Erklärbarkeit von Modellen

Wie transparent sind maschinell gelernte Modelle?

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

GEFÖRDERT VOM
INNOVET
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Beispiel lineare Regressionsmodelle

Einfache lineare Regression: Eisverkäufe und Höchsttemperatur

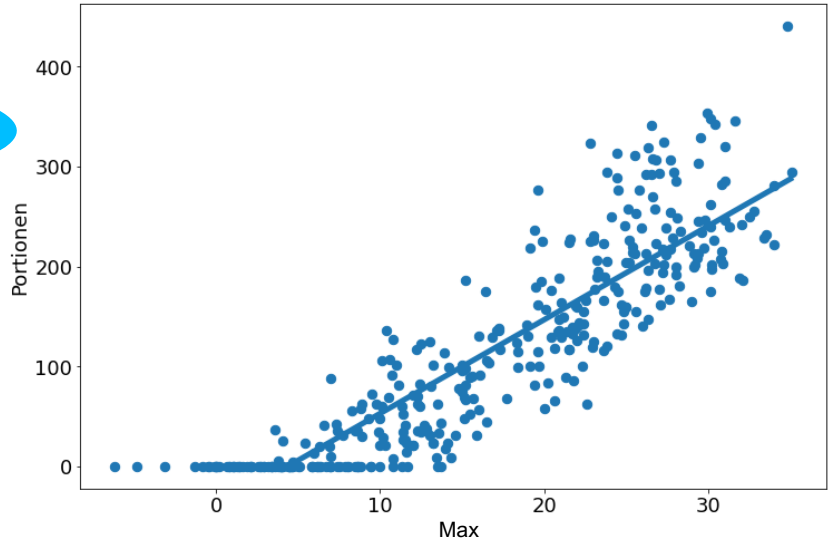
- Modell:

$$y = 9{,}4\,x - 40{,}8$$

Portionen
„Erklärte“ Variable

Max
„Erklärende“ Variable

$$y = b \cdot x + c$$

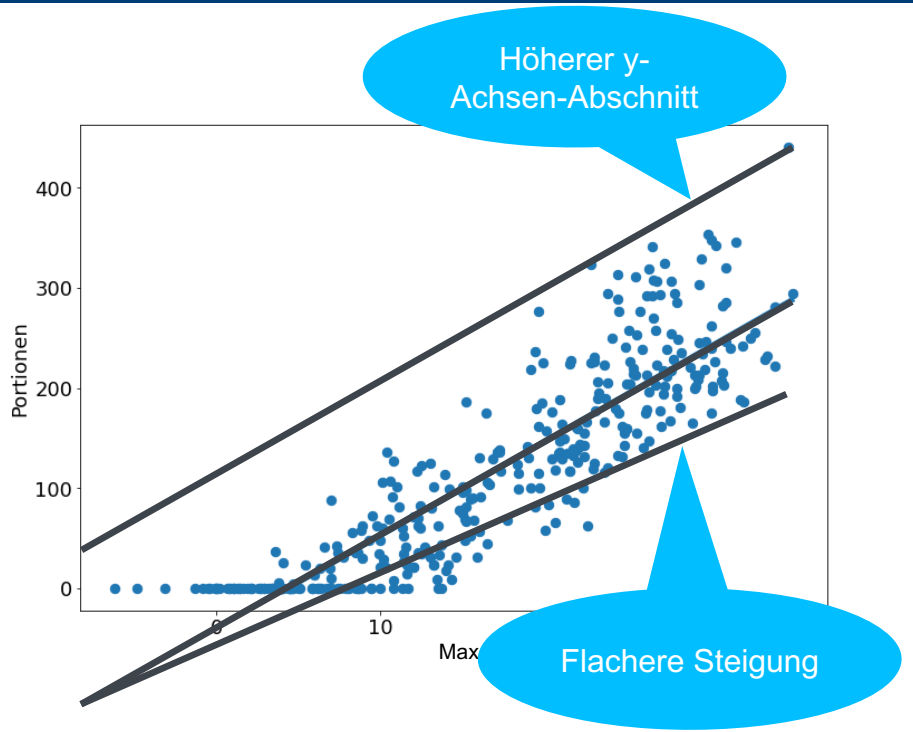


- Hier $b = 9{,}4$
  ⇒ jede Erhöhung von x um eins bewirkt eine Erhöhung von y um 9,4
- Hier $c = -40{,}8$
  ⇒ dieser Wert wird erwartet, wenn $x = 0$ ist

Erklärbarkeit von linearen Regressionsmodellen

- Vorhersage kann „nachgerechnet“ werden
- Beispielvorhersage für Max = 20

  $x = 20 \Rightarrow y = 9.4 \cdot 20 - 40.8 \approx 150$
- Modell ist eine einfache Funktion
- Modellparameter (hier Steigung und y-Achsenabschnitt der entsprechenden linearen Funktion) leicht zu interpretieren
- Zusammenhang zwischen Eingabe und Vorhersage ist klar



Bei der einfachen linearen Regression gibt es nur 2 Parameter, und beide sind leicht zu interpretieren.

