

# Neuronale Netze

*Proseminar Data Mining*

Lukas Krenz  
Fakultät für Informatik  
Technische Universität München  
Email: lukas@krenz.land

**Zusammenfassung**—Das Paper ist eine kurze Einführung in Neuronale Netze, insbesondere in Multilayer Perceptrons inklusive Analyse ihrer praktischen Anwendung.

**Schlüsselworte**—Neuronale Netze, Multilayer Perceptrons, Backpropagation,

## I. EINLEITUNG

Die Familie der Künstlichen Neuronalen Netze(ANN) ist wohl aktuell eins der spannendsten Forschungsgebiete im Bereich des maschinellen Lernens. In der letzten Zeit ist es die Methode, die in vielen Bereichen die besten Ergebnisse erzielt:

- Clustering
- Computer Vision
- Natural Language Processing.

Auch, wenn sie auf den ersten Eindruck kompliziert und unverständlich wirken, ist die mathematische und statistische Grundlage simpel.

In Kapitel 2 wird das Perceptron als Inspiration für Neuronale Netze betrachtet; Schwerpunkt ist das Aufzeigen der Schwächen des Modells.

In Kapitel 3 wird die Theorie von Multilayer-Perceptrons beschrieben, inklusiver statistischen Grundlage und mathematischer Optimierung.

In Kapitel 4 werden Probleme und Lösungen aufgezeigt, die in der Praxis auftreten können.

## II. VOM LINEAREN MODELL ZUM NEURONALEN NETZ

### A. Das Perceptron als additives lineares Modell

Eines der klassischsten Modelle der Regressionsanalyse sind lineare Regression und die logistic Regression.

Es sind mathematisch simple Verfahren, die sogar analytisch lösbar sind - aber eben nicht immer ideale Ergebnisse erzielen: es gibt Probleme, die mit den Verfahren nicht lösbar sind.

Eine Weiterentwicklung ist das so genannte Perceptron. Es besteht aus einem Eingabe-Vektor  $x$ , einem Gewichtsvektor  $w$  und einem Bias-Wert  $b$ . Es ist ein binärer Klassifizierer, der das Ergebniss mit Hilfe der folgenden Formel berechnet:

$$y = \begin{cases} 1 & wx + b > 0 \\ 0 & \text{ansonsten} \end{cases} \quad (1)$$

To do (1)

### B. Das Perceptron als einfaches ANN

Wie wir später sehen werden, ist das Perceptron eine einfache Variante eines Feed-Forward-Neuronalen-Netzes mit der so genannten Heaviside-Funktion

$$\sigma = \begin{cases} 1 & n \geq 0 \\ 0 & n \leq 0 \end{cases} \quad (2)$$

als Aktivierungsfunktion.

Auf den benutzen Trainingsalgorithmus werden wir nicht eingehen; der später betrachtete Backpropagation-Algorithmus für allgemeine ANNs ist auf Perceptrons übertragbar.

### C. Das XOR-Problem: nicht linear separierbare Probleme

Es gibt viele Probleme, die von einem Perceptron nicht gelöst werden können. Es können nur Funktionen von linearen Modellen approximiert werden, die linear separierbar sind. Da das Perceptron als Verknüpfung von linearen Funktionen auch linear ist, kann auch es diese Klasse von Funktionen nicht lösen

### D. Die Lösung: Neuronale Netze

Neuronale Netze sind eine Verallgemeinerung von Perceptrons. Sie besitzen noch eine (oder mehrere) weitere Schicht(en) Neuronen, die jedoch nicht nur lineare Transformationen durchführen, sondern zusätzlich noch eine nicht lineare.

Diese nicht-linearität ermöglicht es, auch nicht linear separierbare Probleme zu lösen.

Durch die Nicht-linearität und durch die komplexere Verknüpfung resultiert ein nicht lineares Optimierungsproblem, das meistens nicht mehr analytisch zu lösen ist: es werden numerische Verfahren benutzt.

