Foto: Riho Kroll auf Unsplash

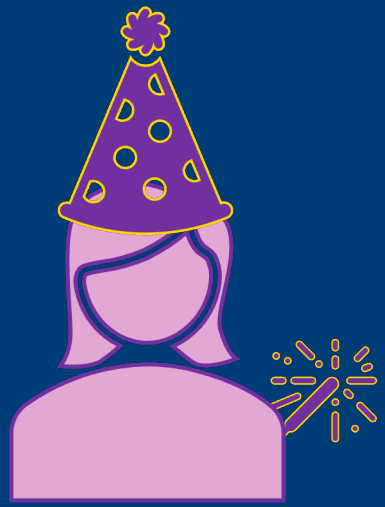
Overfitting – Teil 1

Wie erkennt und vermeidet man Overfitting?

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

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

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



Wenn es eine Wetterfee gäbe…

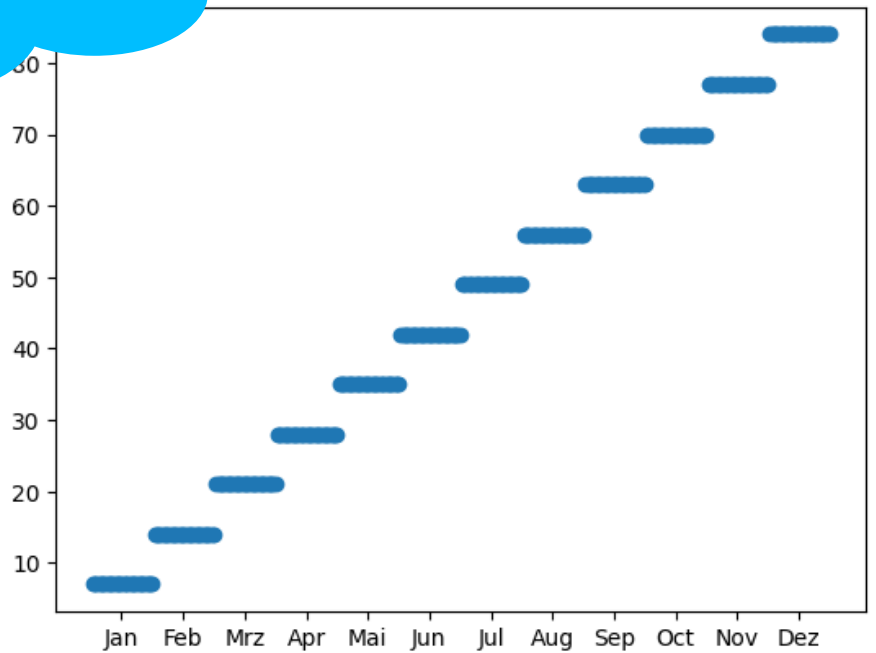
Die Wetterfee

Im Januar kalt, und dann jeden Monat 7 Grad wärmer!

```
def wetter_gen(monat, tag):
  return(month*7)
```

Die Wetterfee

Ab August wieder kälter!

