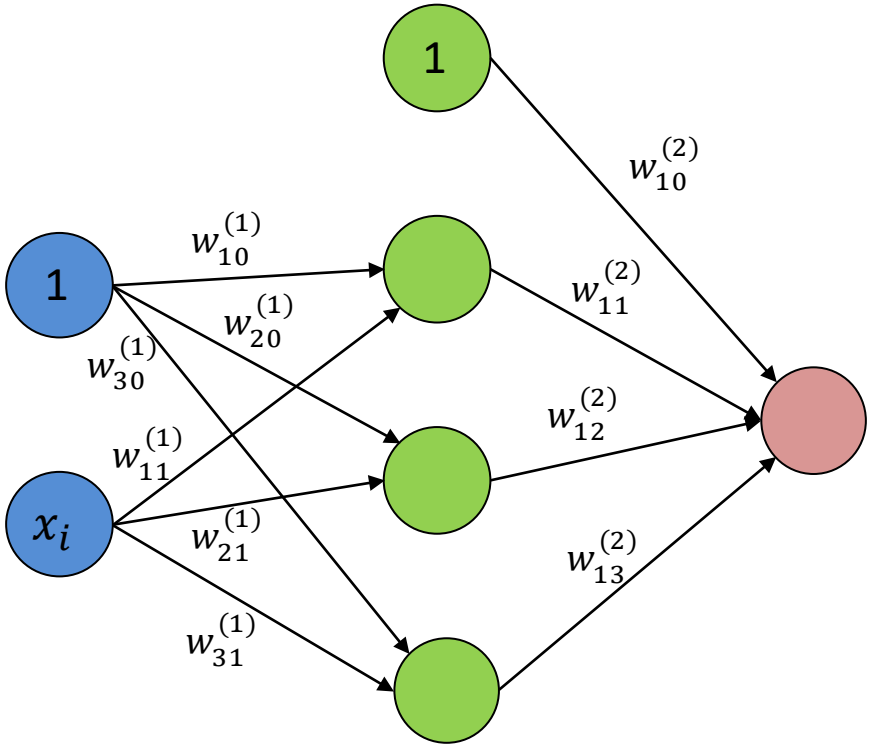
Artificial Neural Network (ANN)

Einfaches Beispiel

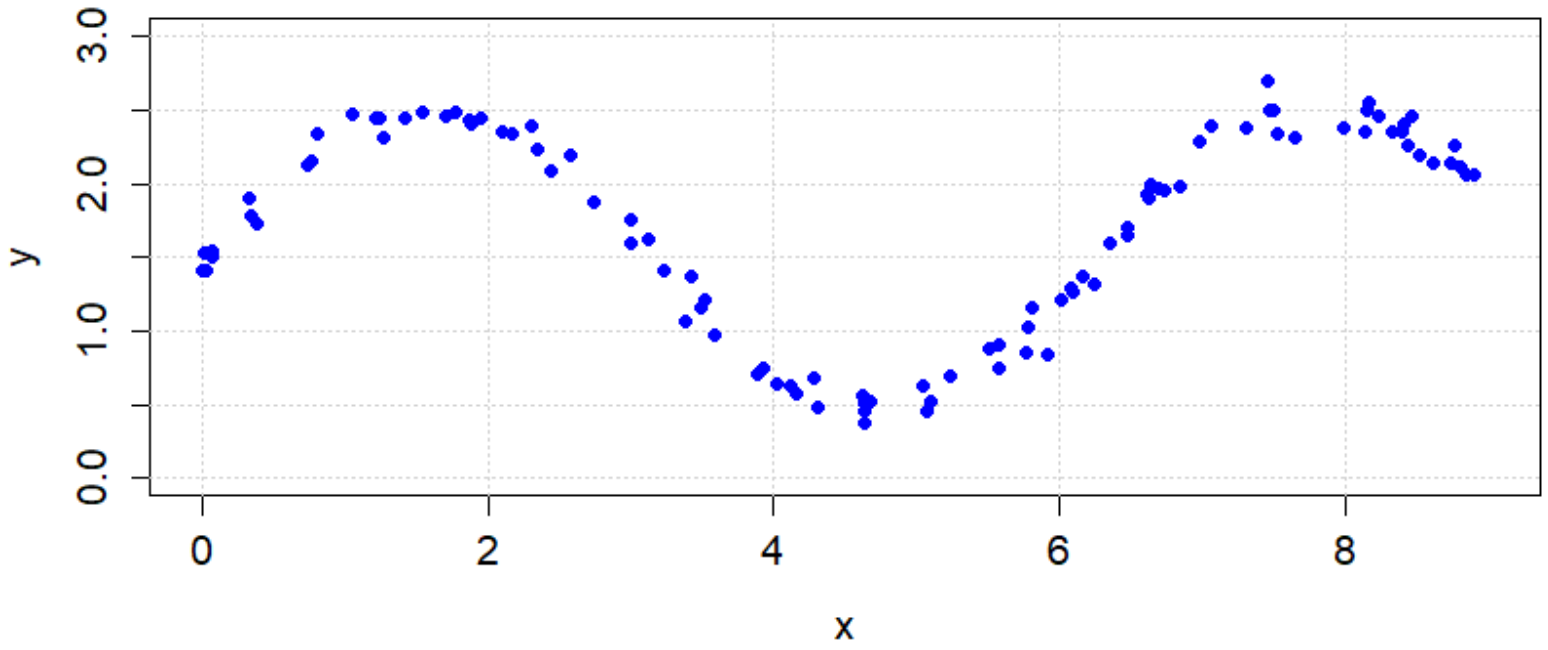


Wie rechnet ein ANN Vorhersagen?

- Ausgangslage
- Architektur ANN
- Schrittweises Vorgehen bei der Berechnung einer Vorhersage
- Vorhersagen im Streudiagramm
- Aktivierungen

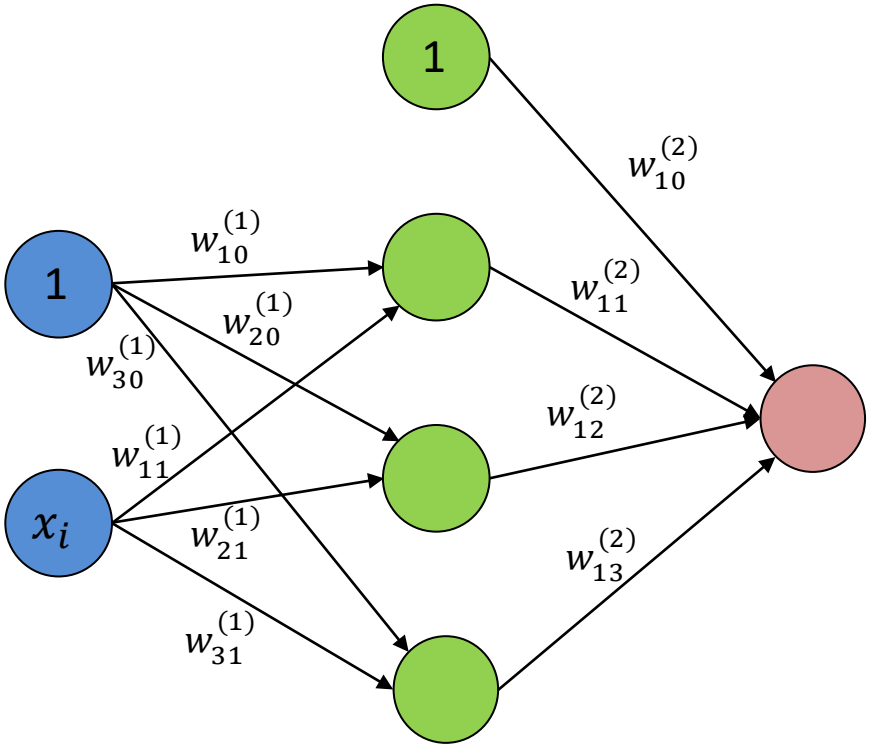
Ausgangslage

- Wir schauen ein **einfaches Regressionsproblem** an:
  - Eine quantitative Zielvariable y_i
  - Eine quantitative Input-Variable x_i
- Der wahre Zusammenhang ist stark **nicht-linear** (Sinus-Kurve).