Single Choice: Parameter bei der linearen Regression

Multiple lineare Regression von Immobilienwerten („California Housing“ Daten)

	Wert	Einkommen	Hausalter	Zimmer/Haushalt	Schlafzimmer/Haushalt	Einwohner	Bewohner/Haushalt	West	Nord
0	4.526	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23
1	3.585	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22
2	3.521	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24
3	3.413	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25
4	3.422	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25
...	...	...	...	...	...	...	...	...	...
20635	0.781	1.5603	25.0	5.045455	1.133333	845.0	2.560606	39.48	-121.09
20636	0.771	2.5568	18.0	6.114035	1.315789	356.0	3.122807	39.49	-121.21
20637	0.923	1.7000	17.0	5.205543	1.120092	1007.0	2.325635	39.43	-121.22
20638	0.847	1.8672	18.0	5.329513	1.171920	741.0	2.123209	39.43	-121.32
20639	0.894	2.3886	16.0	5.254717	1.162264	1387.0	2.616981	39.37	-121.24

erklärte Variable

erklärende Variablen

Modell und Beispielvorhersage

	Einkommen	Hausalter	Zimmer/Haushalt	Schlafzimmer/Haushalt	Einwohner	Bewohner/Haushalt	West	Nord
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23

	Variable	Koeffizient
0	Einkommen	0.436693
1	Hausalter	0.009436
2	Zimmer/Haushalt	-0.107322
3	Schlafzimmer/Haushalt	0.645066
4	Einwohner	-0.000004
5	Bewohner/Haushalt	-0.003787
6	West	-0.421314
7	Ost	-0.434514
y-Achsenabschnitt:		-36.941920

Wert = 0.436693 * 8.3252
+ 0.009436 * 41
- 0.107322 * 6.984127
+ 0.645066 * 1.023810
- 0.000004 * 322
- 0.003787 * 2.555556
- 0.421314 * 37.88
-0.434514 * (-122.23)
-36.94192021
= 4.13169

Vorhersage: 4.13169

Erklärbarkeit von linearer Regressionsmodellen

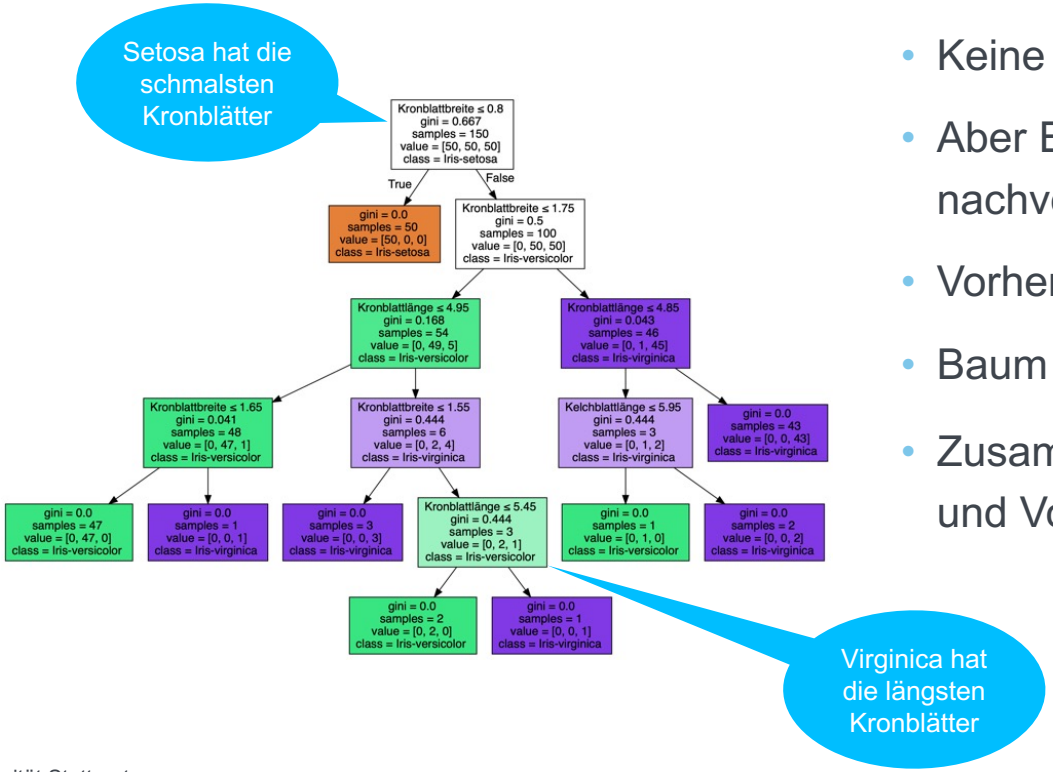
- Vorhersage kann „nachgerechnet“ werden
- Modellparameter (hier Koeffizient für jede erklärende Variable und y-Achsenabschnitt) leicht zu interpretieren
- Zusammenhang zwischen Eingabewerten und Vorhersage ist klar

	Variable	Koeffizient
0	Einkommen	0.436693
1	Hausalter	0.009436
2	Zimmer/Haushalt	-0.107322
3	Schlafzimmer/Haushalt	0.645066
4	Einwohner	-0.000004
5	Bewohner/Haushalt	-0.003787
6	West	-0.421314
7	Ost	-0.434514
y-Achsenabschnitt:		-36.941920

Bei der multiplen linearen Regression gibt es pro erklärender Variable einen Parameter, sowie den y-Achsen-Abschnitt, und alle sind leicht zu interpretieren.

Beispiel Entscheidungsbäume

Beispiel Entscheidungsbäume: Klassifikationsbaum für Irisarten



- Keine Parameter (keine “Berechnung“)
- Aber Entscheidungsfindung nachvollziehbar:
- Vorhersage kann nachgerechnet werden
- Baum ist interpretierbar
- Zusammenhang zwischen Eingabewerten und Vorhersage ist klar

