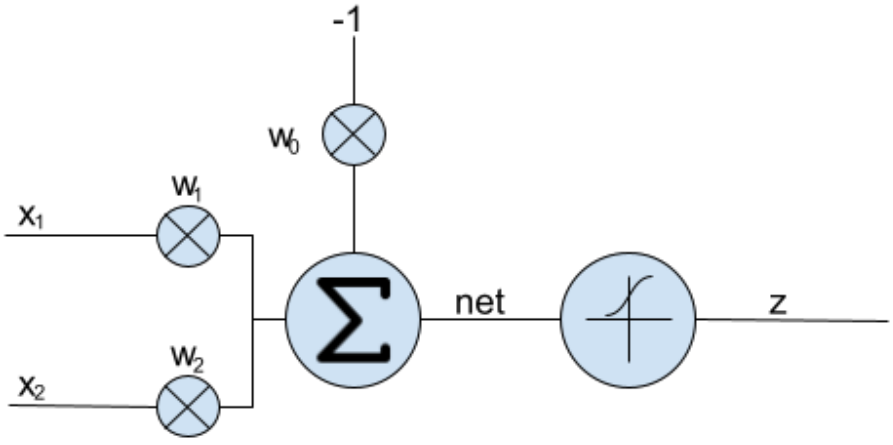
Die Mathematik der neuronalen Netze

Eine Einführung

By
Luca Ritz



2021
Luca Ritz

Inhaltsverzeichnis

1 Einführung **2**

2 Die Mathematik der neuronalen Netze **3**
2.1 Das Perceptron 3
2.1.1 Lernverfahren 4
2.1.2 Das Problem mit XOR und nichtlinearen Funktionen 4
2.2 Neuronale Netze 4
2.2.1 Lernverfahren mit Gradientenabstieg 4
2.2.2 Lernverfahren mit Backpropagation 4
2.2.3 XOR und die Lösung 4

Abbildungsverzeichnis **5**

Glossar **6**

3 Anhang **7**
3.1 Die Ableitung 1. Grades 7
3.2 Die partielle Ableitung 7
3.3 Die Sigmoide 8
3.4 Der Gradient 10
3.5 Das Gradientenabstiegsverfahren 10

Kapitel 1

Einführung

In Anbetracht der Tatsache, dass neuronale Netzwerke in Zukunft eine grössere Rolle spielen, wird in diesem Dokument das Ziel verfolgt, die Mathematik dahinter verständlich zu erklären.

Kapitel 2

Die Mathematik der neuronalen Netze

2.1 Das Perceptron

Zu Beginn steht das Perceptron. Es wird hier wie in Abbildung 2.1 veranschaulicht. Es besteht aus der Summe mehrerer gewichteter Eingabewerte wie auch einem Bias, welcher die Schwelle einer Aktivierung verschiebt. Die Gewichte werden mit w_i, die Eingabewerte mit x_i bezeichnet. Diese Summe, im folgenden als net bezeichnet, wird in eine Aktivierungsfunktion, hier eine Sigmoide, gegeben. Der resultierende Wert wird als z bezeichnet. Der Bias hat den festen Wert -1, er wird wiederum über ein Gewicht w_0 trainiert.