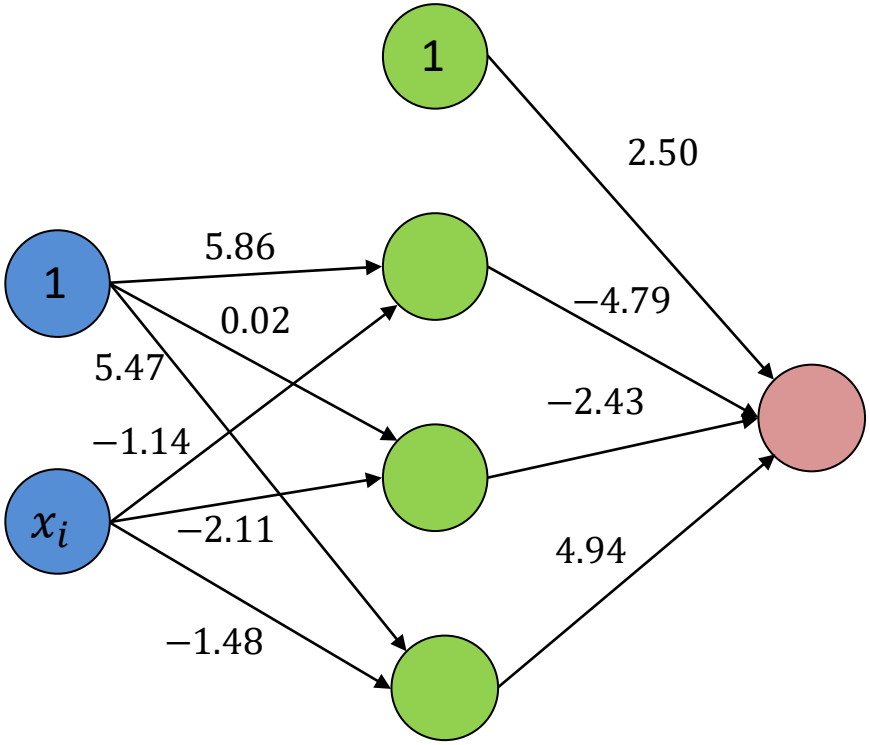
Architektur ANN

- Wir wollen dieses Regressionsproblem mit folgender **ANN Architektur** lösen:

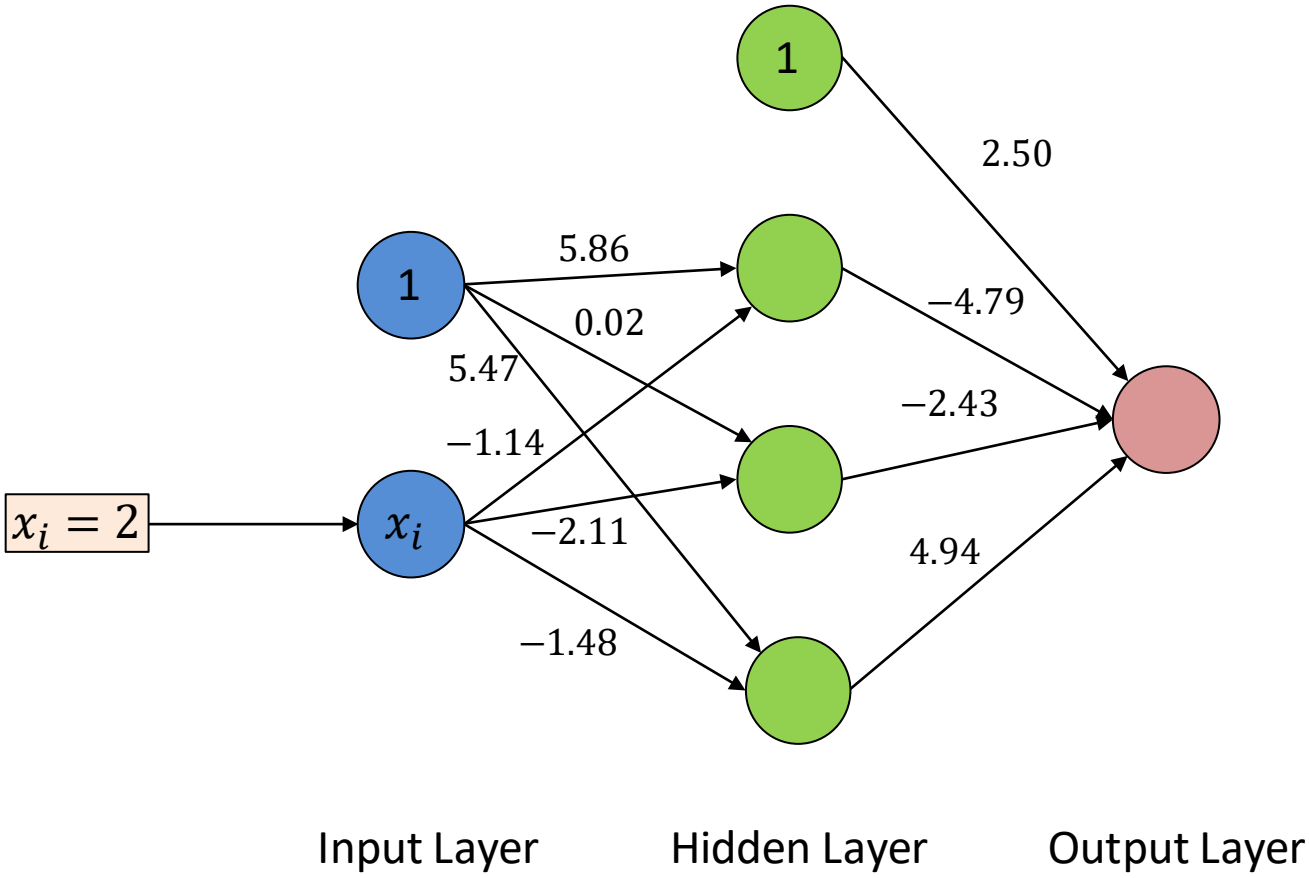


Optimale Gewichte sind gegeben

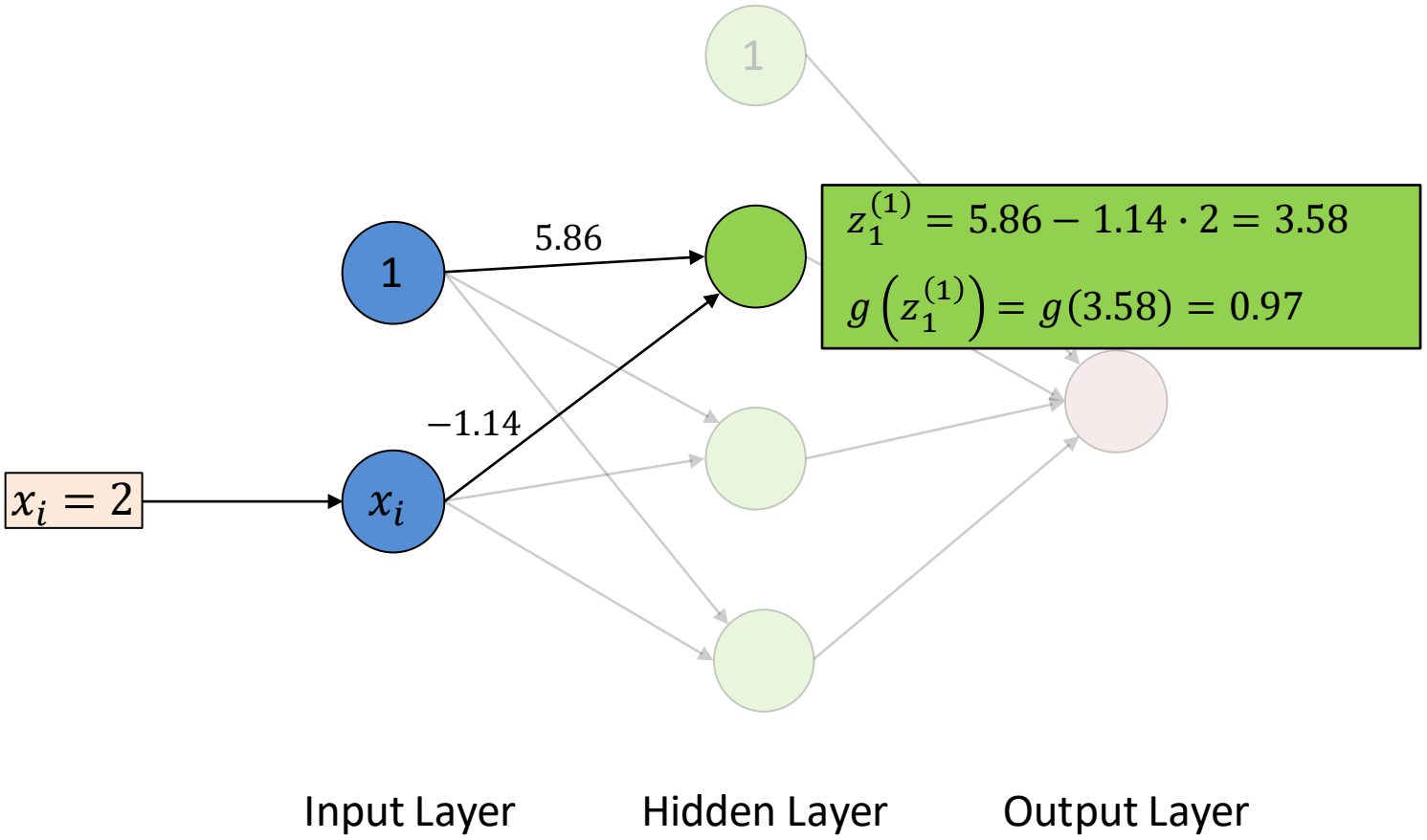
- Wir nehmen in diesem ersten Kapitel an, dass alle **Gewichte** (Parameter) des ANN bereits optimiert wurden und **gegeben** sind:



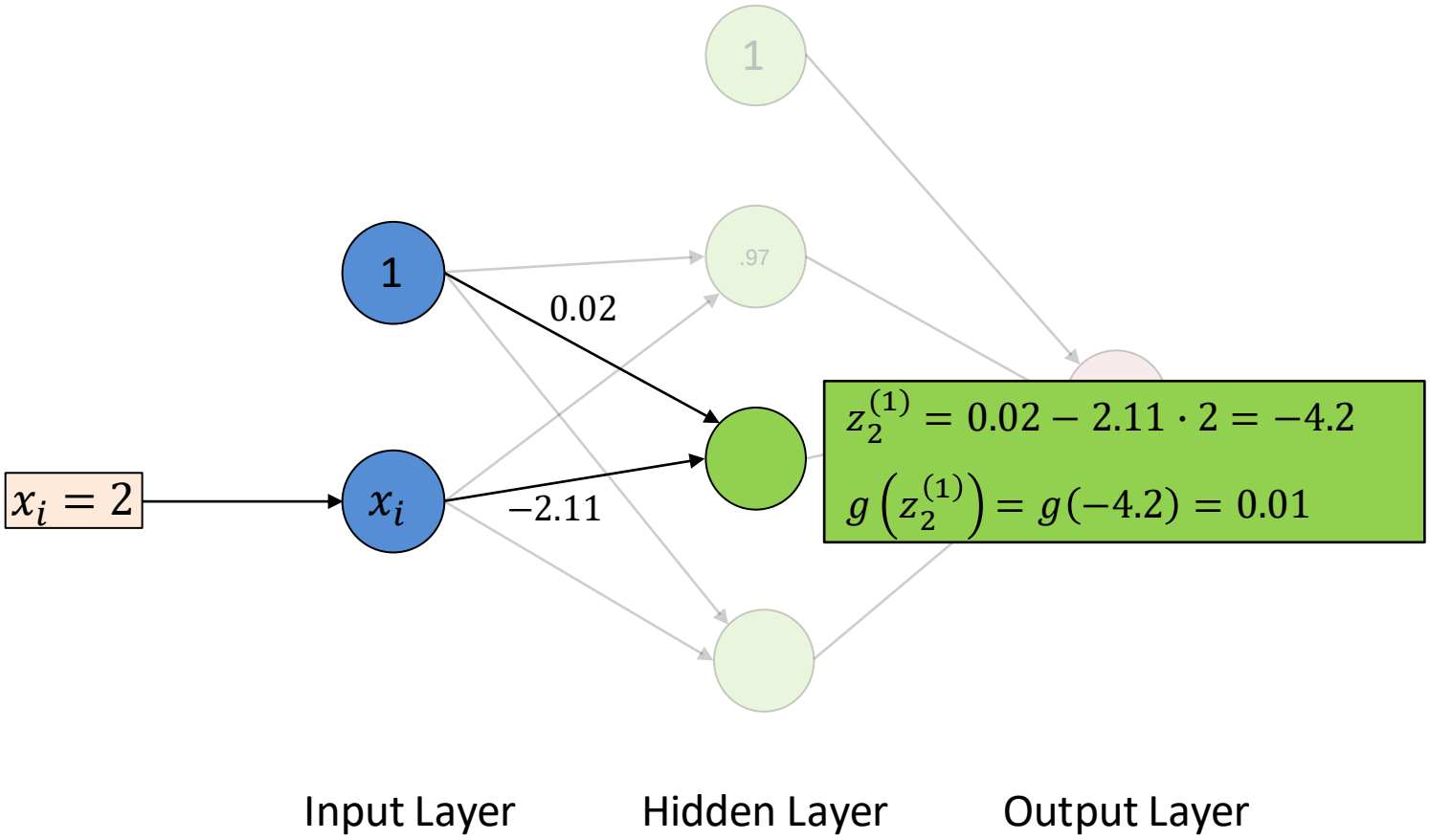
Vorhersage für $x_i = 2$ (Schritt 1)