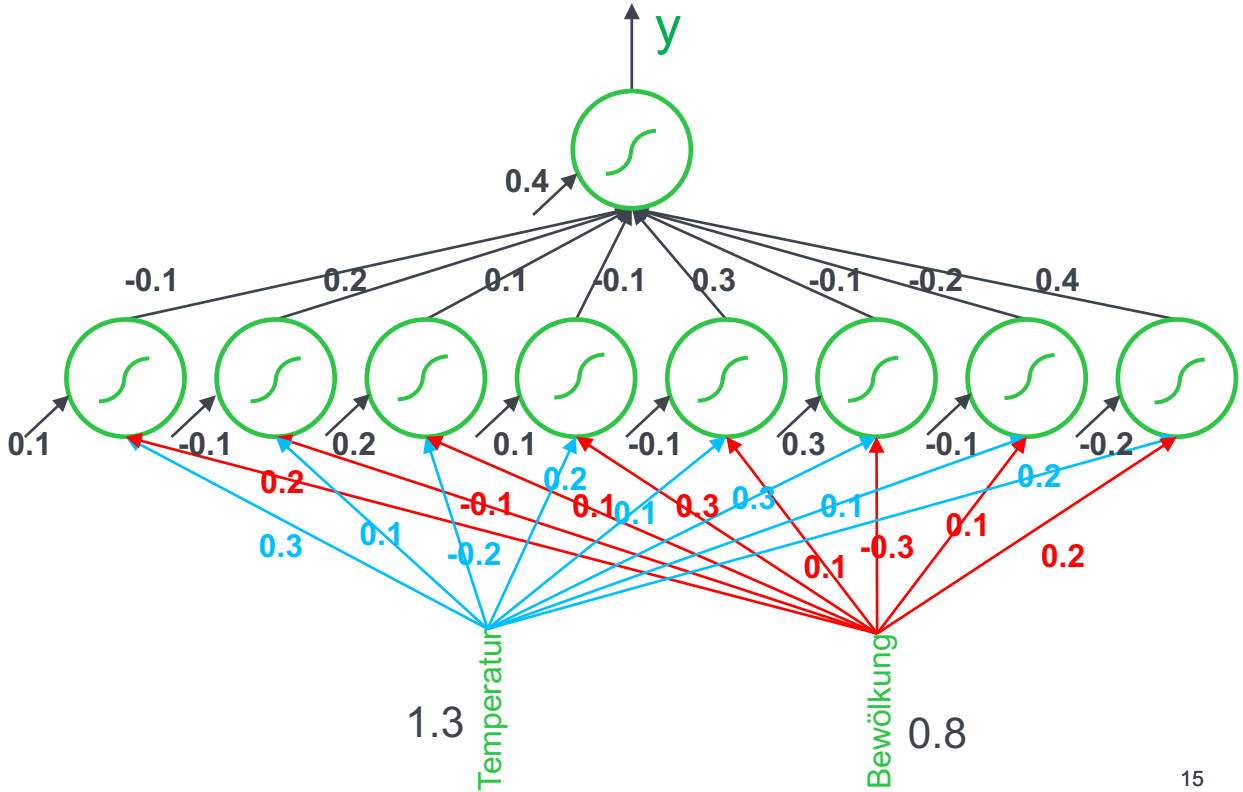
Entscheidungsbäume sind interpretierbar. Es ist klar, aufgrund welcher Eigenschaften wie klassifiziert wird.

Der Zusammenhang zwischen Eingabewerten und Vorhersage ist klar.

Beispiel neuronale Netze

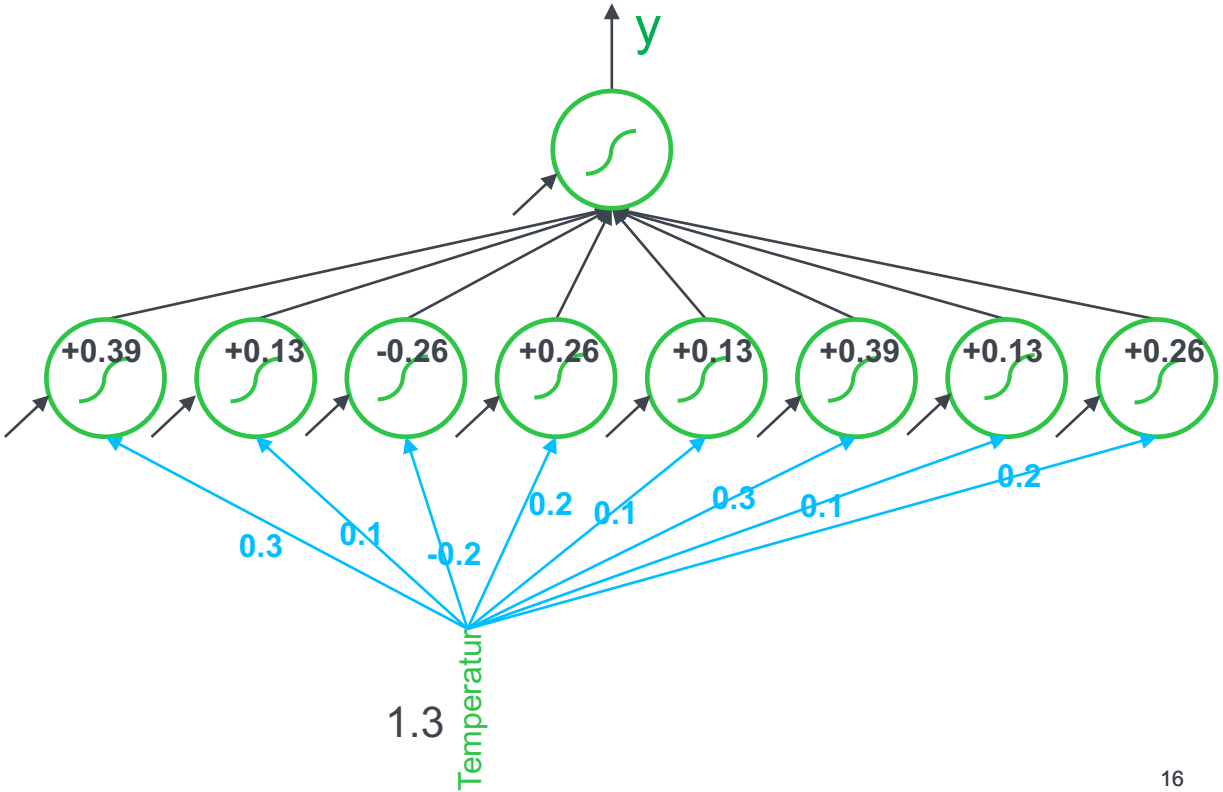
Das Trainingsbeispiel für Eisverkäufe

- Mini-Beispiel
- Netzwerk mit erfundenen Parametern
- Forward Pass für einen Datenpunkt
  - Temperatur 1.3
    Bewölkung 0.8
    (standardisierte Werte)