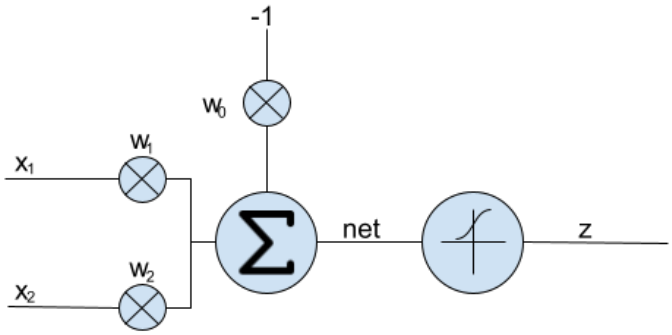


Abbildung 2.1: Das Perceptron

2.1.1 Lernverfahren

2.1.2 Das Problem mit XOR und nichtlinearen Funktionen

2.2 Neuronale Netze

2.2.1 Lernverfahren mit Gradientenabstieg

2.2.2 Lernverfahren mit Backpropagation

2.2.3 XOR und die Lösung

Abbildungsverzeichnis

2.1 Das Perceptron 3

3.1 Steigung an einem bestimmten Punkt der Funktion $f(x)$ 7
3.2 Die Sigmoid-Funktion 8
3.3 Die Ableitung der Sigmoid-Funktion 9
3.4 Gradient an der Position $(1, 1)$ 10

Glossar

Kapitel 3

Anhang

3.1 Die Ableitung 1. Grades

Die Ableitung 1. Grades beschreibt die Steigung an einem bestimmten Punkt der Funktion. In der Abbildung 3.1 wird eine Funktion $f(x)$ (in grün) gegeben. Die Steigung $f'(x)$ an einem bestimmten Punkt x ist rot markiert. Die Ableitung selbst ist wiederum eine Funktion und kann über diverse Ableitungsregeln aufgrund der gegebenen Funktion selbst gebildet werden.

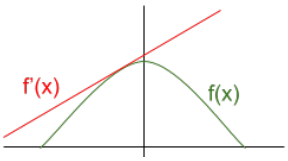


Abbildung 3.1: Steigung an einem bestimmten Punkt der Funktion $f(x)$

Die Ableitungsregeln, welche im Zuge der Erklärung des Lernprozesses eines neuronalen Netzwerks benötigt werden, sind nachfolgend ersichtlich.

Potenzregel $f(x) = x^n \longrightarrow f'(x) = n \cdot x^{n-1}$

Kettenregel $f(x) = u(v(x)) \longrightarrow f'(x) = u'(v(x)) \cdot v'(x)$

Es existieren viele weitere Ableitungsregeln, auf die hier nicht weiter eingegangen wird. Die Schreibweise der Ableitung einer Funktion nach einer Variablen lautet $f'(x) = \frac{\delta f(x)}{\delta x}$.

3.2 Die partielle Ableitung

Ist der Input einer Funktion mehrdimensional, das heisst, die Funktion f ist abhängig von mehreren Variablen, dann kann die Ableitung jeweils lediglich nach einer Variablen gebildet werden. Die übrigen Variablen werden als konstant angesehen. In dem Fall beschreibt die partielle Ableitung die Steigung an einem bestimmten Punkt der abgeleiteten Dimension. Es sei als Beispiel die Funktion $f(x,y) = x^2 + y^2 + 10$ gegeben. Diese wird nun partiell nach x sowie

nach z abgeleitet.

$$f^x(x,y) = 2x \tag{3.1}$$
$$f^y(x,y) = 2y \tag{3.2}$$

Die hierbei angewendete Ableitungsregel ist die Potenzregel, welche bereits im vorangegangenen Kapitel erwähnt wurde.

3.3 Die Sigmoide

Als Aktivierungsfunktion wird die Sigmoidfunktion benutzt. Diese wird heutzutage meist nicht mehr eingesetzt aufgrund des schlechten Lernverhaltens in einigen Bereichen der Funktion. Die erwähnte Eigenschaft wird bei der Betrachtung der Ableitung ersichtlich.

$$sig(t) = \frac{1}{1+e^{-t}} \tag{3.3}$$

Geometrisch lässt sich die Funktion wie in Abbildung 3.2 so interpretieren, dass eine gewisse Schwelle existiert, ab der die Funktion den Eingabewert auf 1 abbildet, in dem Jargon der Neuronen also „feuert“. Das Resultat lautet entweder 0 oder 1.

