Einführung in die Neuroinformatik

Tim Luchterhand, Paul Nykiel (Gruppe P)

2. Juli 2018

1 Transferfunktionen

1.1

a)

$$\frac{1+\tanh\left(\frac{x}{2}\right)}{2} = \frac{1+\frac{e^{\frac{x}{2}}-e^{-\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}}}{2} = \frac{\frac{e^{\frac{x}{2}}+e^{-\frac{x}{2}}+e^{x}-e^{-\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}}}{2} = \frac{e^{\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}} = \frac{1}{1+e^{-x}} = \mathrm{sig}(x)$$

b)

i. Positive Gradienten in grün, negative Gradienten in rot.

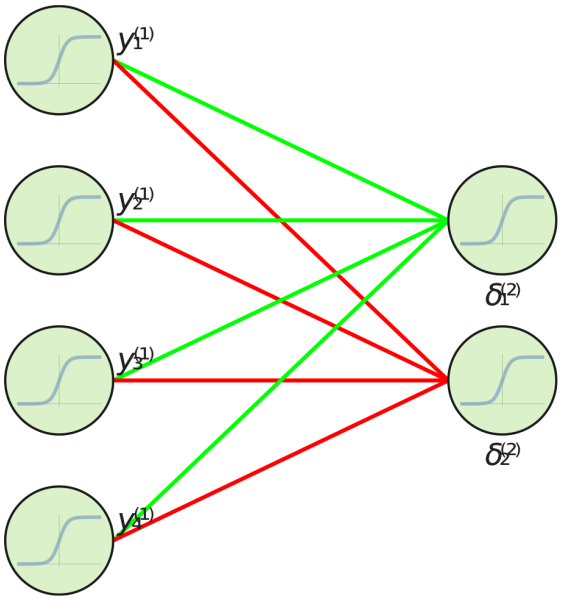

Abbildung 1: Vorzeichen des Gradienten mit Sigmoid-Aktivierung

ii. Positive Gradienten in grün, negative Gradienten in rot.

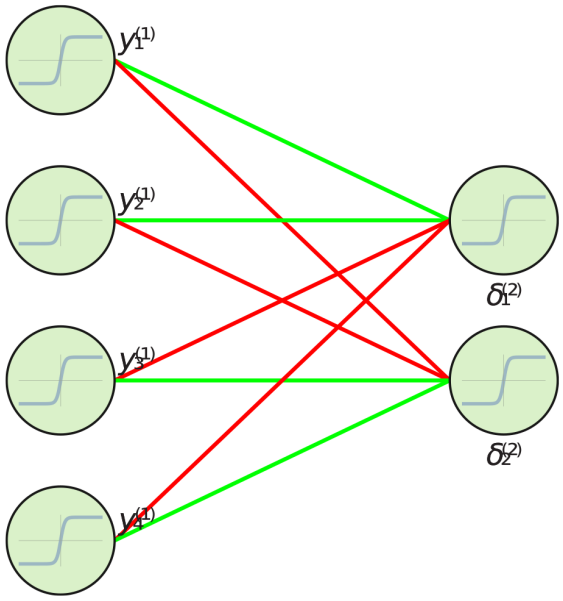

Abbildung 2: Vorzeichen des Gradienten mit Tanh-Aktivierung

iii. Wenn eines der Ausgangsneuronen einen betragsmäßig kleinen Fehler hat, so kann es passieren, dass bei der Sigmoid-Funktion der Gradient des anderen Neurons und damit sein Vorzeichen dominiert. Dadurch würde der Erwartungswert verschoben werden.

Bei der Tanh-Funktion ist dies weniger problematisch, da das Vorzeichen des Gradienten nicht nur vom Fehler δ abhängt.

1.2

a) Für negative dendritische Potentiale u ist der Ausgang des Layers 0. Außerdem ist die Ableitung für negative u ebenfalls 0 und im aktuellen Layer und in allen vorhergehenden Layern werden die Gewichte nicht adaptiert.

b) Der Rechenaufwand für die Berechnung der ReLU-Funktion ist deutlich geringer als die Berechnung des tanh. Die ReLU-Funktion ist quasi eine einzige Fallunterscheidung, während beim tanh aufwendige Gleitkommaoperationen durchgeführt werden müssen.

