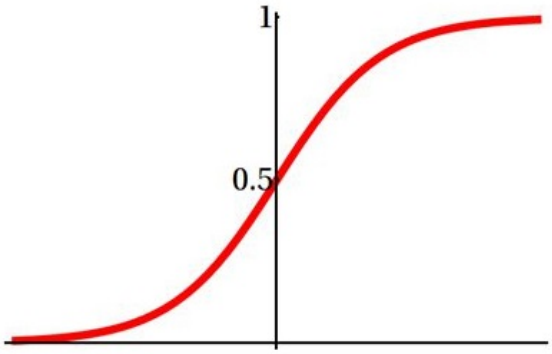
We use basis functions to map samples to a space where they are (almost) linearly separable.

1.2 Problem 2

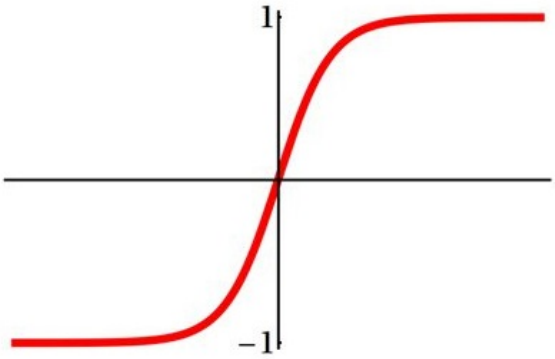
When we take a look at the two non-linear activation functions, we recognize two similarities

1. The range of $tanh(x)$ is $2\times$ bigger than $\sigma(x)$
2. The functions are shifted to each other by 1

$$\sigma(\Sigma) = \frac{1}{1+e^{-\Sigma}} \qquad \tanh(\Sigma) = \frac{e^{\Sigma}-e^{-\Sigma}}{e^{\Sigma}+e^{-\Sigma}}$$



logistic (sigmoid, unipolar)



tanh (bipolar)

$$\begin{aligned}
2\sigma(y) - 1 &= tanh(x) \\
2\frac{1}{1+e^{-y}} - 1 &= \\
2\frac{e^{y}}{1+e^{y}} - \frac{1+e^{y}}{1+e^{y}} &= \frac{e^{2x}-1}{e^{2x}+1} \\
\frac{e^{y}-1}{e^{y}+1} &= \\
y &= 2x \\
2\sigma(2x) - 1 &= tanh(x)
\end{aligned}$$

1.3 Problem 3

$$\begin{aligned}
\frac{d}{dx}tanh(x) &= \frac{d}{dx}\frac{sinh(x)}{cosh(x)} \\
&= \frac{cosh(x)\frac{d}{dx}sinh(x) - sinh(x)\frac{d}{dx}cosh(x)}{cosh^2(x)} \\
&= \frac{cosh(x)cosh(x) - sinh(x)sinh(x)}{cosh^2(x)} \\
&= 1 - tanh^2(x)
\end{aligned}$$

The derivative of the tanh activation function can also be expressed in terms of the function value itself. This fact is usually used for an efficient implementation of the gradient.

1.4 Problem 4

$E_{old}(w) = -\sum_{p=1}^{n} y_p log f(x_p, w) + (1 - y_p) log[1 - f(x_p, w)]$
If we consider a network having an output $-1 \leq f(x_p, w) \leq 1$ and target values $y_{new} \in \{-1, 1\}$ instead of $y_{old} \in \{0, 1\}$, we have to scale and shift the outputs $f_{new}(a) = 2f_{old}(a) - 1$

$$E_{new}(w) = -\sum_{p=1}^{n} \frac{1 + y_p}{2} log \frac{1 + f(x_p, w)}{2} + \left(1 - \frac{1 + y_p}{2}\right) log \left(1 - \frac{1 + f(x_p, w)}{2}\right)$$
In this case an appropriate choice of activation function is

$$\begin{aligned} f(a) &= 2\sigma(a) - 1 \\ &= tanh\left(\frac{a}{2}\right) \end{aligned}$$

2 Optimization

2.1 Problem 5

If $|\eta| < 1$ then $f(\eta) = \frac{1}{2}\eta^2$ & $f'(\eta) = \eta$
else $f(\eta) = |\eta| - \frac{1}{2}$ & $f'(\eta) = \frac{\eta}{|\eta|}$

If $|y_i - wx_i| < 1$ then $f(w) = \frac{1}{2}(y_i - wx_i)^2$ & $f'(w) = -x_i(y_i - wx_i)$
else $f(w) = |y_i - wx_i| - \frac{1}{2}$ & $f'(w) = -x_i \frac{y_i - wx_i}{|y_i - wx_i|}$

$E'(w) = \frac{1}{m} \sum_{i=1}^{m} f'(w) + \lambda w$

2.2 Problem 6

Overfitting typically occurs when we try to model the training data perfectly (low training error). Overfitting means poor generalization! The validation performance tells us how well our model generalizes. For this reason, when the error of the validation rises again (after 50 iterations) we stop training. We only touch the test set once at the end to report final performance!

3 Numerical stability

3.1 Problem 7

$$\begin{aligned} log\sum_{i=1}^{N} e^{x_i} &= a + log\sum_{i=1}^{N} e^{x_i - a} \\ &= a + log\sum_{i=1}^{N} e^{x_i} e^{-a} \\ &= a + loge^{-a} + \sum_{i=1}^{N} e^{x_i} \\ &= a - a + \sum_{i=1}^{N} e^{x_i} \\ &= \sum_{i=1}^{N} e^{x_i} \end{aligned}$$

3.2 Problem 8

$$\begin{aligned} \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}} &= \frac{e^{x_i - a}}{\sum_{i=1}^{N} e^{x_i - a}} \\ &= \frac{e^{x_i} e^{-a}}{\sum_{i=1}^{N} e^{x_i} e^{-a}} \\ &= \frac{e^{-a} e^{x_i}}{e^{-a} \sum_{i=1}^{N} e^{x_i}} \\ &= \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}} \end{aligned}$$

3.3 Problem 9

$$\begin{aligned} & -(ylog(\sigma(x)) + (1-y)log(1-\sigma(x))) && (1) \\ = & -\left(ylog(\frac{1}{1+e^{-x}}) + log(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}) - ylog(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}})\right) && (2) \\ = & -\left(y(log(1) - log(1+e^{-x}) + log(e^{-x}) - log(1+e^{-x}) - y(log(e^{-x}) - log(1+e^{-x})))\right) && (3) \\ = & \ x - xy + log(1+e^{-x}) && (4) \\ & \ if\ x > 0 && (5) \\ & \ x - xy + log(1+e^{-x}) && (6) \\ = & \ max(x,0) - xy + log(1+e^{-|x|}) && (7) \\ & \ if\ x \leq 0 && (8) \\ & \ log(e^{x}) + log(1+e^{-x}) - xy && (9) \\ = & \ log(1+e^{x-x}) - xy + log(1+e^{x}) && (10) \\ = & \ max(x,0) - xy + log(1+e^{-|x|}) && (11) \end{aligned}$$