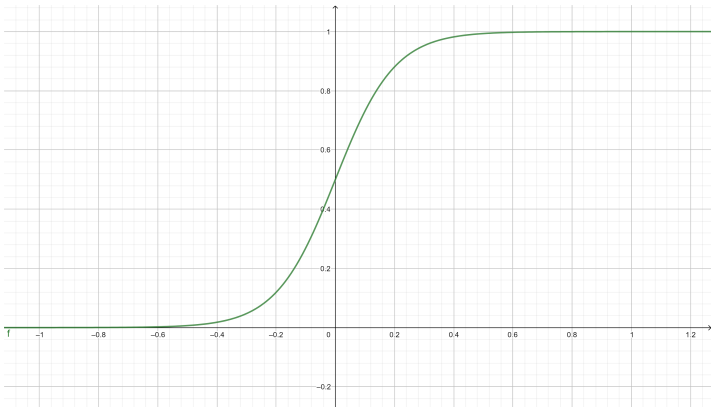


Abbildung 3.2: Die Sigmoid-Funktion

Für die weitere Verwendung ist nun vor allem die Ableitung der Sigmoide interessant, welche in den nachfolgenden Zeilen behandelt wird. Zuerst wird der Bruch durch eine andere Schreibweise (Exponent -1) dargestellt. Weiterhin lässt sich nun die Sigmoide als Verkettung von zwei Funktionen f und y schreiben.

$$sig(t) = (1+e^{-t})^{-1} \longrightarrow y(t) = 1+e^{-t}, f(t) = y(t)^{-1} \tag{3.4}$$

Es handelt sich also um eine äussere und innere Funktion, wobei nun die Ableitungsregel 3.1 angewendet wird. Für die innere Ableitung wird nun noch die Regel $f(x) = e^x \longrightarrow f'(x) = e^x$ verwendet.

$$\frac{\delta f(t)}{\delta t} = (-1)\cdot(y(t))^{-2} \tag{3.5}$$
$$\frac{\delta y(t)}{\delta t} = (e^{-t})\cdot(-1) \tag{3.6}$$

Nach der Kettenregel resultiert:

$$\frac{\delta sig(t)}{\delta t} = (-1) \cdot (1 + e^{-t})^{-2} \cdot (e^{-}t) \cdot (-1) \tag{3.7}$$

$$\frac{\delta sig(t)}{\delta t} = \frac{e^{-t}}{(1 + e^{-t})^2} \tag{3.8}$$

Nun wird $\frac{1}{1+e^{-t}}$ ausgeklammert.

$$\frac{\delta sig(t)}{\delta t} = \frac{1}{1 + e^{-t}} \cdot \frac{e^{-t}}{1 + e^{-t}} \tag{3.9}$$

Beim Zähler wird 1 dazuaddiert und abgezogen, damit der Faktor umgeformt werden kann.

$$\frac{\delta sig(t)}{\delta t} = \frac{1}{1 + e^{-t}} \cdot \frac{e^{-t} + 1 - 1}{1 + e^{-t}} \tag{3.10}$$

$$\frac{\delta sig(t)}{\delta t} = \frac{1}{1 + e^{-t}} \cdot (\frac{1 + e^{-t}}{1 + e^{-t}} - \frac{1}{1 + e^{-t}}) \tag{3.11}$$

Es resultiert die Ableitung in bekannter Form 3.12

$$\frac{\delta sig(t)}{\delta t} = sig(t) \cdot (1 - sig(t)) \tag{3.12}$$

Auch hier kann eine geometrische Interpretation in Abbildung 3.3 erfolgen. Ersichtlich wird nun, dass die Steigung nur in einem sehr kleinen Intervall stark ungleich 0 ist. Dies bedeutet, dass sich im Falle eines solchen Variablenwerts, wo die Steigung fast 0 ist, kaum eine Korrektur durch den Gradienten[1] durchzuführen ist. Daher werden heutzutage Aktivierungsfunktionen wie die „ReLU“ oder „LeakyReLU“ verwendet, welche dieses Problem beheben.