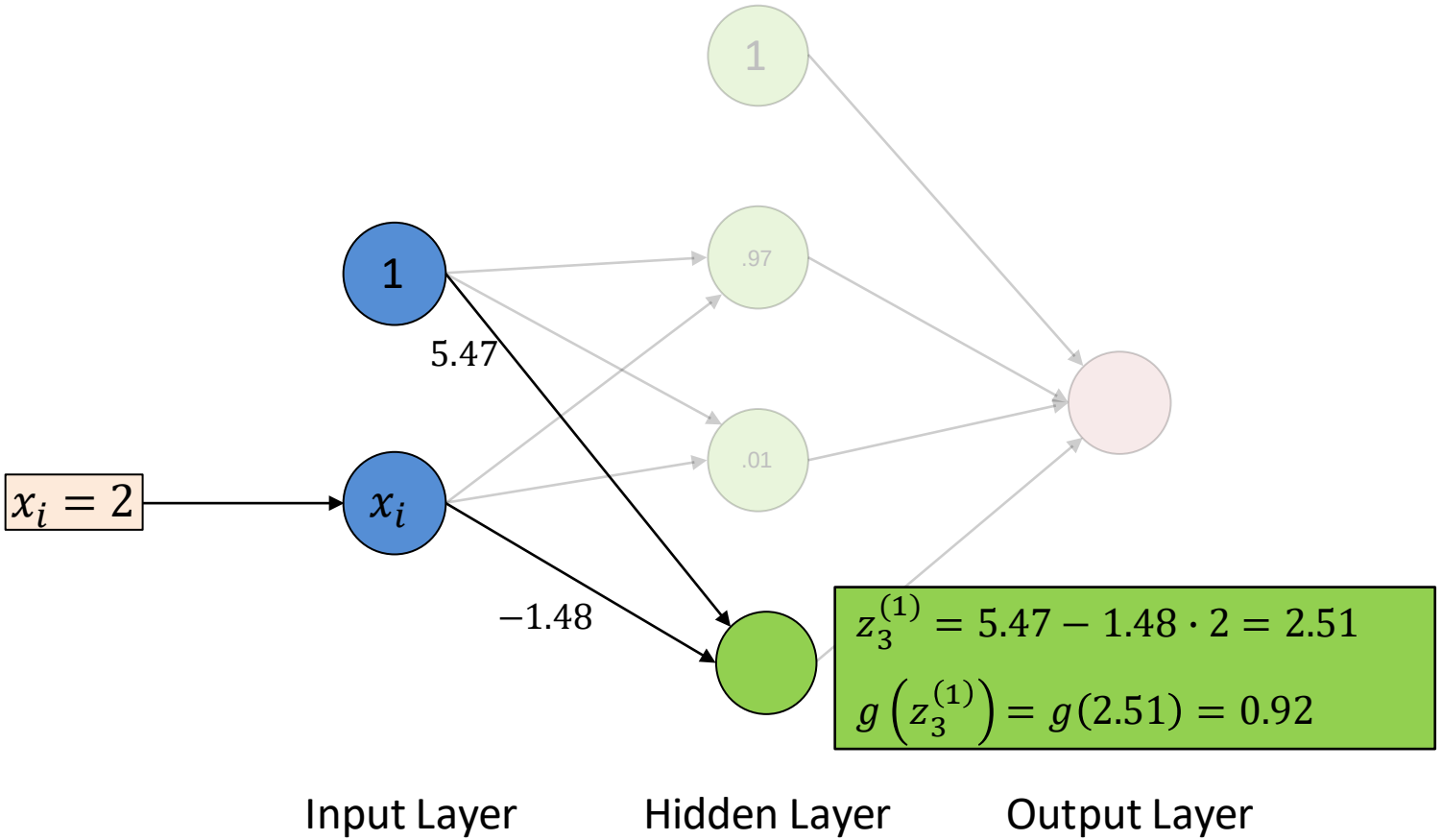
Vorhersage für $x_i = 2$ (Schritt 2)



Vorhersage für $x_i = 2$ (Schritt 3)

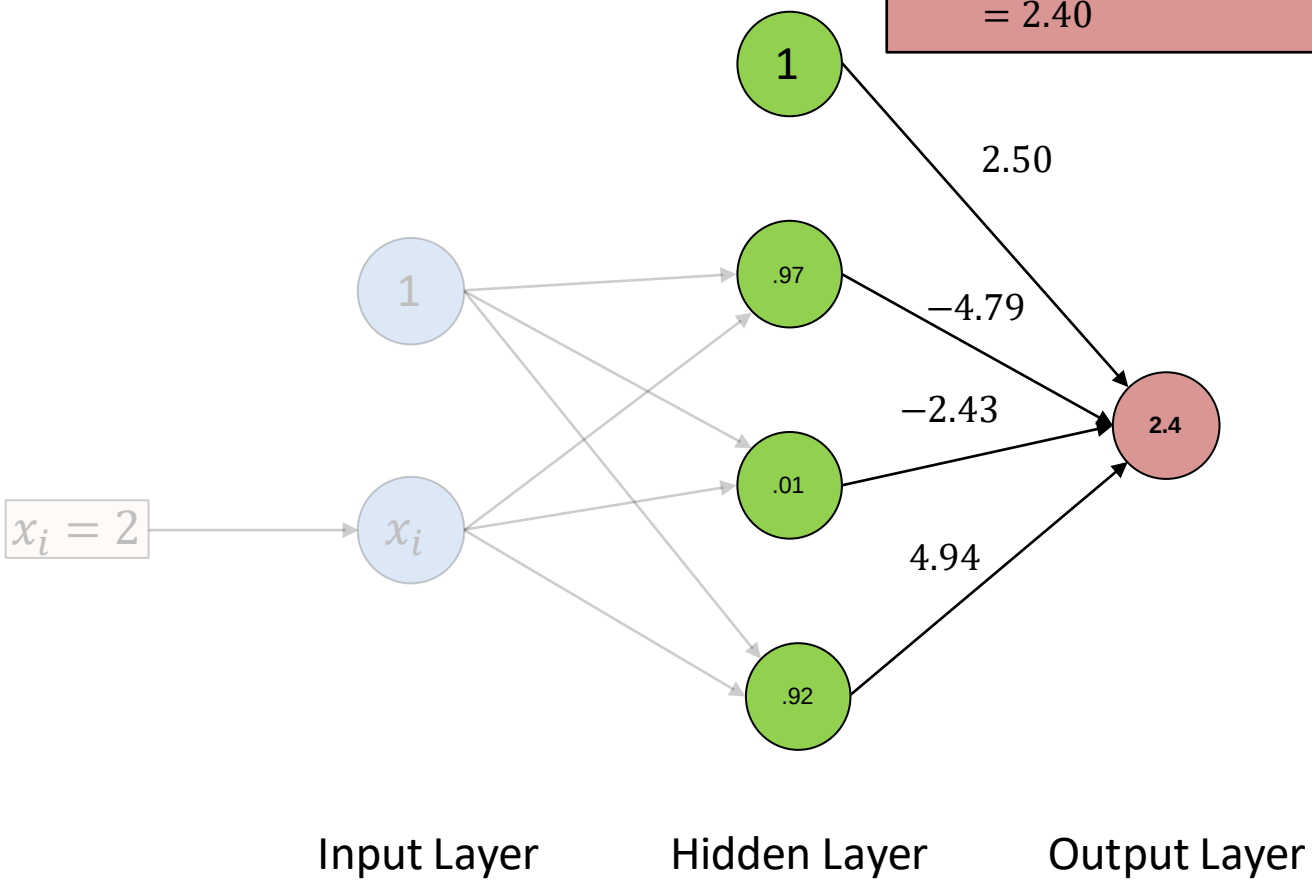


Vorhersage für $x_i = 2$ (Schritt 4)