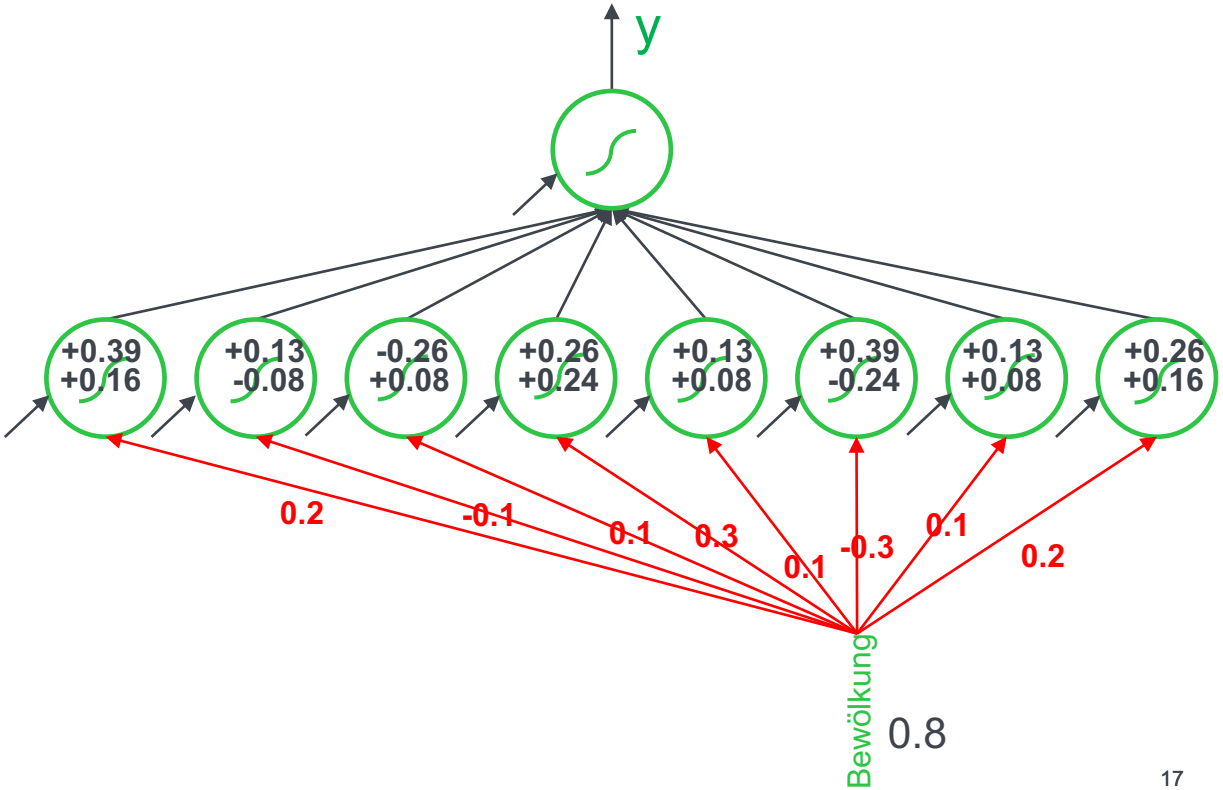
Forward Pass

- Logits der ersten Schicht (z_i^1)
  - Inputs von Temperatur



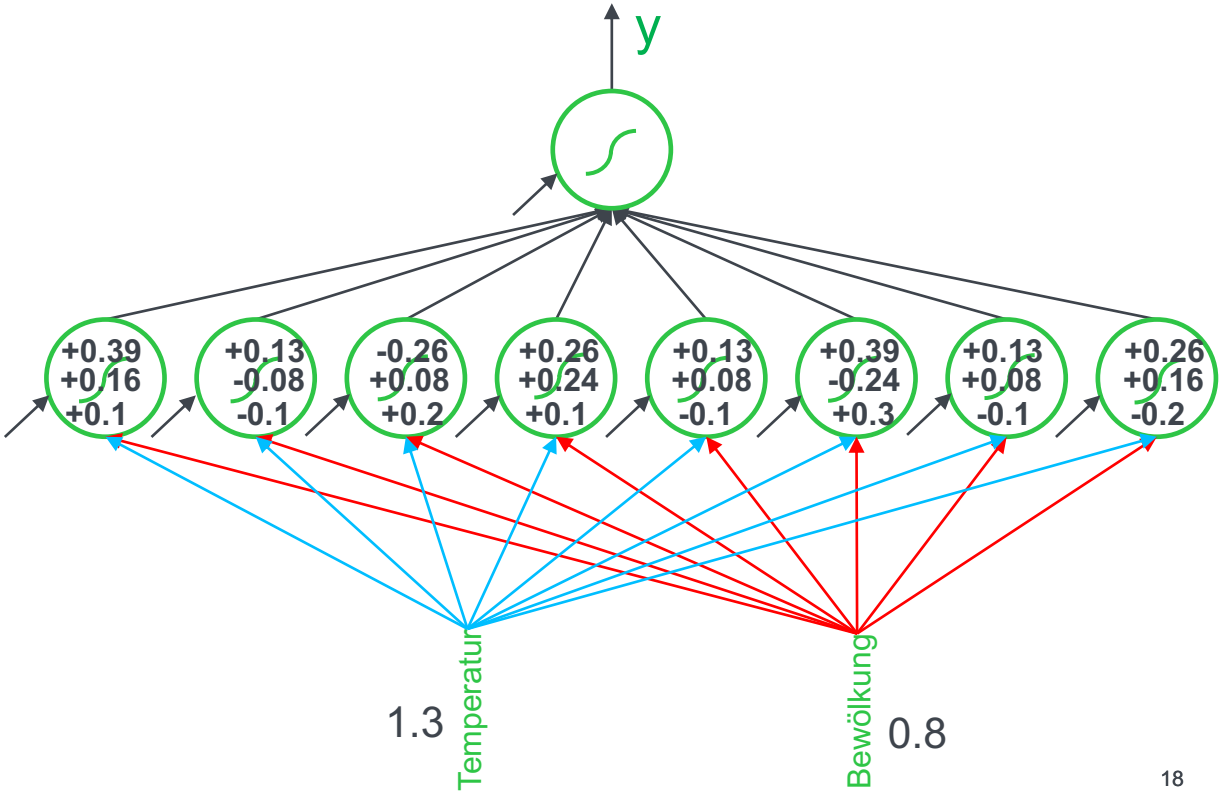
Forward Pass

- Logits der ersten Schicht (z_i^1)