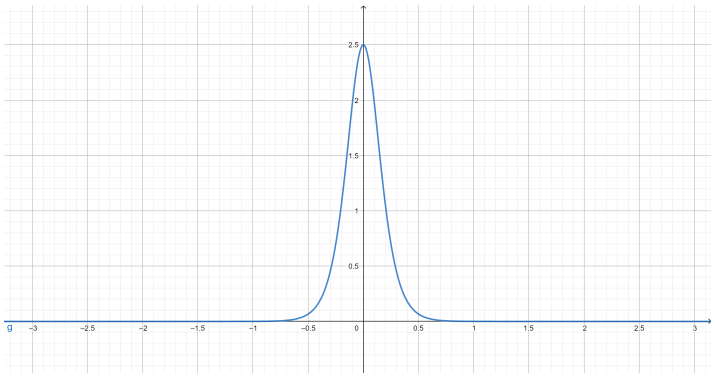


Abbildung 3.3: Die Ableitung der Sigmoid-Funktion

[1] Siehe Kapitel 3.4

3.4 Der Gradient

Der Gradient beschreibt einen Vektor, welcher in Richtung des steilsten Anstiegs einer Funktion zeigt. Die Komponenten des Gradientenvektors bestehen aus den partiellen Ableitungen der Funktion an der jeweiligen Variablen.

$$\nabla f(x,y) \longrightarrow \begin{pmatrix} \frac{\delta f(x,y)}{\delta x} & \frac{\delta f(x,y)}{\delta y} \end{pmatrix} \tag{3.13}$$

Geometrisch kann dies an der Position $(x = 1, y = 1)$ für die Funktion $f(x,y) = x^2 + y^2$ wie in Abbildung 3.4 aussehen. Zu beachten sei hier, dass der eigentliche Gradient in der XY-Ebene liegt (blau). Der schwarze Vektor soll lediglich anzeigen, was eine Verschiebung in dieser Richtung bei der Eingabe der Variablen für den Ausgabewert der Funktion bedeutet. Die Länge des Gradienten entspricht der Stärke der Steigung an diesem Punkt.

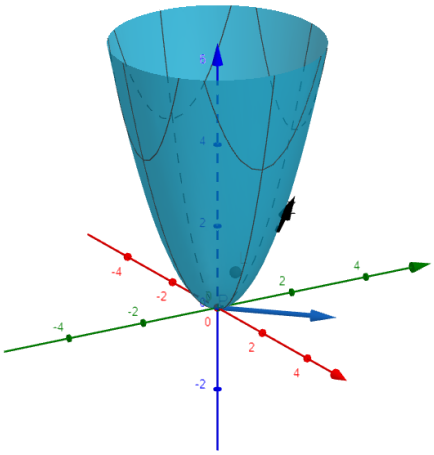


Abbildung 3.4: Gradient an der Position $(1,1)$

3.5 Das Gradientenabstiegsverfahren

Hierbei handelt es sich um ein Optimierungsverfahren, um einen Maximal- oder Minimalwert einer gegebenen Zielfunktion zu finden. Es wird in dem Fall der Gradient, wie in Kapitel 3.4 besprochen, verwendet. Die Idee ist, dass man bei einer Maximierung in kleinen Schritten in Richtung des Gradienten folgt. Als Beispiel wird eine Funktion angegeben, die von zwei Variablen x, y abhängig ist. Beim Wert λ handelt es sich um die Lernrate, welche die Länge der zu gehenden Schritte beeinflusst. Der $\vec{\nabla}$ steht hierbei für den Gradienten.

$$\begin{pmatrix} x_{neu} & y_{neu} \end{pmatrix} = \begin{pmatrix} x_{alt} & y_{alt} \end{pmatrix} + \lambda \cdot \vec{\nabla} \tag{3.14}$$

Geometrisch kann wiederum die Abbildung 3.4 hinzugezogen werden. Dort würde man in einem ersten Schritt in Richtung des schwarzen Vektors gehen, respektive in Richtung des Blauen, wenn man nur die Variablen beachtet.