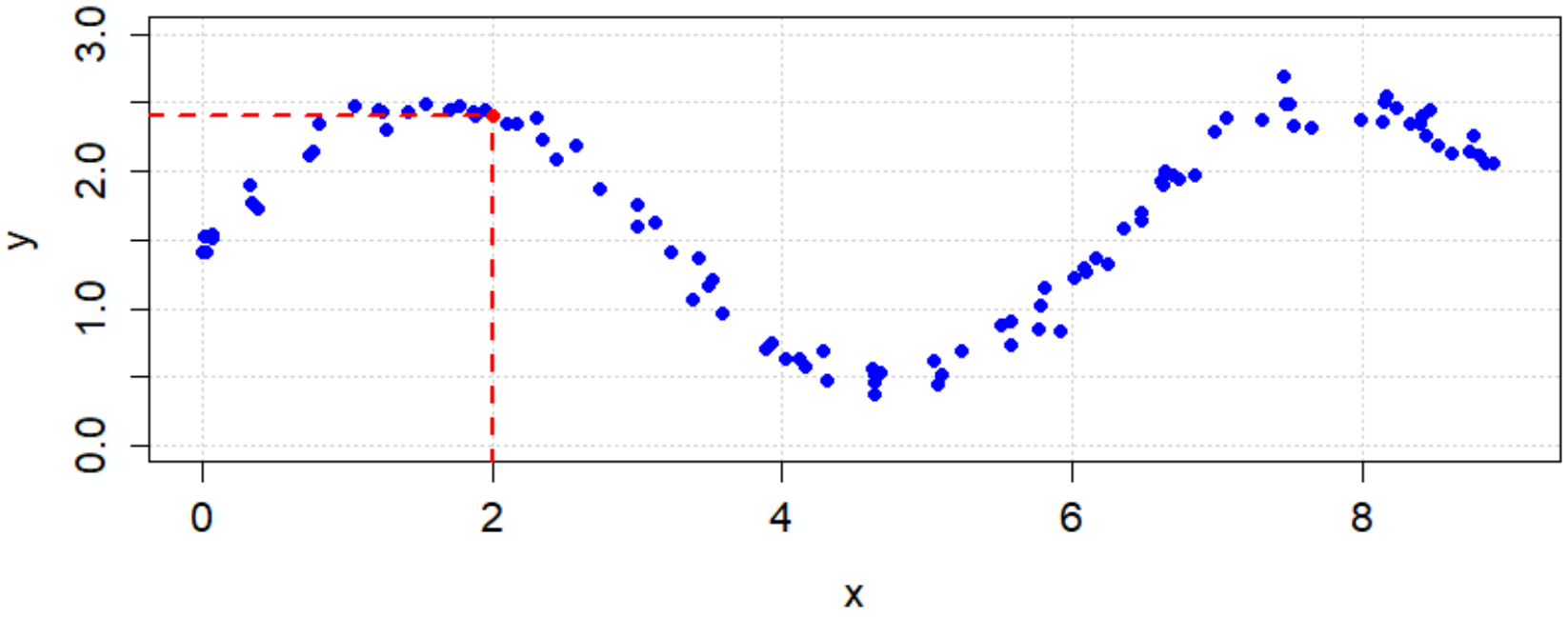
Vorhersage für $x_i = 2$ (Schritt 5)

$$z_1^{(2)} = 2.50 - 4.79 \cdot 0.97 - 2.43 \cdot 0.01 + 4.97 \cdot 0.92 = 2.40$$



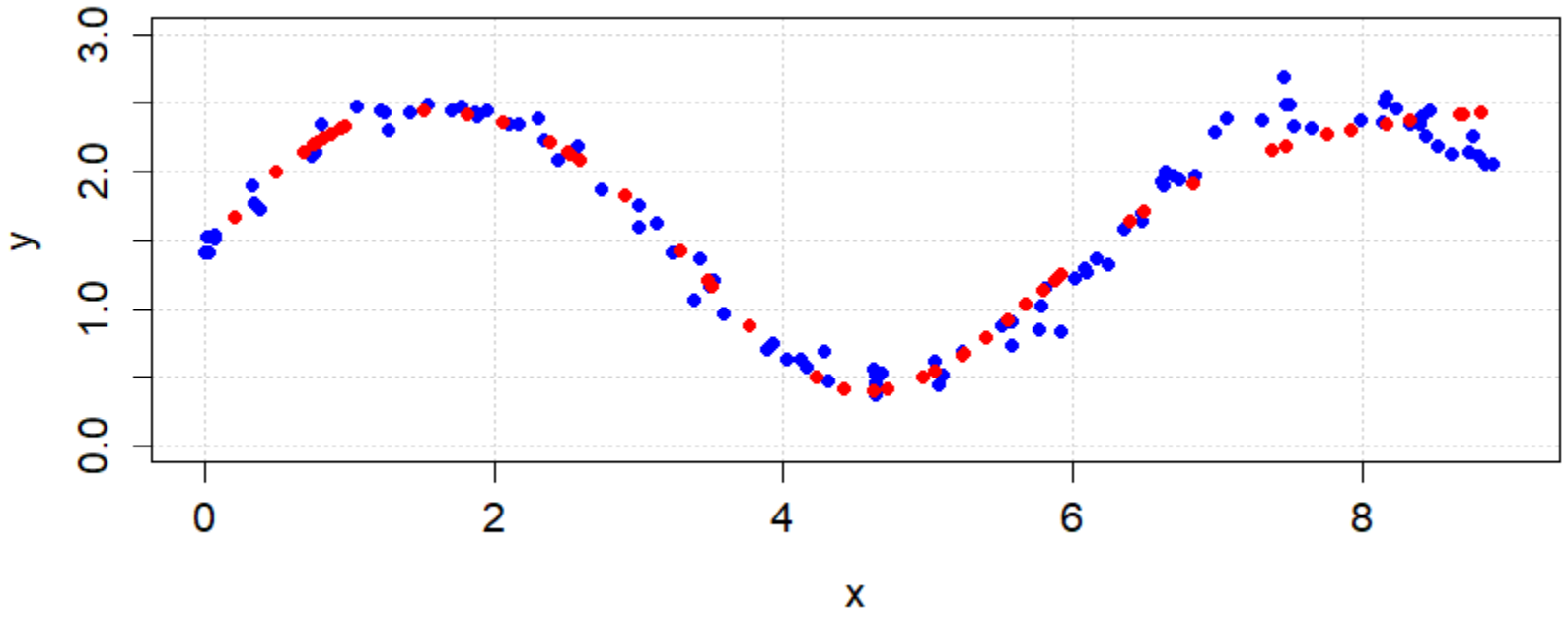
Vorhersagen im Streudiagramm (1)

- Wir haben oben eine **Vorhersage** für die Beobachtung $x_i = 2$ gerechnet.
- Der rote Punkt zeigt den vorhergesagten Wert $\hat{y}_i = 2.4$.