  - Inputs von Temperatur
  - Inputs von Bewölkung



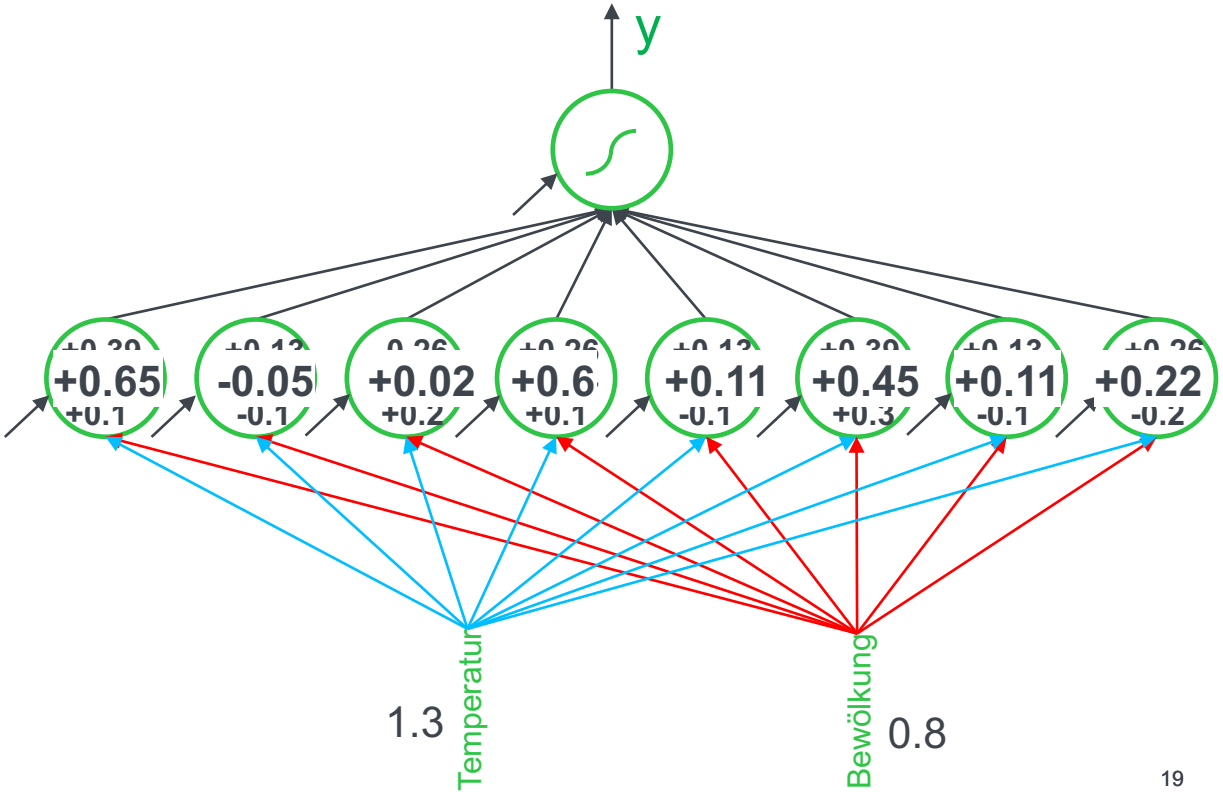
Forward Pass

- Logits der ersten Schicht (z_i^1)
  - Inputs von Temperatur
  - Inputs von Bewölkung
  - Bias