## III. NEURONALE NETZE

### A. Aufbau und Namenskonventionen

Ein Neuronales Netz besteht aus mehreren Schichten Neuronen. Einem Input Layer, einen (oder mehreren) Hidden Layer, und einem Output Layer. Wenn man von

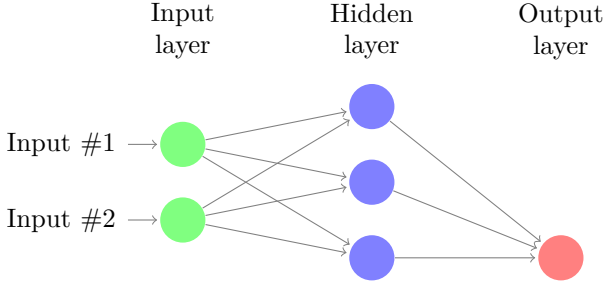


Abbildung 1. Ein 2-schichtiges MLP.

einem k-schichtigen ANN spricht, bezeichnet k die Anzahl der Schichten, ohne Input-Layer.

Sowohl Input, als auch Hidden-Layer bestehen aus Neuronen mit einer Aktivierungsfunktion, die die Eingangsdaten transformiert.

Jeder Layer ist mit den anderen Elementweise verbunden. Jede Verbindung hat ein so genanntes Gewicht. Daten fließen nur von links nach rechts.

Wir benutzen folgende Namenskonvention:  $m$  ist die Anzahl der Parameter,  $x \in \mathbb{R}^{m \times 1}$  ist der Input-Vektor,  $w_i \in \mathbb{R}^{\text{Anzahl Neuronen} \times \text{Anzahl Neuronen Layer davor}}$  der Gewichtsvektor, der die  $i$ -te Schicht mit der  $i + 1$  Schicht verbindet;  $b \in \mathbb{R}^{\text{Anzahl Neuronen} \times 1}$  der Bias-Vektor;  $\hat{y}$  ist der Output des Netzwerkes,  $y$  der Referenzwert (Ergebnis im Trainingsset).

### B. Die Aktivierungsfunktion

Die Funktion, die pro Knoten die nicht-lineare Transformation durchführt, heißt Aktivierungsfunktion. Häufig wird eine Funktion mit sigmoiden Erscheinungsbild gewählt, das heißt eine Funktion, die die Ergebnisse in ein bestimmtes Intervall "zusammenquetscht" (Quelle) und ein an ein "S" erinnerndes Erscheinungsbild hat. Die einfachste Funktion dieses Typs ist die so genannte Sigmoid-Funktion

$$\sigma_1(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

oder der hyperbolische Tangens

$$\sigma_2(x) = \tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} = \sigma_1(2x) - 1 \quad (4)$$

Die Sigmoid-Funktion ist nicht 0-zentriert. Das hat den Effekt, dass, falls alle Eingangswerte des Neurons, der Gradient entweder komplett positiv, oder negativ ist. Das führt unter Umständen zu unerwünschten Oszillationen bei der Optimierung. Der hyperbolische Tangens ist 0-zentriert, hat aber einen Nachteil mit der Sigmoid-Funktion gemeinsam: sehr kleine oder sehr große Werte haben einen sehr kleinen Gradienten zur Folge; dies führt zum "Absterben" des Neurons.

Eine alternative Aktivierungsfunktion ist die Rectifier-Funktion

$$\sigma_3(x) = \max(0, x). \quad (5)$$

Glorot et al argumentieren, dass die ReLU-Funktion zwar biologisch plausibler ist als der hyperbolische Tangens, aber die Rectifier-Funktion noch näher an der Funktionsweise biologischer Neuronen ist. Was für uns relevanter ist: sie ist effizienter zu berechnen; also in der Praxis oft eine bessere Wahl. [1]

### C. Die Kostenfunktion

Wir benutzen eine Kostenfunktion, die den durchschnittlichen Fehler des ANN-Ergebnis quantifiziert.

