

## Einführende Texte

Wir möchten in dieser Studie eine Reihe von Verfahren vorstellen, die sich zur statistischen Datenauswertung verwenden lassen. Während dir die „lineare Regression“ und ähnliche Verfahren vermutlich im Zuge deiner Methodenausbildung bereits näher gebracht wurden, wollen wir hier mit „Neuronalen Netzen“ auch eine Auswertungsmöglichkeit vorstellen, die dir vielleicht weniger geläufig ist.

Unter Neuronalen Netzen versteht man mathematische Modelle von Systemen aus Neuronen („Units“) und den sie verbindenden Synapsen („Koeffizienten“). Man unterscheidet unter den Units meist zwischen „Inputunits“ –über die z.B. Prädiktorvariablen eingespeist werden können- und „Outputunits“ –an denen sich die errechnete Reaktion des Netzes auf den Input ablesen lässt.



(Rechts sieht man ein Neuronales Netz mit zwei grünen Inputunits und einer roten Outputunit.)

Man kann Neuronale Netze –genau wie die lineare Regression- zur Vorhersage von Variablen nutzen. Z.B. könnte man das Alter einer Person als Input einspeisen, und das Netz so einstellen, dass die Aktivität der Outputunit möglichst genau eine andere Variable –z.B. die Leistung in einem Gedächtnistest- vorhersagt).

**Wichtig:** Für alle hier gezeigten Verfahren gilt:  
Je weniger gut die Daten mit einem Verfahren vorhergesagt werden können, desto größer ist der **Vorhersagefehler RMSE** („root mean squared error“).

Im Folgenden möchten wir einige ausgewählte Verfahren genauer vorstellen:

### - lineare Regression

Die lineare Regressionsanalyse ist ein Verfahren, mit dem man einen *linearen Zusammenhang* zwischen einem oder mehreren vorhersagenden Variablen (auch „Prädiktoren“ oder „UVs“ genannt) und einer vorhergesagten Variable (auch „Kriterium“ oder „AV“ genannt) beschreiben kann.

Zum Beispiel könnte man sich fragen, ob sich die Gedächtnisleistung vorhersagen lässt, wenn man das Alter kennt. Abbildung 1 zeigt hierzu fiktive Daten: Das Alter –der Prädiktor- ist an der x-Achse angetragen, die Fehler im Gedächtnistest –das Kriterium- an der y-Achse.

Man sieht einen deutlichen Zusammenhang:  
Je höher das Alter, desto häufiger die Fehler im Gedächtnistest.  
Dieser Zusammenhang lässt sich optimal durch die eingezeichnete Gerade beschreiben.

Eine Geradengleichung hat die Form  $y = b_1 \cdot x + b_0$   
Wobei  $b_1$  die Steigung der Geraden angibt,  
und  $b_0$  den y-Achsenabschnitt.

In der linearen Regression werden diese beiden *Koeffizienten*  $b_1$  und  $b_0$  so bestimmt, dass die Gerade optimal zu den Daten passt, bzw. optimal in der Punktwolke liegt. Wie diese Schätzmethode genau funktioniert soll hier nicht näher besprochen werden.

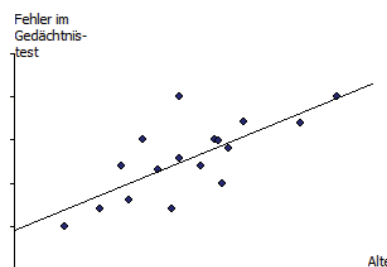


Abb. 1: Lineare Regressionsgerade mit positiver Steigung.

*Wichtig ist nur*, dass man mit dem Verfahren „Lineare Regression“ eine *lineare Funktion* erhält, mit der man aus einem Prädiktor (z.B. Alter) ein Kriterium (z.B. Gedächtnisleistung) vorhersagen kann.

Eine solche lineare Regressionsgleichung könnte z.B. lauten:

$$\text{Gedächtnisleistung} = 1,5 \cdot \text{Alter} + 2$$

### - polynomiale Regression

Natürlich kann es vorkommen, dass sich ein Zusammenhang nicht treffend über eine einfache Gerade beschreiben lässt.

Nehmen wir z.B. an, der Zusammenhang zwischen Alter und Gedächtnisleistung sähe eher aus wie in Abbildung 2 eingezeichnet. Demnach werden vor allem von jüngeren (z.B. kleinen Kindern) und älteren Menschen viele Fehler in einem Gedächtnistest begangen, während Menschen mittleren Alters (z.B. 20- bis 50-Jährige) die besten Leistungen im Gedächtnistest erzielen.