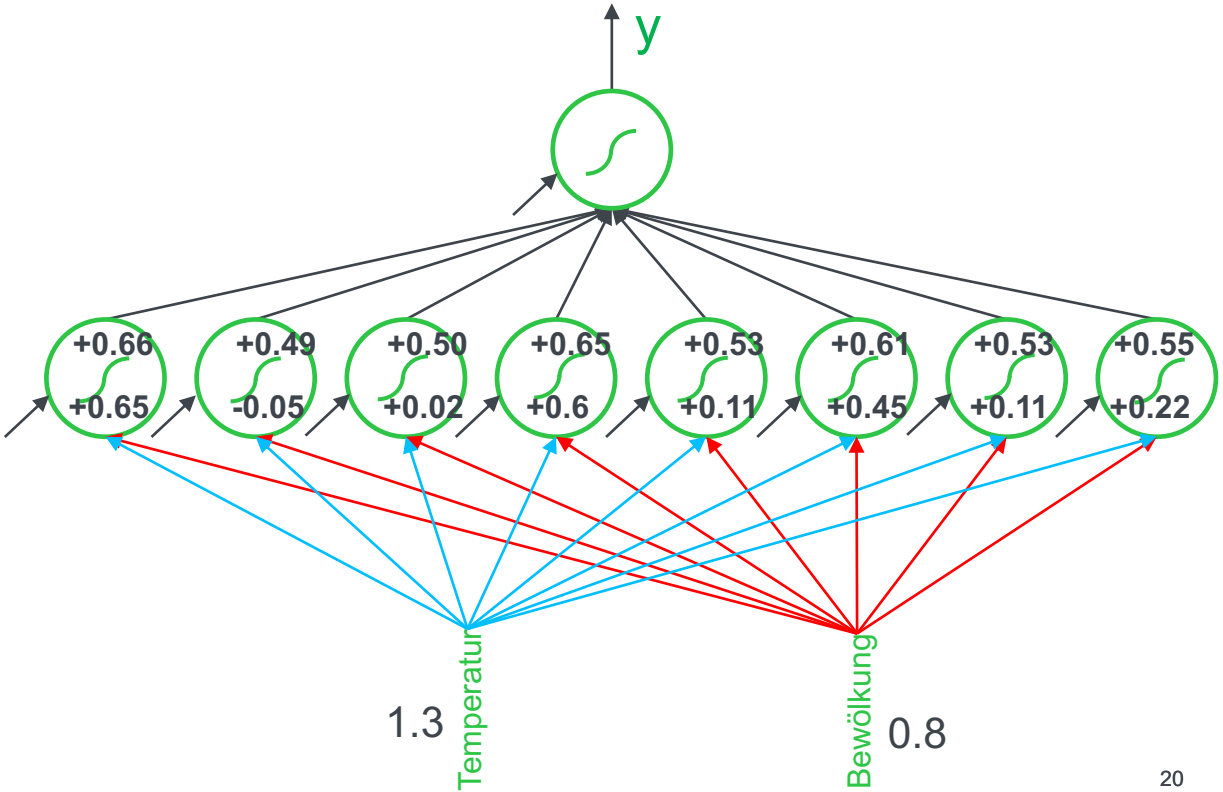
Forward Pass

- Logits der ersten Schicht (z_i^1)
  - Inputs von Temperatur
  - Inputs von Bewölkung
  - Bias



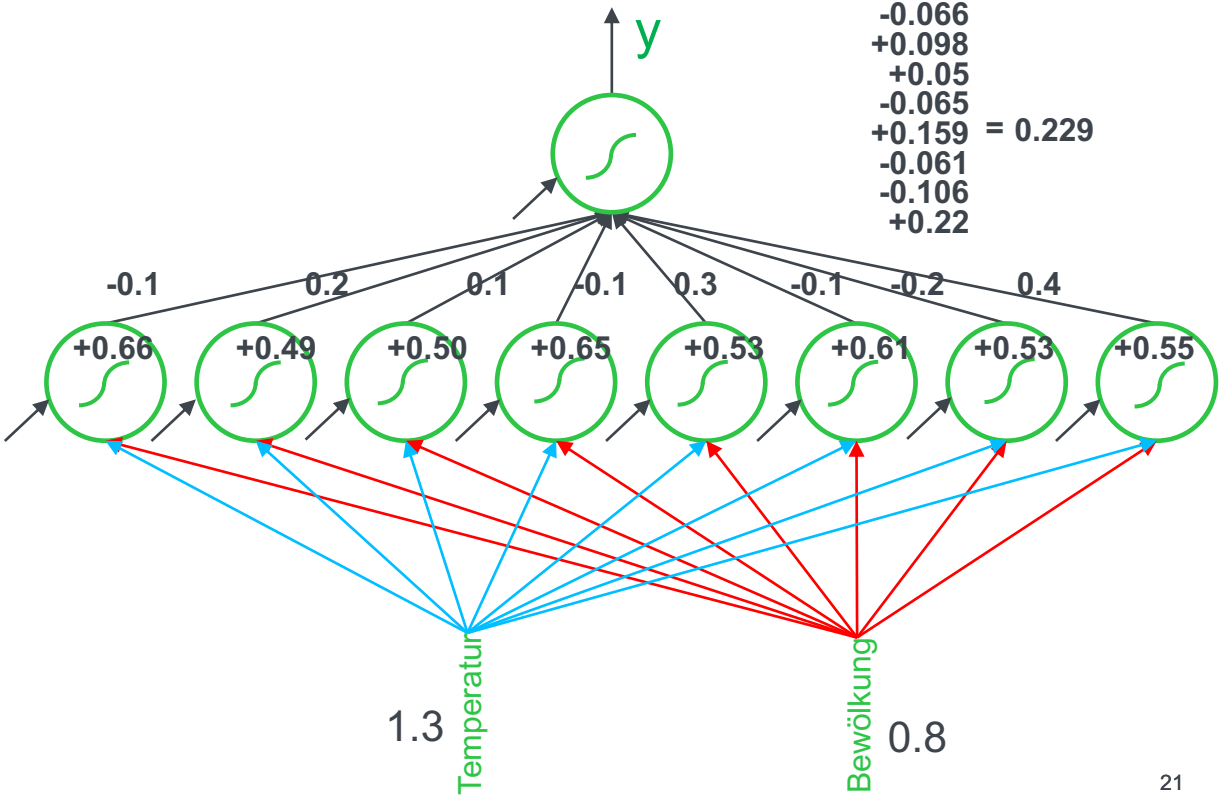
Forward Pass

- Logits der ersten Schicht (z_i^1)
  - Inputs von Temperatur
  - Inputs von Bewölkung
  - Bias
- Sigmoid-Aktivierung der ersten Schicht (a_i^1)