Obwohl der Rechenaufwand für die ReLU-Funktion kleiner ist, ist das Lernen nicht zwingend schneller, da für negative u gar nicht gelernt wird.

c) Wie oben bereits erläutert werden die Gewichte für negative u nicht adaptiert. LeakyReLU löst dieses Problem, indem es denn Wert von $f(x)$ für $x < 0$ nicht fix zu 0 setzt, sondern für $x \to \infty$ leicht abfallen lässt. Dadurch nimmt $f'(x)$ einen Wert $\neq 0$ $\forall x$ an.

1.3

a) Da die Funktion differenzierbar ist, kann mit Gradientenabstieg gelernt werden.

Die Ableitung der Leaky-ReLU-Funktion ist in 0 nicht stetig. Dadurch kommt es durch eine beliebig kleine Änderung des dendritischen Potentials zu einer komplett unterschiedlichen Adaption der Gewichte. Dadurch kann das Netz anfällig zu Rauschen auf den Daten werden.

b) i. Für $l \to \infty$ sollen die Gewichte gegen feste Werte konvergieren. Das heißt auch der Erwartungswert und die Varianz müssen gegen konvergieren. Das heißt für $l \to \infty$ müssen die Differenzen $|\mu_l - \mu_{l+1}|$ und $|\sigma_l - \sigma_{l+1}|$ gegen 0 konvergieren (Konvergenz nach Cauchy).

Wenn die Parameter der Verteilung so variieren, dauert es demzufolge lange bis die Gewichte konvergieren und das Training dauert lange.

ii. Die Verteilung der positiven dendritischen Potentiale bleibt durch SELU weitestgehend unverändert. Die negativen Werte werden in ihrer Verteilung jedoch stark gestaucht. Dadurch gibt es kaum noch extreme negative Werte, sondern viele, betragsmäßig kleine negative Werte. Durch die Stauchung wird das Histogramm schmaler und die Varianz dadurch kleiner. Außerdem reduziert sich der Erwartungswert, da es weniger extreme negative Werte gibt.

1.4

a)