Da neuronale Netze als Modellierung der Wahrscheinlichkeit der Vorhersage, dass ein Output Vektor  $t$  für neue Eingabewerte  $x$  angesehen werden, kann man den Fehler bezüglich der bedingten Wahrscheinlichkeit  $p(t|x)$  darstellen. Um eine Fehlerfunktion zu bilden benutzt man nun die Maximum-Likelihood-Funktion, die Parameter einer Wahrscheinlichkeitsverteilung schätzt.

$$L = \prod_n p(x^n | t^n) = \prod_n p(t^n | x^n) (x^n) \quad (6)$$

Anstatt die Maximum-Likelihood-Funktion zu maximieren, minimieren wir die negative, logarithmische Maximum-Likelihood-Funktion. Das funktioniert, weil  $\ln$  eine monotone Funktion ist, es erleichtert uns außerdem mit großen Werten zu rechnen.

$$\begin{aligned} J &= -\ln L \\ &= -\sum_n \ln p(t^n | x^n) - \sum_n \ln p(x^n) \\ &= -\sum_n \ln p(t^n | x^n) \end{aligned} \quad (7)$$

Aus dieser allgemeinen Fehlerfunktion lassen sich spezielle Fehlerfunktionen ableiten, die in der Praxis benutzt werden [2]:

Wir wollen jetzt wieder den Fehler für  $p(t|k)$  darstellen. Wir nehmen an, dass die Variable  $t_k$  aus einer deterministischen Funktion, und normalverteiltem Rauschen (mit dem arithmetischen Mittel 0 und Varianz  $\sigma$  gebildet wird. Deswegen gilt:

$$p(t_k | x) = \frac{1}{(2\pi\sigma^2)^{0.5}} \exp\left(-\frac{\{y_k(x; w) - t_k\}^2}{2\sigma^2}\right) \quad (8)$$

Den Fehlerterm können wir jetzt einfach als Summe über alle Fehlerwahrscheinlichkeiten darstellen:

$$E = \frac{1}{2\sigma^2} \sum_{n=1}^N \sum_{k=1}^c \{y_k(x^n; w) - t_k^n\}^2 + Nc \ln \sigma + \frac{Nc}{2} \ln(2\pi) \quad (9)$$

Ignorieren wir nun alle Faktoren, die nicht von den Parametern des Netztes abhängen, erhalten wir

$$J(\theta) = \sum_{k=1}^K \sum_{i=1}^N \left( \hat{Y}_i - Y_i \right)^2, \quad (10)$$

was der Sum-of-squared-errors Fehlerfunktion entspricht. [2]

$$J(\theta) = \sum_{k=1}^K \sum_{i=1}^N \left( \hat{Y}_i - Y_i \right)^2 \quad (11)$$

$\hat{Y}_I$  entspricht dem Ergebniss des ANN,  $Y$  dem richtigen Ergebniss.

Bei Klassifizierungsproblemen wird üblicherweise die Sum-of-squared-errors Fehlerfunktion, oder aber die negative Kreuzentropie benutzt [3]:

$$J(\theta) = - \sum_{k=1}^K \sum_{i=1}^N y_{ik} \ln f_k(x_i) \quad (12)$$

Bei Klassifikationsproblem ist Cross entropy fast immer die bessere Wahl. To do (2)

#### D. Feedforward

Um ein ANN zur Vorhersage neuer Input-Vektoren zu benutzen, wird es im sogenannten Feedforward Modus betrieben. In diesem wird der Input Vektor zuerst den Neuronen des Input-Layers präsentiert. Alle anderen Neuronen des nächsten Layers bilden dann eine lineare Kombination aus den Neuronen ihrer Vorgängerschicht:

$$\text{net}_j = \sum_{i=1}^d x_i w_{j,i} + w_{j,0} = \sum_{i=0}^d x_i w_{j,i} = w_j^t x. \quad (13)$$

Nachdem diese lineare Transformation abgeschlossen ist, wird bei allen Neuronen (außer bei denen im Output-Layer) die Aktivierungsfunktion benutzt, die eine nicht-lineare Transformation durchführt:

$$y_j = \sigma(\text{net}_j). \quad (14)$$

Dieser Prozess wird so lange durchgeführt, bis die Daten das komplette Netz durchlaufen haben, und im Output-Layer ankommen.