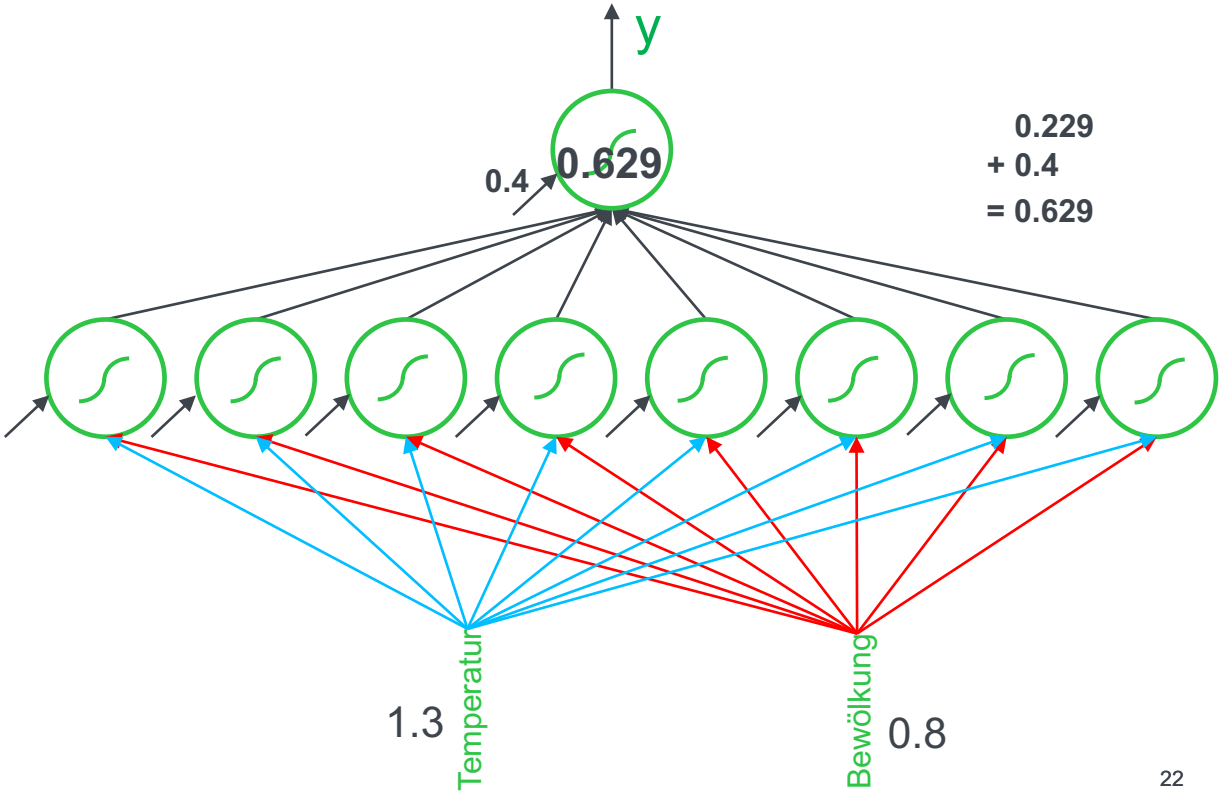
Forward Pass

- Logits der ersten Schicht (z_i^1)
  - Inputs von Temperatur
  - Inputs von Bewölkung
  - Bias
- Sigmoid-Aktivierung der ersten Schicht (a_i^1)
- Logits der zweiten Schicht (z_i^2)
  - Inputs von Schicht 1



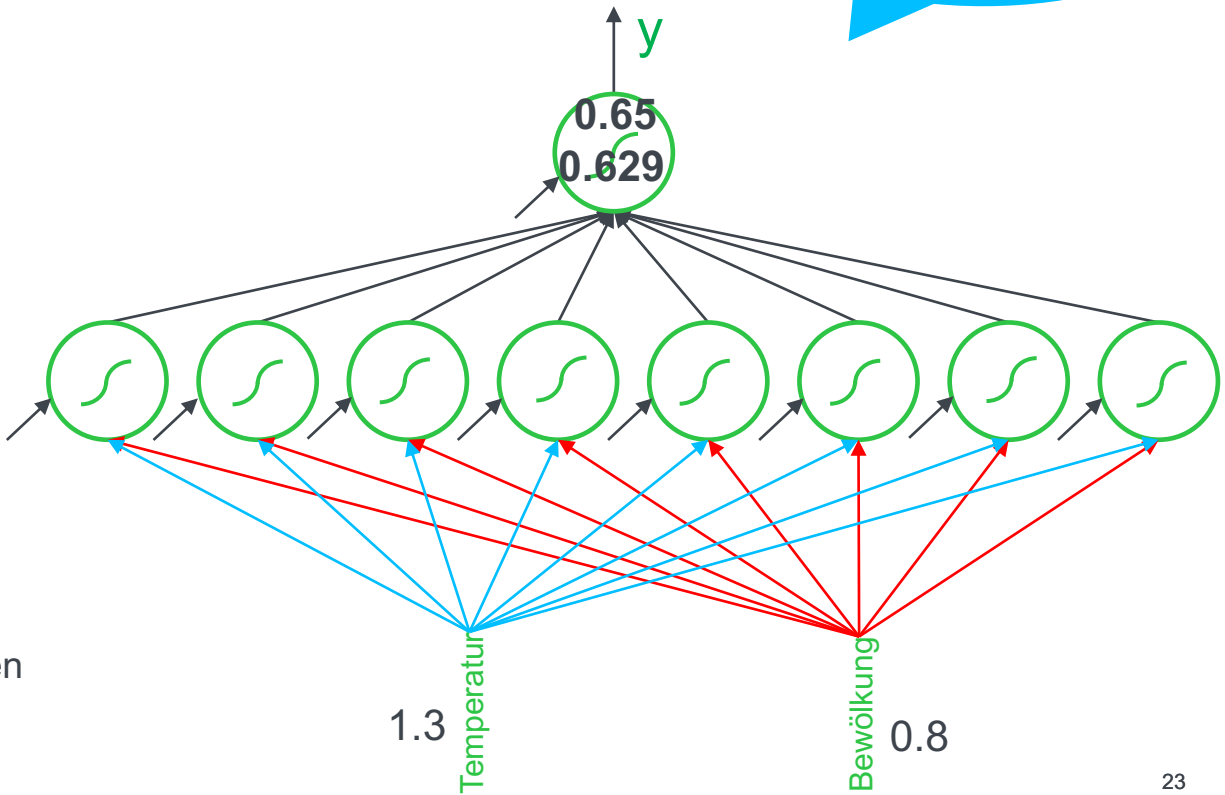
Forward Pass

- Logits der ersten Schicht (z_i^1)
  - Inputs von Temperatur
  - Inputs von Bewölkung
  - Bias
- Sigmoid-Aktivierung der ersten Schicht (a_i^1)
- Logits der zweiten Schicht (z_i^2)
  - Inputs von Schicht 1
  - Bias