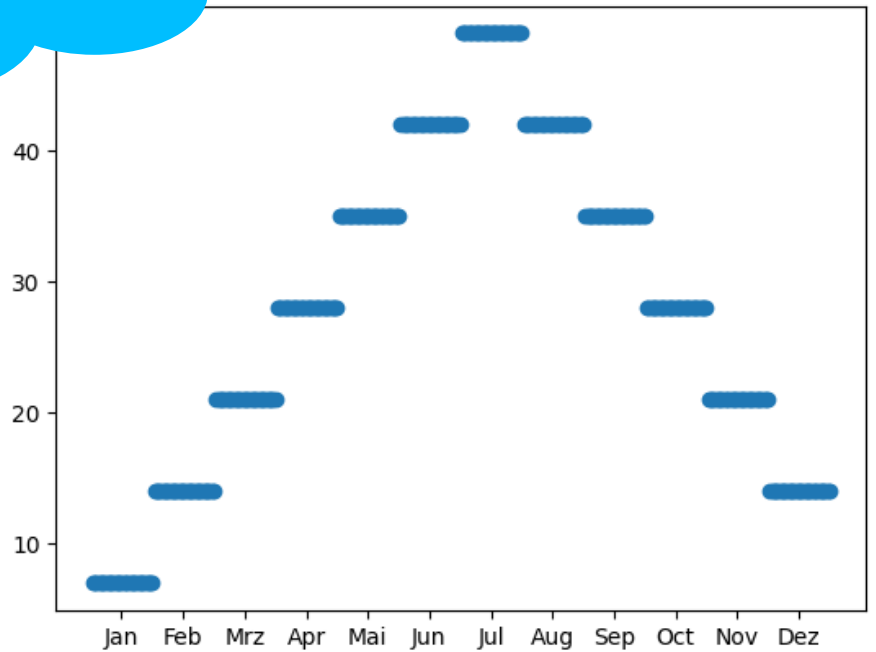


Die Wetterfee

```
def wetter_gen(monat, tag):
  if month > 7:
    return((14-month)*7)
  else:
    return(month*7)
```

Die Wetterfee

Insgesamt noch zu heiß!

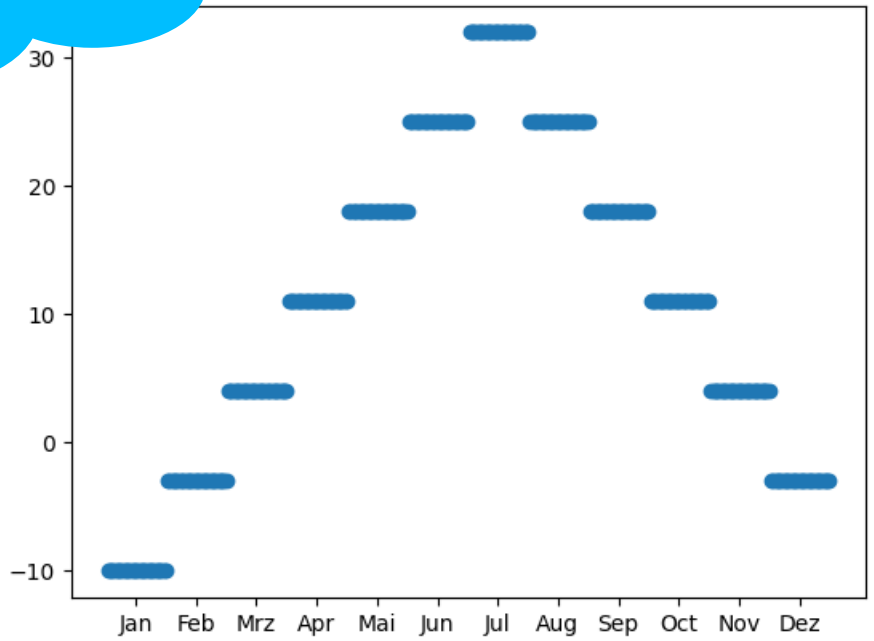


Die Wetterfee

```
def wetter_gen(monat, tag):
  if month > 7:
    return((14-month)*7-17)
  else:
    return(month*7-17)
```

Die Wetterfee

Jeden Tag ein bisschen kälter bzw. wärmer!