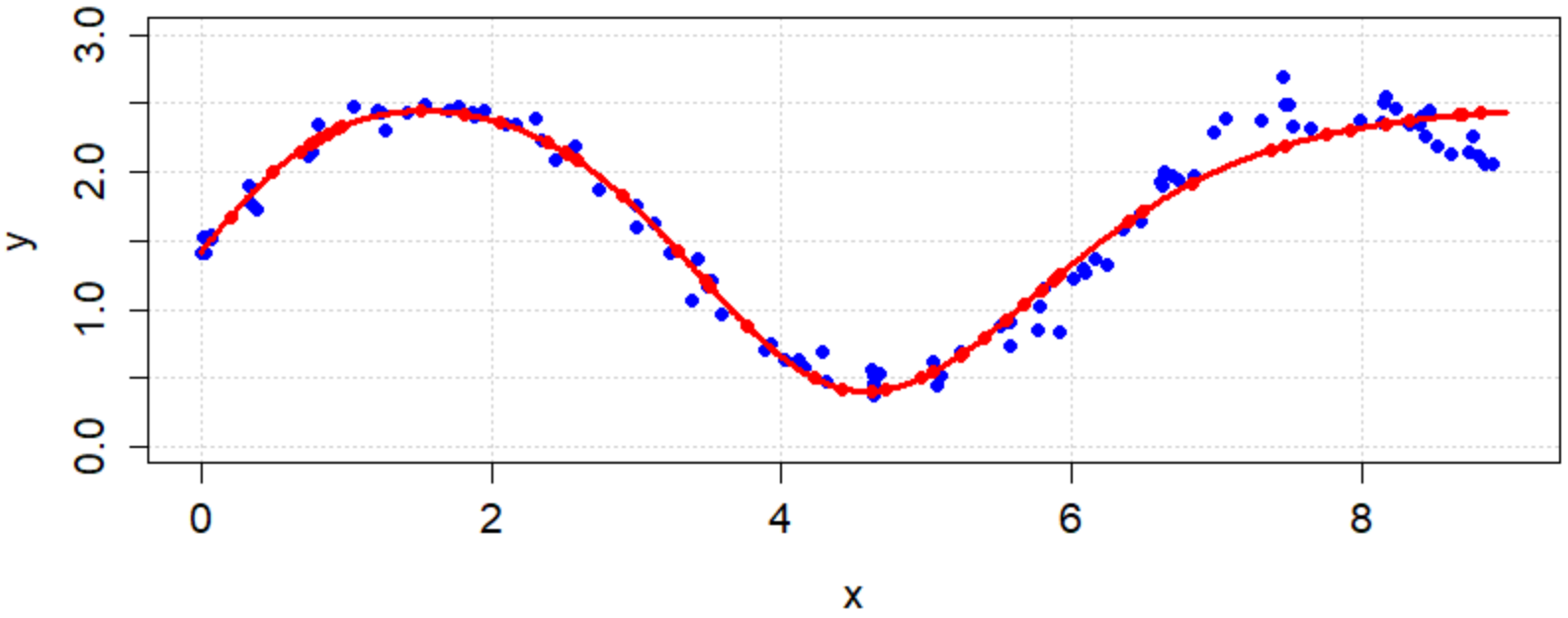
Vorhersagen im Streudiagramm (2)

- Wir können Vorhersagen für viele verschiedene x_i Werte einzeichnen.



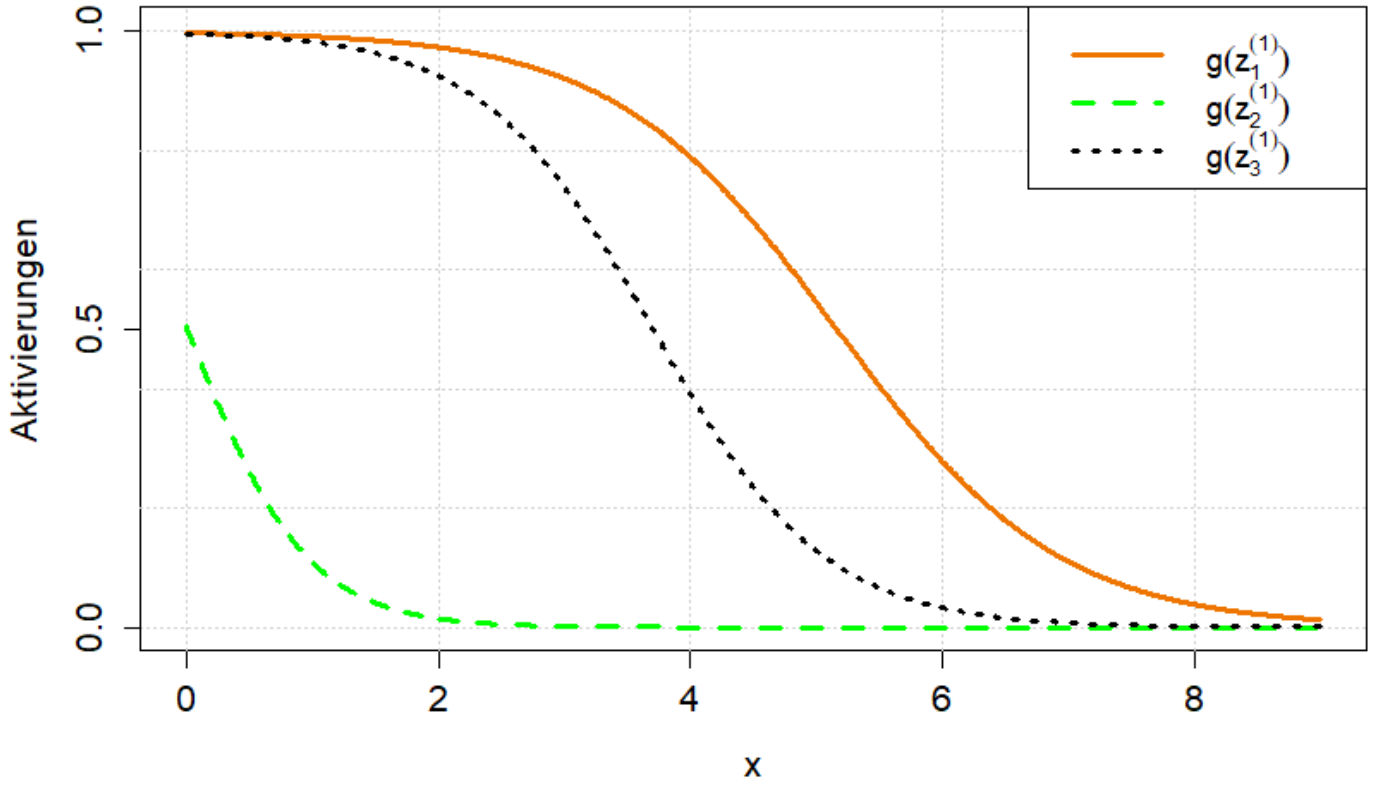
Vorhersagen im Streudiagramm (3)

- Wenn wir die Punkte interpolieren, dann kriegen wir eine **Kurve**, die das Modell repräsentiert.
- Wow! Unser kleines ANN kann den nicht-linearen Zusammenhang gut fitten ohne dass wir dem Modell sagen müssen, was es zu tun hat.