Forward Pass

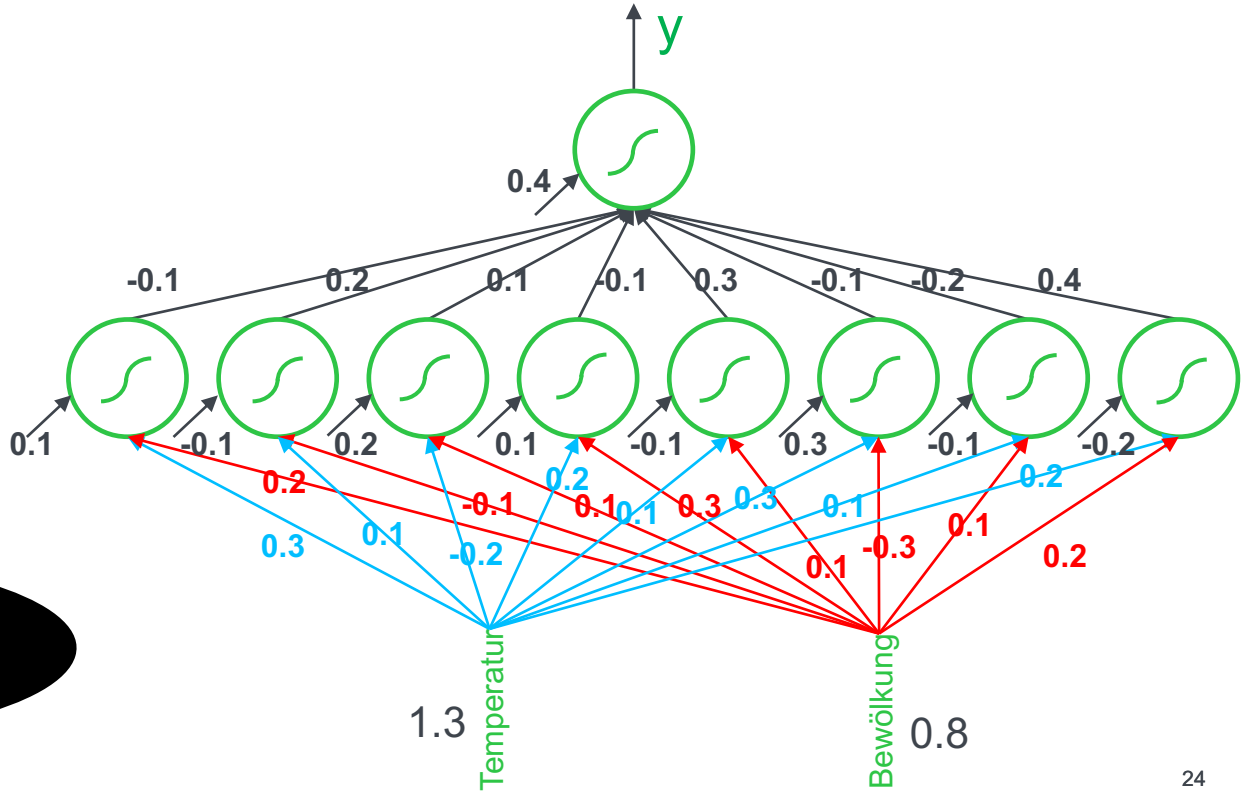
- Logits der ersten Schicht (z_i^1)
  - Inputs von Temperatur
  - Inputs von Bewölkung
  - Bias
- Sigmoid-Aktivierung der ersten Schicht (a_i^1)
- Logits der zweiten Schicht (z_i^2)
  - Inputs von Schicht 1
  - Bias
- Sigmoid-Aktivierung der zweiten Schicht (a_i^1)



Erklärbarkeit von neuronalen Netzen

- Vorhersage kann nachgerechnet werden
- Einzelne Parameter schwer interpretierbar
- Zusammenhang zwischen Eingabewerten und Vorhersage unklar

Oft Millionen bis Milliarden von Parametern ⇒ effektiv eine „Black Box“



Neuronale Netze sind schlecht interpretierbar. Die Vorhersage kann zwar nachgerechnet werden, aber der Zusammenhang zwischen Eingabewerten und Vorhersage ist aufgrund der sehr großen Anzahl von Parametern oft unklar.

Neuronale Netze werden daher oft als Black-Box-Modelle bezeichnet.

Question Set: Parameter in maschinell gelernten Modellen

Sort the Paragraphs: Interpretierbarkeit von Modellen

INNOVET

KI B³

Künstliche Intelligenz in die berufliche Bildung bringen

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering

IHK Industrie- und Handelskammer Reutlingen
Reutlingen | Tübingen | Zollernalb

IHK Region Stuttgart

IHK Industrie- und Handelskammer Karlsruhe

LMU Ludwig-Maximilians-Universität München

GEFÖRDERT VOM

INNOVET

Bundesministerium für Bildung und Forschung

bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Lizenzbestimmungen

“Erklärbarkeit von Modellen" von Antje Schweitzer, KI B³ / Uni Stuttgart

Das Werk - mit Ausnahme der folgenden Elemente:

- Logos der Verbundpartner und des Förderprogramms
- im Quellenverzeichnis aufgeführte Medien

ist lizenziert unter:

 CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.de)
(Namensnennung 4.0 International)

Quellenverzeichnis

Titelfoto: Markus Spiske (https://unsplash.com/de/@markusspiske), ohne Titel, auf Unsplash (https://unsplash.com/de/fotos/iar-afB0QQw), lizenziert unter Unsplash-Lizenz (https://unsplash.com/license). Bildausschnitt verändert.