Die Wetterfee

```
def wetter_gen(monat, tag):
  if month > 7:
    return((14-month)*7-17-tag*0.25)
  else:
    return(month*7-17+tag*0.25)
```

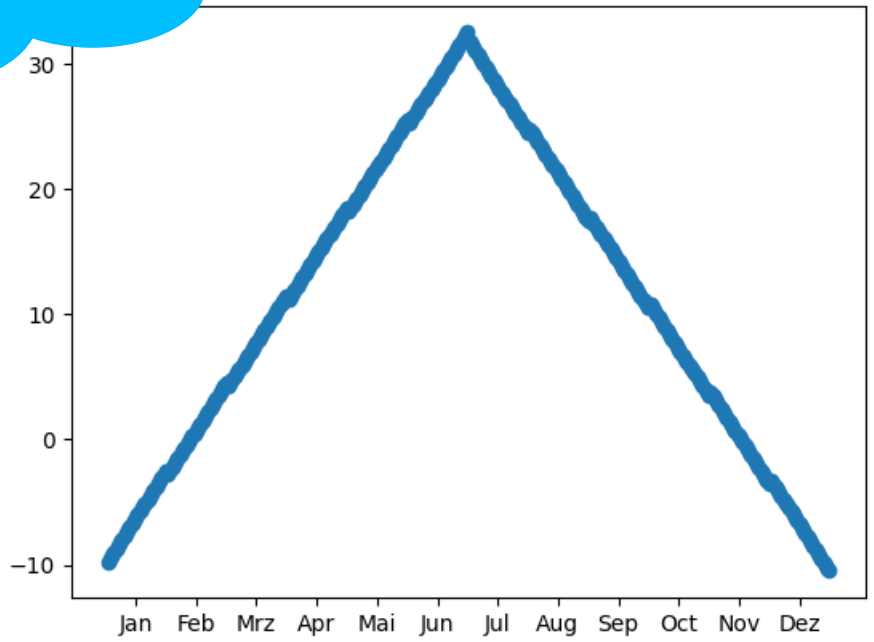
Die Wetterfee

Und jetzt noch ein bisschen Abwechslung!!

Zahl würfeln

Kopf: von Tagestemperatur abziehen

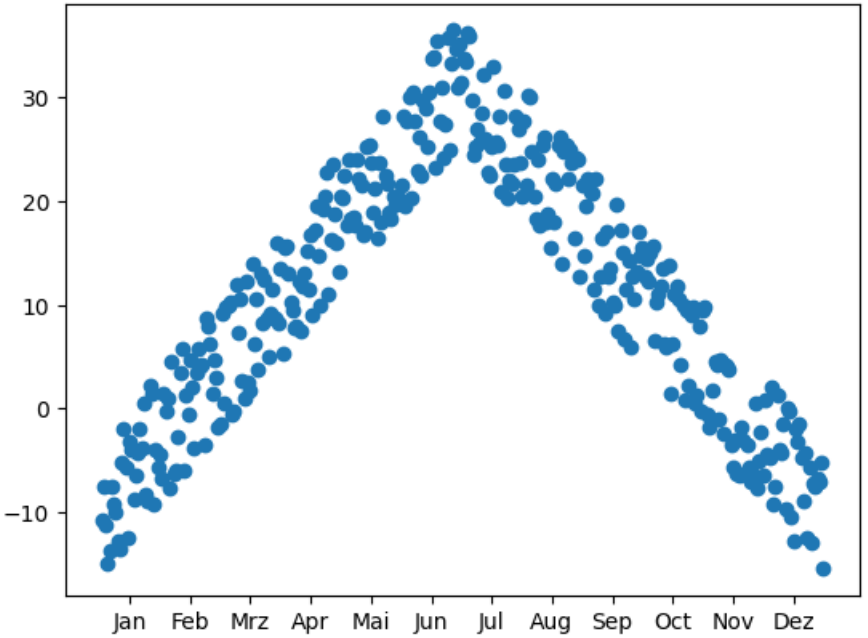
Zahl: zu Tagestemperatur addieren



Die Wetterfee

```
def wetter_gen(monat, tag):
  muenze = random.sample([-1,1],1)[0]
  wuerfel = random.randint(1,6)
  if month > 6:
    return((14-month)*7-day*0.25-17
      +muenze*wuerfel)
  else:
    return(month*7+day*0.25-17
      +muenze*wuerfel)
```

Die Wetterfee



Die Wettervorhersage

Wir sammeln Trainingsdaten und trainieren ein Modell!