[4]

#### E. Die Minimierung der Kostenfunktion - Backpropagation

Natürlich reicht es für maschinelles Lernen nicht aus, nur Vorhersagen zu treffen - das Modell muss sich auch anpassen können. Bei Neuronalen Netzen heißt der benutzte Trainingsalgorithmus Backpropagation.

Es ist eine Möglichkeit, Fehler, die bei den Ausgabeneuronen auftreten auch für die Neuronen der Schichten davor zu berücksichtigen - der Fehler wird zurückgesendet.

Dazu wird als erster Schritt für einen Trainingsvektor ein Forward-Pass durchgeführt - damit wird die Vorhersage des Netzwerkes berechnet, und anschließend mit der vorgegeben, korrekten Lösung des Trainingssets verglichen.

Diese Werte werden dann benutzt, um den Fehler zu berechnen, und dann die Parameter(also die Gewichte) anzupassen.

Am Ende wird das Ergebniss der Optimierungsmethode benutzt, um die Gewichte zu aktualisieren. [5]

Alle relevanten Fehlerfunktionen haben die Form

$$J = \sum_n J_n(y_1, \dots, y_c), \quad (15)$$

, das heißt, der Gesamtfehler ist die Summe aller Einzelfehler. Wir nehmen weiterhin an, dass der Fehler als ableitbare Funktion der Output-Werte darstellbar ist.

Wir suchen jetzt die partielle Ableitung der Fehlerfunktion unter Berücksichtigung eines bestimmten Gewichts:

$$\frac{\partial J}{\partial w_{i,j}} = \frac{\partial J^n}{\partial a_j} \frac{\partial a_j}{\partial w_{i,j}} \quad (16)$$

Durch mehrfaches Anwenden der Kettenregel bekommen wir die so genannte Backpropagation-Formel:

$$\theta_j = \begin{cases} \sigma'(\text{net}_j)(y_j - \hat{y}_j) & \text{wenn j Ausgabeneuron ist} \\ \sigma'(\text{net}_j) \sum_k w_{k,j} \theta_k & \text{j hidden Layer} \end{cases} \quad (17)$$

Dadurch können wir den Gradienten einfach rekursiv berechnen. [2]

Wichtig ist noch die Unterscheidung zwischen:

**Stochastic Backpropagation** Bei diesem Verfahren wird immer nur ein zufälliger Eingabevektor evaluiert, und dann werden sofort die Gewichte aktualisiert.

**Batch Backpropagation** Bei der zweiten Methode werden alle Eingabevektoren des Trainingssets präsentiert und die Gewichtsänderungen summiert. Erst dann werden alle Gewichte aktualisiert.

[4]

Die stochastische Methode ist die in der Praxis bevorzugte, weil die Konvergenz nicht nur von der Größe des Trainingsset, sondern vielmehr von der Anzahl der Iterationen und der Verteilung der Trainingsdaten - für große Sets ist ein Benutzen der Batch Methode nicht mehr sinnvoll. [6]

#### F. Die numerische Methode - Gradient Descent und verwandte

Der Gradient wird oft mit dem Gradientenverfahren, einer Methode um Funktionen zu minimieren, benutzt.

Er ist relativ simpel:

$$x_{n+1} = x_n - \gamma_n \nabla F(x_n) \quad (18)$$

Das Verfahren funktioniert, weil der negative Gradient in die Richtung zeigt, in der die Funktion am schnellsten kleiner wird. Daher kommt auch die Bezeichnung Verfahren des steilsten Abstiegs. Die Konstante  $\nabla$  heißt Lernrate. Je höher Sie ist, desto schneller konvergiert das Verfahren - bei einer zu hoch gewählten sind aber oft Oszillationen zu beobachten.