Ganz offensichtlich kann man so einen Zusammenhang nur sehr schlecht mit Hilfe einer Geradengleichung beschreiben. Statt einer linearen Funktion bietet sich hier ein Polynom höheren Grades an, z.B. ein Polynom zweiten Grades ( $y = x^2$ ).

Auch komplexere Zusammenhänge lassen sich über Polynom-Gleichungen beschreiben. Eine polynomiale Funktion hat die Form  $y = b_0 + b_1 \cdot x + \dots + b_n \cdot x^n$ .

Wie das Schätzen der optimalen *Koeffizienten*  $b_0, b_1, \dots, b_n$  im Detail aussieht soll auch hier wieder nicht näher besprochen werden.

*Wichtig ist nur*, dass man mit dem Verfahren „Polynomiale Regression“ eine *nicht-lineare Funktion* erhält, mit der man aus einem Prädiktor (z.B. Alter) ein Kriterium (z.B. Gedächtnisleistung) vorhersagen kann.

Die Regressionsgleichung einer polynomialen Regression könnte z.B. lauten

$$\text{Gedächtnisleistung} = 1,5 \cdot \text{Alter}^2 + 1,3 \cdot \text{Alter} + 2$$

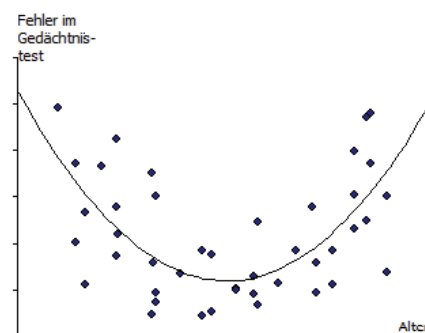
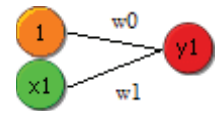


Abb. 2: Quadratische Regressionsgleichung

## - Neuronales Netz ohne Hiddenunits

Auch mit künstlichen Neuronalen Netzen lässt sich ermitteln, ob/wie zwei Variablen miteinander zusammenhängen. Rechts sieht man ein einfaches Neuronales Netz mit *einer Inputunit* (die dem Prädiktor bzw der UV entspricht), *einer Biasunit* (die in der folgenden Gleichung einer Konstanten entspricht), und *einer Outputunit* (die dem Kriterium bzw der AV entspricht).



Die Ausgabe der Outputunit errechnet sich im einfachsten Fall über folgende Geraden-Gleichung:

$$y = w_1 * x + w_0$$

Auch bei Neuronalen Netzen werden über ein mathematisches Verfahren die optimalen Koeffizienten  $w_1$  und  $w_0$  bestimmt, so dass die resultierende Gerade bestmöglich zu den Daten passt, bzw optimal in der Punktwolke liegt..

In diesem Fall sind die Ergebnisse eines Neuronalen Netzes mit jenen einer Regression identisch (siehe Abbildung 1).

Bei Neuronalen Netzen gibt es jedoch eine Besonderheit:

Über die sog. „Aktivitätsfunktion“ der Outputunit, kann man festlegen, welcher Art der gesuchte Zusammenhang zwischen Prädiktor und Kriterium sein soll.

Angenommen ein annähernd linearer Zusammenhang zwischen Alter und Gedächtnisleistung bestünde nur in einer bestimmten Lebensspanne, und sowohl davor als auch danach sei Altern nicht mehr mit weiterer Verbesserung oder Verschlechterung verbunden –z.B. weil dann ein Maximum/Minimum der Gedächtnisleistung erreicht ist.

Ein solcher Zusammenhang ließe sich durch eine S-förmige „logistische“ Kurve beschreiben (siehe Abbildung 3), nicht jedoch durch eine Gerade. In diesem Fall könnte man eine logistische Kurve als Aktivitätsfunktion wählen. Die Ausgabe der Outputunit lässt sich dann über folgende –leicht andere- Gleichung errechnen:

$$y = \text{logistic}(w_1 * x + w_0)$$

Es wird in diesem Fall also nicht mehr die optimale Gerade gesucht, sondern die optimale logistische Kurve, die bestmöglich zu den Daten passt, bzw die optimal in der Punktwolke liegt. Wie genau der Schätzmethode für die Koeffizienten  $w_1$  und  $w_0$  abläuft soll an dieser Stelle nicht näher besprochen werden.