Aktivierungen

- Unser ANN **kombiniert** die drei **Aktivierungskurven** auf eine clevere Weise zu der roten Kurve auf vorheriger Abbildung!

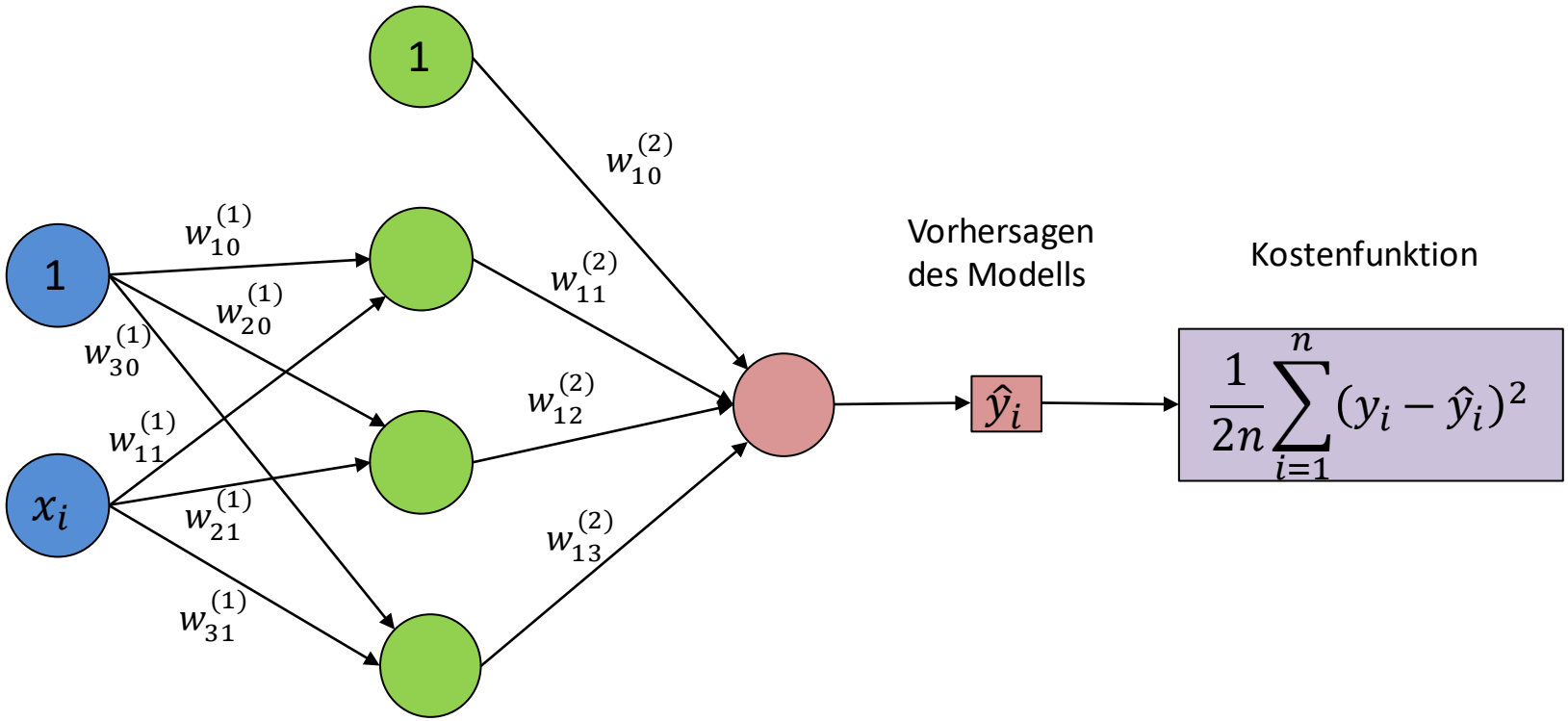


Wie optimieren wir die Gewichte?

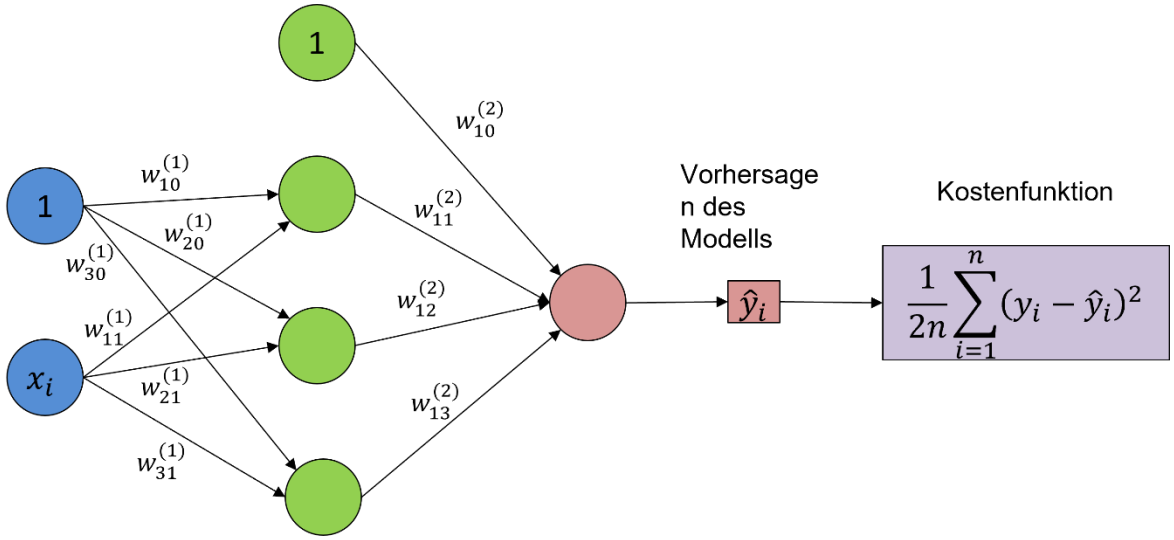
- Gewichte optimieren mit Gradient Descent
- Ableitung der Kostenfunktion nach Gewichten
- Ableitungsregeln
- Ableitung nach Gewichten in ZWEITEM Layer
- Ableitung nach Gewichten in ERSTEM Layer
- Alle Ableitungen auf einen Blick
- Forward Pass
- Backward Pass

Gewichte optimieren mit Gradient Descent

- Kostenfunktion (Regression): **Mean Squared Error** (MSE)



Ableitung der Kostenfunktion nach Gewichten



Als erstes müssen wir die Kostenfunktion umschreiben, so dass die **Gewichte explizit** auftauchen:

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \underbrace{\Big(w_{10}^{(2)} + w_{11}^{(2)} \cdot g\left(z_1^{(1)}\right) + w_{12}^{(2)} \cdot g\left(z_2^{(1)}\right) + w_{13}^{(2)} \cdot g\left(z_3^{(1)}\right)\Big)}_{\hat{y}_i}\Big)^2$$

Ableitungsregeln

- **Summenregel**
  - $(\sum_{i=1}^{n} f_i(x))' = \sum_{i=1}^{n} f'(x)$
  - Ableitung einer Summe (von Funktionen) = Summe der Ableitungen
- **Kettenregel**
  - $\left[f\left(g(x)\right)\right]' = f'\left(g(x)\right) \cdot g'(x)$ für zwei Funktionen $f()$ und $g()$.
  - Beispiel: Ableitung von $\ln(3x^2)$ nach x?