Auf ANNs angewandt, sehen die Formeln dann folgendermaßen aus:

$$\Delta w_{j,i} = -\eta \delta_j x_i \quad (19)$$

$$\Delta w_{j,i} = -\eta \sum_n \delta_j^n x_i^n \quad (20)$$

[2]

Man unterscheidet beim Training von ANNs zwischen Batch und Stochastic Gradient Descent.

Bei der Batch Methode werden alle Gewichte auf einmal aktualisiert, bei Stochastic GD nur einzelne, zufällig ausgewählte Gewichte.

Das Gradientenverfahren hat das Problem, dass es oft in lokalen Minima steckenbleibt - das ist vor allem bei ANNs problematisch, weil die gewählte Kostenfunktion in den meisten Fällen mehrere lokale Minima besitzt. Auch wenn es nicht unbedingt notwendig, oder sogar erwünscht ist, einen globalen Extremwert zu finden (vgl. Overfitting), existieren bessere numerische Verfahren.

Eine Verbesserung zum gewöhnlichen Gradientenverfahren ist es, einen so genannten Momentumparameter zu benutzen, der auch vorherige Ergebnisse einbezieht, und somit das oft beobachtete oszillierende Verhalten des Gradientenverfahrens behebt.

$$\mathbf{b} = \mathbf{a} - \gamma \nabla F(\mathbf{a}) + \text{Momentum} \quad (21)$$

#### IV. NEURONALE NETZE IN DER PRAXIS

Weight Decay, Anfangswert

##### A. Das Gradientenverfahren in der Praxis

Beim Gradientenverfahren ist es wichtig eine sinnvolle Schrittweite zu verwenden. Ist sie kleiner, als die optimale, ist die benötigte Rechenzeit höher. Ist sie größer, kann das Verfahren divergieren. Eine gute Heuristik für die Wahl des Parameters ist, mit einem großen Wert zu starten, und bei Divergenz es mit einer dreimal kleineren Schrittweite erneut zu probieren. Eine andere gute Optimierung ist, weder Batch noch Stochastic Backpropagation zu nutzen, sondern eine Mini-batch-Variante, bei der mehrere, aber nicht alle Eingabevektoren evaluiert werden vor jedem Gewichtsupdate. [6]

Die Kostenfunktion hat oft mehrere lokale Minima - das Gradientenverfahren kann dann in einem solchen steckenbleiben, das heißt, es konvergiert unter Umständen nicht

zum gesuchten globalen Minimum. Auch wenn das globale Minimum nicht immer erwünscht ist (vgl. early stopping), ist es unter Umständen doch besser, es zu finden.

Die einfachste Lösung wäre natürlich, ein anderes numerisches Verfahren zu benutzen. Doch andere Verfahren sind meistens komplizierter zu implementieren, und unter Umständen sogar langsamer als SGD. Wenn man beim SGD in einem Minimum stecken bleibt, bietet es sich an, einfach mit neu initialisierten Gewichtsparametern neuzustarten.

##### B. Anzahl Hidden Layer

##### C. Initialisierung der Gewichte

Die Initialisierung der Gewichte ist ein schwieriges Problem. Eine oft benutzte Heuristik ist

$$W_{ij} \sim U\left[-\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}\right], \quad (22)$$

wobei  $U[-a, a]$  gleichverteilten Werten zwischen  $-a$  und  $a$  entspricht, und  $n$  der Anzahl der Neuronen in der vorherigen Schicht.

Ein Vorschlag, um dieses Problem zu umgehen ist die normalisierte Initialisierung

$$W \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}, \frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}\right]. \quad (23)$$

In der Praxis ergeben sich durch sie bessere Ergebnisse. [7]

##### D. Vermeidung von Overfitting

Ein Problem, dass in der Praxis auftreten kann, ist das so genannte Overfitting, das ist die übermäßige Anpassung an die Trainingsmenge. Oft ist das Modell zu stark angepasst und daher nicht mehr allgemein effizient.