Wissenschaftler in der Feen-Welt

Hihi, selbst das beste Modell wird im Mittel einen absoluten Fehler von 3.5 machen…

Augen	Fehler	abs.
1	1 oder -1	1
2	2 oder -2	2
3	3 oder -3	3
4	4 oder -4	4
5	5 oder -5	5
6	6 oder -6	6

gleich häufig!

Mittel: 3.5

Disclaimer

Anders als in der Feen-Welt weiß man bei Anwendungen von Maschinellem Lernen im echten Leben normalerweise nicht, wie gut die bestmögliche Performanz ist.

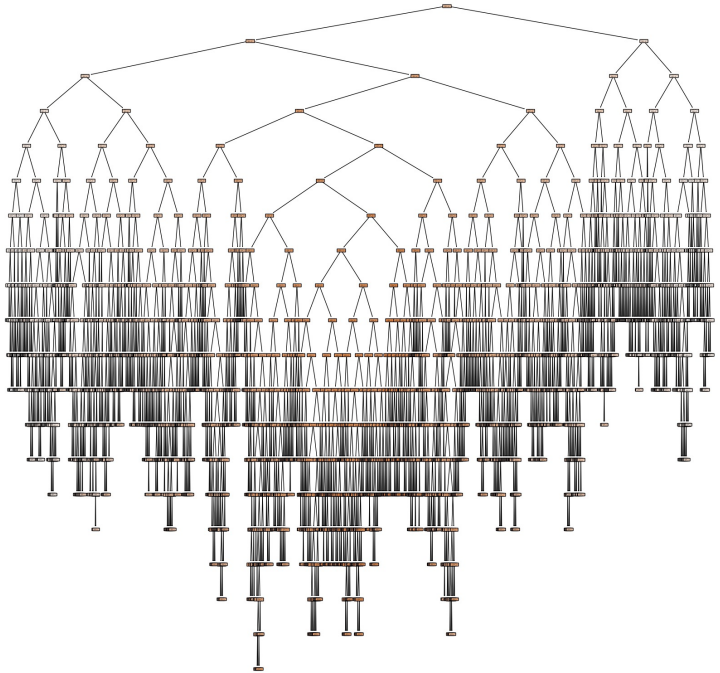
Die Daten für die Wissenschaftler

- Vereinfachte Annahme: jeder Monat besteht aus 30 Tagen
  - 360 Tage pro Jahr
- Trainingsdaten für 4 Jahre: 1440 Datenpunkte
- Validierungsdaten (Daten aus dem Vorjahr): 360 Datenpunkte
- Testdaten (Daten aus dem Folgejahr): 360 Datenpunkte

Ein Regressionsbaum

- Dieser Regressionsbaum macht auf den Daten keinen Fehler
- Mittlerer (absoluter) Fehler: 0
- Hat aber 1403 Blätter bei 1440 Datenpunkten
- Mittlerer Fehler im nächsten Jahr: 4.57

Overfitting



Beste Performanz wäre 3.5

➡ Dieser Baum generalisiert nicht gut

Bei Overfitting sind die Modelle unnötig komplex, weil sie auch irrelevante Details der Trainingsdaten modellieren.