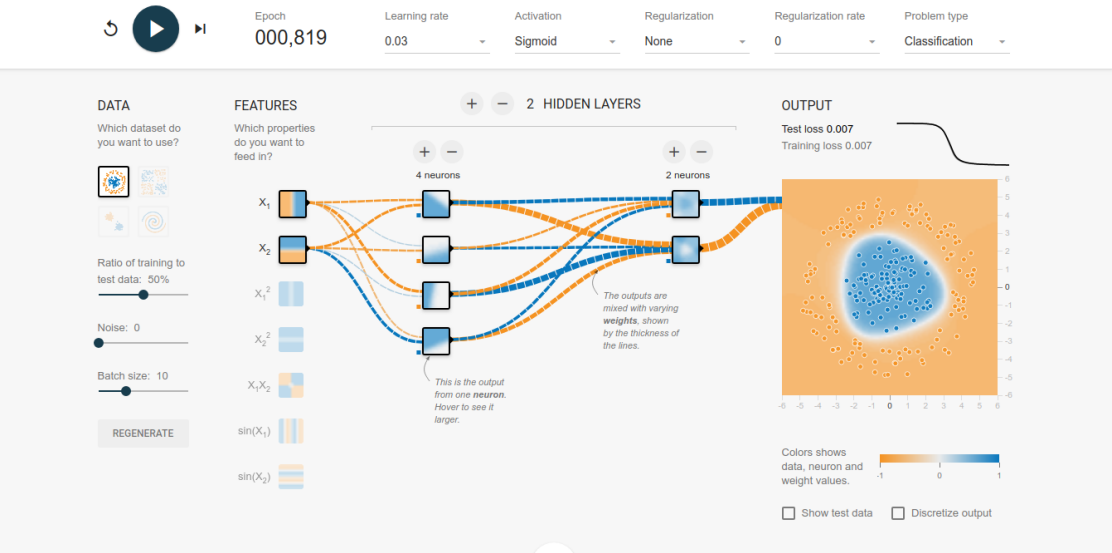

Abbildung 3: Training mit Sigmoid-Transferfunktion

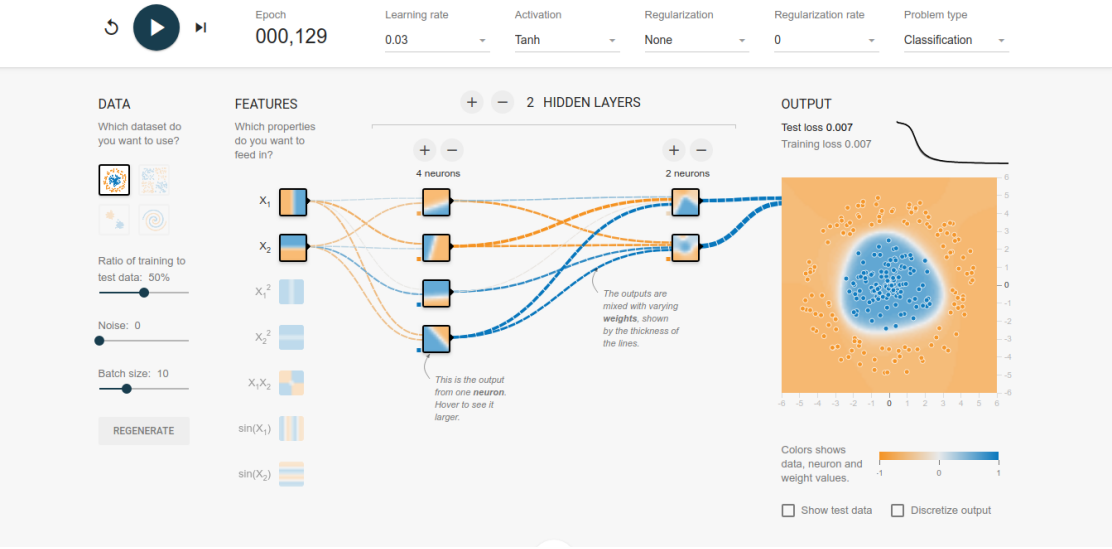

Abbildung 4: Training mit Tanh-Transferfunktion

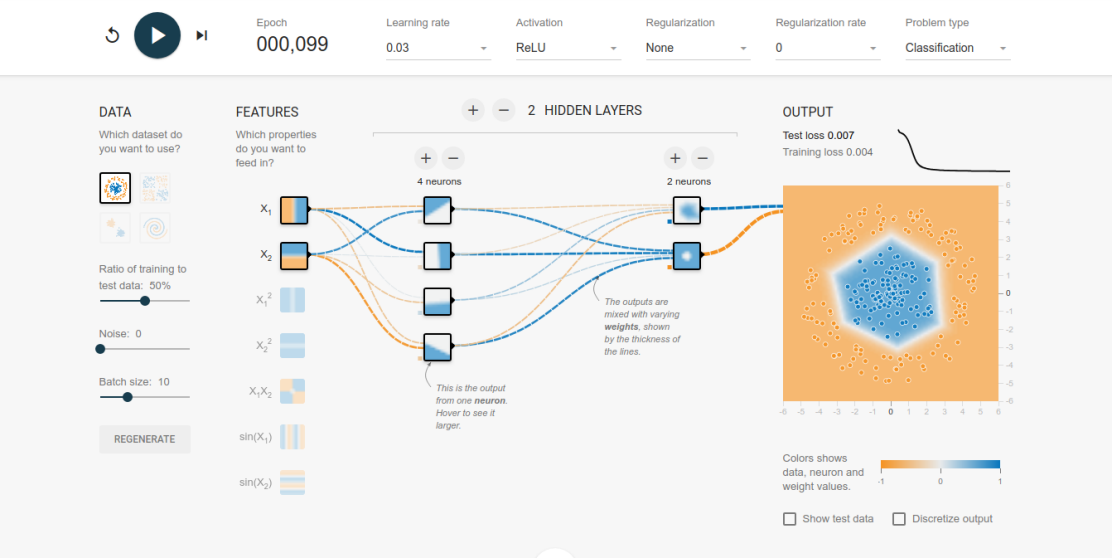

Abbildung 5: Training mit ReLU-Transferfunktion

b)