Eine mögliche Lösung ist die L1/L2-Regularisierung. Bei ihr wird an die Kostenfunktion ein weiterer Term angehängt:

$$J_{L1} = J + \lambda \sum_i \theta_i \quad (24)$$

$$J_{L2} = J + \lambda \sum_i \theta_i^2 \quad (25)$$

Er skaliert mit der Summe aller Gewichte, es werden also komplexe Modelle bestraft. Der Name Weight-Decay kommt daher, dass die Gewichte gegen 0 gehen. [3]

##### To do (3)

Aus der Bayesianischen Perspektive betrachtet, ist eine L2-Regulierung gleichbedeutend damit, dass wir eine A-priori-Verteilung bezüglich der Parameter annehmen. Die Fehlerfunktion entspricht dann:

$$J(z, \theta) = -\ln P(z|\theta). \quad (26)$$

[6]

Eine weitere Lösung ist das so genannte early stopping, bei dem man nicht mit der Backpropagation aufhört, wenn ein (lokales) Minimum gefunden wurde, sondern dann, wenn die Performance des Modells bei dem Validierungsset optimal ist. Dabei wird oft die Iteration oft beendet, wenn bei der Fehlerfunktion gar kein Minimum vorliegt. **To do** (4)

Dabei entspricht die L2-Regularisierung early stopping, es sollten also nicht beide Methoden gleichzeitig genutzt werden; L1-Regularisierung kann jedoch mit beiden Regularisierungsarten kombiniert werden. [6]

#### E. Interpretation des Output-Layers

Bei der Regression gibt es üblicherweise nur einen Knoten im letzten Layer, ohne Aktivierungsfunktion (bzw. mit  $\sigma(x) = x$ ). Der Wert dieses Neurons ist das Ergebniss der Regression.

Bei der Klassifikation gibt es für jede mögliche Klasse einen Austrittsknoten, und es muss folgendes gelten:

$$\forall k : y_k \in [0, 1] \quad (27)$$

und

$$\sum_k^n y_k = 1. \quad (28)$$

Das heißt, sie sind direkt als Wahrscheinlichkeiten interpretierbar. Eine mögliche Wahl für die Aktivierungsfunktion ist die SoftMax-Funktion, eine Verallgemeinerung der logistischen Funktion:

$$\sigma_c = \frac{\exp(a_k)}{\sum_{k'} \exp(a_{k'})}. \quad (29)$$

Sie hat für die Wahrscheinlichste Klasse den Wert 1, für alle anderen den Wert 0. [2]

### V. ZUSAMMENFASSUNG UND AUSBLICK

Wir haben nur MLPs betrachtet, in neueren Forschungsergebnissen werden oft ANNs mit anderer Topologie benutzt, das heißt, die Daten fließen nur in eine Richtung. Relevant sind aber auch andere Netzwerkarten wie RNNs, RBM, etc. Sie besitzen einen komplizierteren Aufbau; aber im Grunde sind die meisten genannten Grundlagen auch für diese Arten von Netzwerken relevant. Die Zukunft bleibt spannend, ANNs sind sehr vielversprechend.

TO DO...

- ☐ 1 (p. 1): Quelle: Perceptron
- ☐ 2 (p. 3): Warum ist Cross entropy besser?
- ☐ 3 (p. 4): Ja, genau. Die L2 ist halt direkt als Gaussian-prior interpretierbar. Zusammen mit der per NLL-Fehlerfunktion als Likelihood entspricht das dann eben einem Maximum A Posteriori-Ansatz. Das macht das ganze dann zB auch mit Gaussian Processes vergleichbar und bettet das halt einfach in ein solides stochastisches Framework ein.
- ☐ 4 (p. 5): Quelle early stopping!

#### LITERATUR

- [1] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, vol. 15, 2011, pp. 315–323.
- [2] C. M. Bishop, *Neural networks for pattern recognition*. Clarendon press Oxford, 1995.
- [3] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.
- [4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.
- [5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, 1988.
- [6] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 437–478.
- [7] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010, pp. 249–256.