*Wichtig ist nur*, dass man mit dem Verfahren „Neuronales Netz ohne Hiddenunits“ eine *lineare oder eine nonlineare Funktion* erhalten kann –je nachdem welche Funktion man als Aktivitätsfunktion der Outputunit wählt-, mit der man aus einem Prädiktor (z.B. Alter) ein Kriterium (z.B. Gedächtnisleistung) vorhersagen kann, z.B.:

$$\text{Gedächtnisleistung} = \text{logistic}(1,5 * \text{Alter} + 2)$$

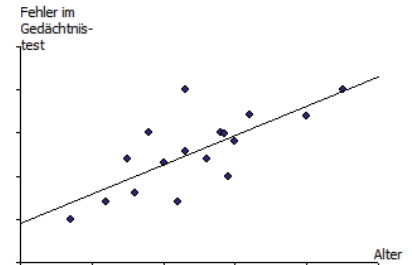


Abb. 1: Lineare Regressionsgerade mit positiver Steigung.

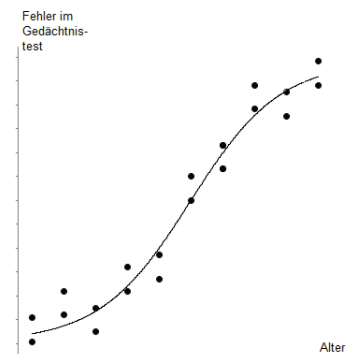


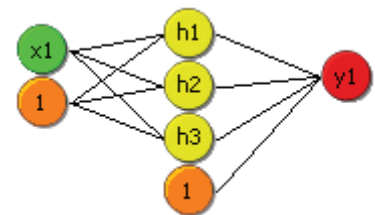
Abb.3: Logistische Regressionskurve

## - Neuronales Netz mit Hiddenunits

Ein Verfahren zum Auffinden besonders komplexer Zusammenhänge stellt ein Neuronales Netz mit sog. „Hiddenunits“ dar. Hiddenunits sind Einheiten, die sich zwischen Input- und Outputunits befinden (in der Abbildung rechts die drei gelben Einheiten zwischen Inputunits und Outputunit).

Durch Hiddenunits wird das Netz sehr flexibel und kann unterschiedlichste Zusammenhänge aufdecken – je mehr Hiddenunits man wählt, desto komplexer kann der gefundene Zusammenhang werden.

Prinzipiell lassen sich durch Neuronale Netze mit Hiddenunits beliebige Zusammenhänge aufdecken.



Will man die Ausgabe eines Neuronalen Netzes mit Hiddenunits als Gleichung darstellen, sieht das zunächst einmal sehr unübersichtlich aus, z.B. so:

$$y = w_1 * \text{logistic}(w_{10} * x + w_{00}) + w_2 * \text{logistic}(w_{11} * x + w_{01}) + w_3 * \text{logistic}(w_{12} * x + w_{02}) + w_4$$

Jedoch sollte man sich durch die Gleichung nicht abschrecken lassen.

Wenn man diese Funktion in seinen Datensatz einzeichnet, zeigt sich wie gut die komplexe Funktion den Zusammenhang beschreiben kann (siehe z.B. Abbildung 4).

Auch bei diesem Verfahren soll nicht näher darauf eingegangen werden, wie die Schätzung der Koeffizienten vor sich geht.

*Wichtig ist nur*, dass man mit dem Verfahren „Neuronales Netz mit Hiddenunits“ eine *beliebig komplexe idR nicht-lineare Funktion* erhalten kann, mit der man aus einem Prädiktor (z.B. Alter) ein Kriterium (z.B. Gedächtnisleistung) sehr exakt vorhersagen kann, z.B.:

$$\text{Gedächtnisleistung} = 2,2 * \text{logistic}(1,5 * \text{Alter} + 1,3) + 2,1 * \text{logistic}(1,7 * \text{Alter} + 1,4)$$

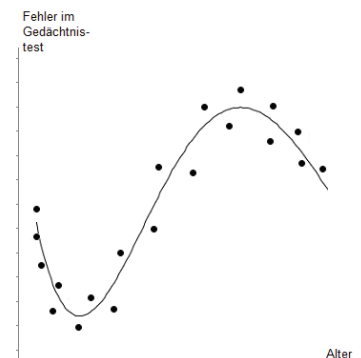


Abb. 4: Nonlineare Regressionskurve