Ableitung nach Gewichten in ZWEITEM Layer

$$J(\boldsymbol{w}) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \left(w_{10}^{(2)} + w_{11}^{(2)} \cdot g\left(z_1^{(1)}\right) + w_{12}^{(2)} \cdot g\left(z_2^{(1)}\right) + w_{13}^{(2)} \cdot g\left(z_3^{(1)}\right)\right)\right)^2$$

Ableitungen nach Gewichten zwischen Hidden und Output Layer (rot markiert):

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(2)}} = \frac{2}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot (-1) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(2)}} = \frac{2}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot \left(-g\left(z_1^{(1)}\right)\right) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{12}^{(2)}} = \frac{2}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot \left(-g\left(z_2^{(1)}\right)\right) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{13}^{(2)}} = \frac{2}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot \left(-g\left(z_3^{(1)}\right)\right) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g\left(z_3^{(1)}\right)$$

Ableitung nach Gewichten in ERSTEM Layer

$$J(\boldsymbol{w}) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \left(w_{10}^{(2)} + w_{11}^{(2)} \cdot g\left(w_{10}^{(1)} + w_{11}^{(1)} \cdot x_i\right) + w_{12}^{(2)} \cdot g\left(z_2^{(1)}\right) + w_{13}^{(2)} \cdot g\left(z_3^{(1)}\right)\right)\right)^2$$

Ableitungen nach Gewichten zwischen Input und Hidden Layer (rot markiert):

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(1)}} = \frac{2}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot \left(-w_{11}^{(2)}\right) \cdot g'\left(w_{10}^{(1)} + w_{11}^{(1)} \cdot x_i\right) \cdot 1$$

$$= -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g'\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(1)}} = \frac{2}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot \left(-w_{11}^{(2)}\right) \cdot g'\left(w_{10}^{(1)} + w_{11}^{(1)} \cdot x_i\right) \cdot x_i$$

$$= -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g'\left(z_1^{(1)}\right) \cdot x_i$$

Für die Ableitungen im ersten Layer benötigen wir die Ableitung der Aktivierungsfunktion $g'(z)$. Für die **Sigmoid** Aktivierung lautet die Ableitung wie folgt:

$$g'(z) = g(z) \cdot (1 - g(z))$$

Alle Ableitungen auf einen Blick

Zweiter Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{12}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{13}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_3^{(1)}\right)$$

Erster Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{20}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{21}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{30}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{31}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)\cdot x_i$$

Was benötigen wir, um die Ableitungen zu rechnen?

Zweiter Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{12}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{13}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_3^{(1)}\right)$$

Werte der Zielvariable im Trainingsdatensatz: y_i

Erster Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{20}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{21}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{30}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{31}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)\cdot x_i$$

Was benötigen wir, um die Ableitungen zu rechnen?

Zweiter Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{12}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{13}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_3^{(1)}\right)$$

Aktuelle Vorhersagen unseres Modells: $\hat{y}_i$

Erster Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{20}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{21}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{30}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{31}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)\cdot x_i$$

Was benötigen wir, um die Ableitungen zu rechnen?

Zweiter Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{12}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{13}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g\left(z_3^{(1)}\right)$$

Erster Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g'\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g'\left(z_1^{(1)}\right) \cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{20}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g'\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{21}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g'\left(z_2^{(1)}\right) \cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{30}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g'\left(z_3^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{31}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \cdot g'\left(z_3^{(1)}\right) \cdot x_i$$

Aktivierungen für jede Trainingsbeobachtung: $g\left(z_1^{(1)}\right), g\left(z_2^{(1)}\right), g\left(z_3^{(1)}\right)$

Was benötigen wir, um die Ableitungen zu rechnen?

Zweiter Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{12}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{13}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_3^{(1)}\right)$$

Erster Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{20}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{21}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{30}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{31}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)\cdot x_i$$

Ableitungen der Aktivierungen: $g'\left(z_1^{(1)}\right)$, $g'\left(z_2^{(1)}\right)$, $g'\left(z_3^{(1)}\right)$

Was benötigen wir, um die Ableitungen zu rechnen?

Zweiter Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{12}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{13}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_3^{(1)}\right)$$

Erster Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{20}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{21}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{30}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{31}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)\cdot x_i$$

Aktuelle Gewichte zweiter Layer: $w_{11}^{(2)}, w_{12}^{(2)}, w_{13}^{(2)}$

Was benötigen wir, um die Ableitungen zu rechnen?

Zweiter Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{12}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{13}^{(2)}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g\left(z_3^{(1)}\right)$$

Erster Layer

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{11}^{(1)}} = -\frac{w_{11}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_1^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{20}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{21}^{(1)}} = -\frac{w_{12}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_2^{(1)}\right)\cdot x_i$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{30}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)$$

$$\frac{\partial J(\boldsymbol{w})}{\partial w_{31}^{(1)}} = -\frac{w_{13}^{(2)}}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot g'\left(z_3^{(1)}\right)\cdot x_i$$

Werte der Input-Variable im Trainingsdatensatz: x_i

Forward Pass

- Wir haben eben gesehen, dass wir ganz viele Dinge brauchen, um die **Ableitungen der Kostenfunktion** nach den verschiedenen Gewichten zu rechnen.
- Die Trainingsdaten (x_i, y_i) sind gegeben, ebenso die aktuellen Gewichte des Modells.
- Doch wie kriegen wir die **aktuellen Aktivierungen** (für jede Trainingsbeobachtung) und die **aktuellen Vorhersagen** unseres Modells $\hat{y}_i$?
  - Wir machen einen sogenannten **Forward Pass** durch unser Modell.
  - Das ist nichts anderes, als was wir im ersten Teil dieser Folien gemacht haben.
  - Ganz am Anfang des Trainings wird der erste Forward Pass die zufällig initialisierten Gewichte verwenden.

Backward Pass

- Nach dem Forward Pass haben wir alle «Zutaten», um die **Ableitungen** zu berechnen.
- Mit den Ableitungen können wir nun **für jedes Gewicht** des Modells einen **Gradient Descent Schritt** machen:

Zweiter Layer

$$w_{10}^{(2)} := w_{10}^{(2)} - \alpha \cdot \frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(2)}}$$

etc.

Erster Layer

$$w_{10}^{(1)} := w_{10}^{(1)} - \alpha \cdot \frac{\partial J(\boldsymbol{w})}{\partial w_{10}^{(1)}}$$

etc.

- Nun **wiederholen** wir diese Sequenz von Forward und Backward Pass ganz häufig, so dass sich die Gewichte des Modells immer mehr dem **Optimum** annähern.