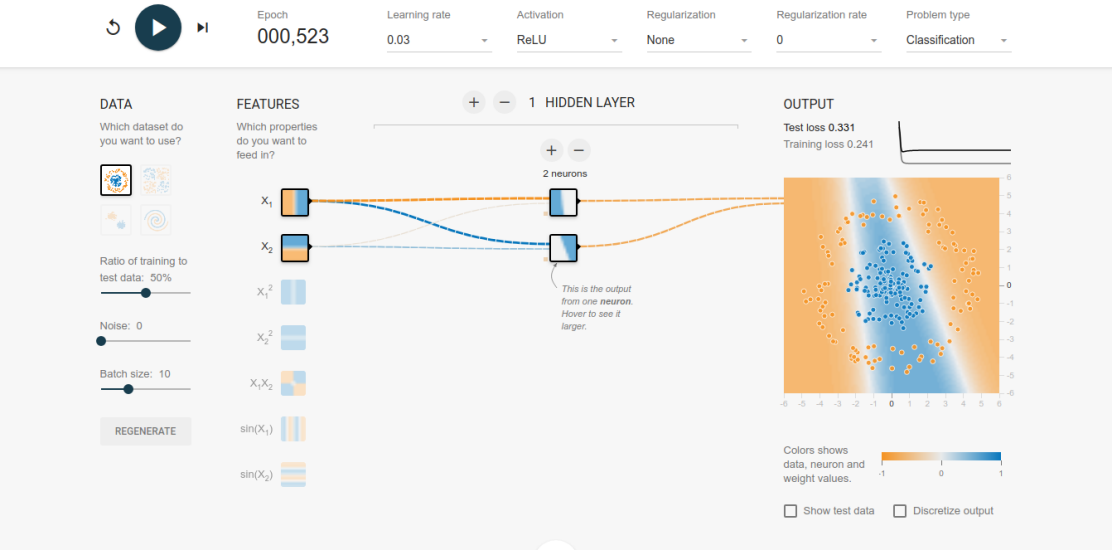

Abbildung 6: Dying Relu

1.5

a)

i. Die Varianz der Ausgangswerte jeden Layers nimmt bei fast allen Aktivierungsfunktionen ab. Die einzige Ausnahme stellt die SELU-Funktion dar. Bei dieser Funktion geht die Varianz gegen 1.

   Auch bei den Gradienten ist die Varianz im Falle der SELU-Aktivierung deutlich größer, bei allen anderen Aktivierungsfunktionen ist die Varianz des Gradienten sehr klein, bleibt aber über alle Layer hinweg relativ konstant. Bei der SELU-Funktion nimmt die Varianz der Gradienten über die Layer hinweg ab, ebenso der Mittelwert.

ii. Bei der Verteilung der Netzwerk-Aktivität ist der Mittelwert mit ReLU bzw. ELU Aktivierung in den ersten Layern positiv. Bei der Aktivierung mit tanh bzw. SELU ist der Mittelwert über alle Layer quasi 0.

   Bei der Verteilung des Gradienten lässt sich kein Unterschied erkennen. Die einzige Ausnahme bildet die Verteilung der Gradienten des Netzwerks mit SELU-Aktivierung. Hier wird der Mittelwert des Gradienten für weiter hinten gelegene Layer kleiner.

iii. Wie in Grafik 7 zu erkennen, ist tanh breiter verteilt, als ELU und RELU. Gerade bei letzterer ist dies deutlich zu sehen. tanh erzeugt also betragsmäßig größere Gradienten, wodurch *learning slowdown* weniger stark auftritt.

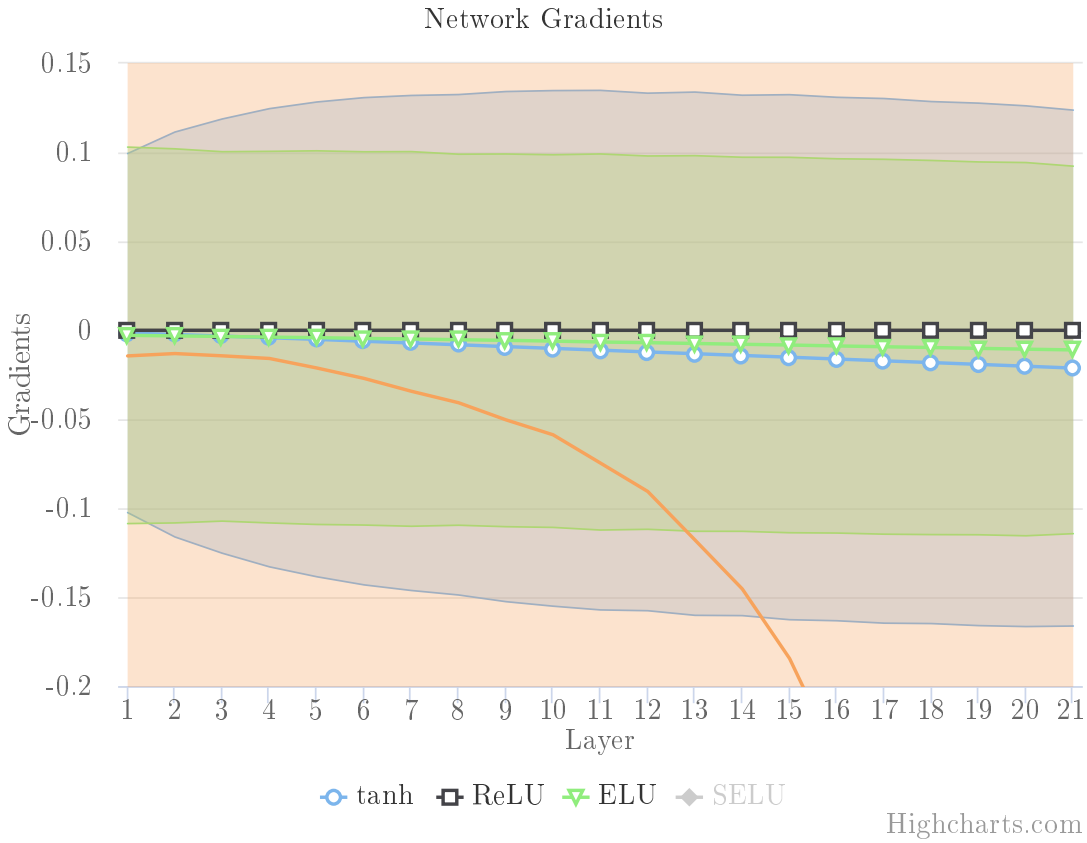

Abbildung 7

iv. In der ersten Epoche wird durch die SELU-Aktivierung am meisten gelernt. Der betragsmäßige Erwartungswert des Gradienten ist in allen Layer größer als bei den anderen Aktivierungsfunktionen, dadurch werden die Gewichte am stärksten adaptiert.

b) Die Histogramme der Aktivierungen mit tanh sind symmetrisch. Die Varianz nimmt für weiter hinten gelegene Schichten im Netz ab. Die ELU-Funktion verhält sich sehr ähnlich.

Für die ReLU-Aktivierung ist zu sehen, dass kein Layer einen negativen Ausgangswert hat. Im Vergleich zur tanh-Funktion ist die Varianz deutlich kleiner und nimmt über ab.

Die Histogramme der Aktivierungen mit SELU-Funktionen sind deutlich breiter verteilt. Die Verteilung ist nicht symmetrisch, bleibt aber über alle Layer quasi konstant.