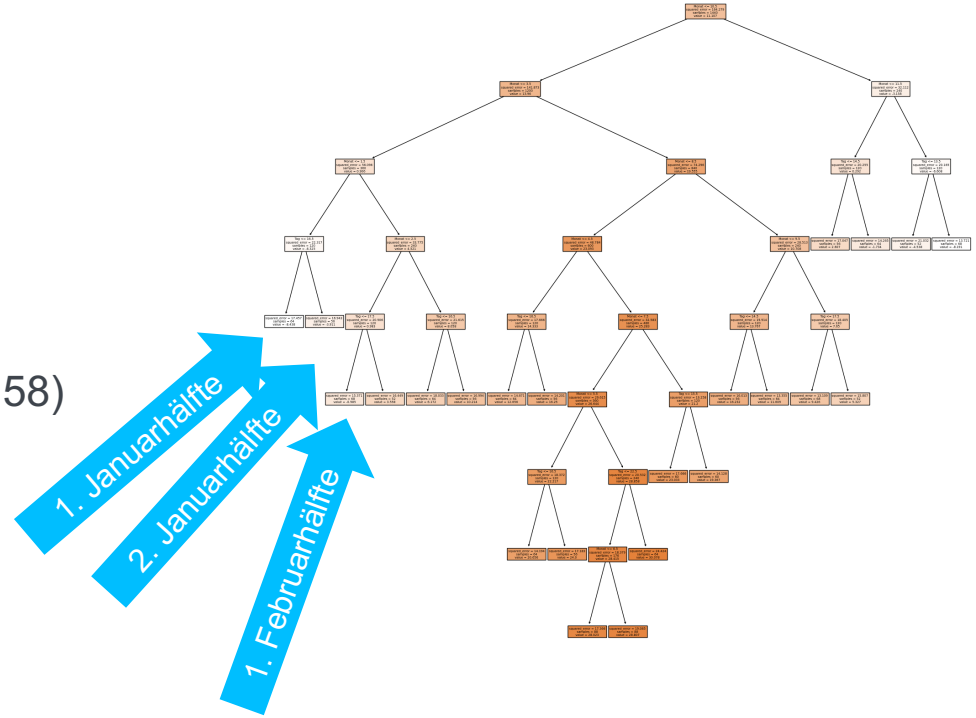
Overfitting erkennt man an der deutlich schlechteren Performanz auf den Testdaten.

Dadurch generalisieren die Modelle schlecht, d.h., sie “gelten“ nicht im Allgemeinen, sondern nur auf den Trainingsdaten.

Ein einfacherer Regressionsbaum

- Mindestens 50 Fälle pro Blatt
- Nur noch 23 Blätter
- Mittlerer Fehler: 3.52
- Mittlerer Fehler im nächsten Jahr: 3.79
- (und auf den Daten vom Vorjahr: 3.58)

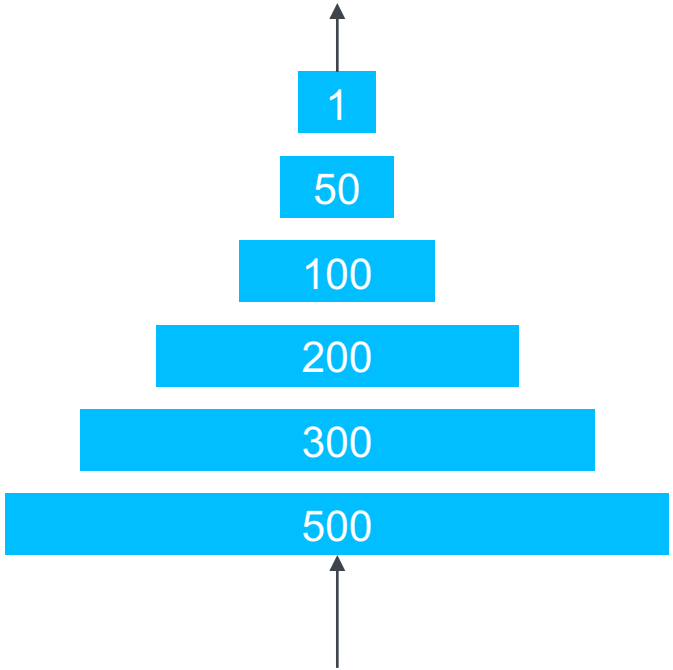
kaum Overfitting



Ein Feed Forward Network

- Fast 240 000 Parameter
- Nach 10 000 Epochen:
- Mittlerer Fehler: 0.15
- Mittlerer Fehler auf Validierungsdaten: 4.28

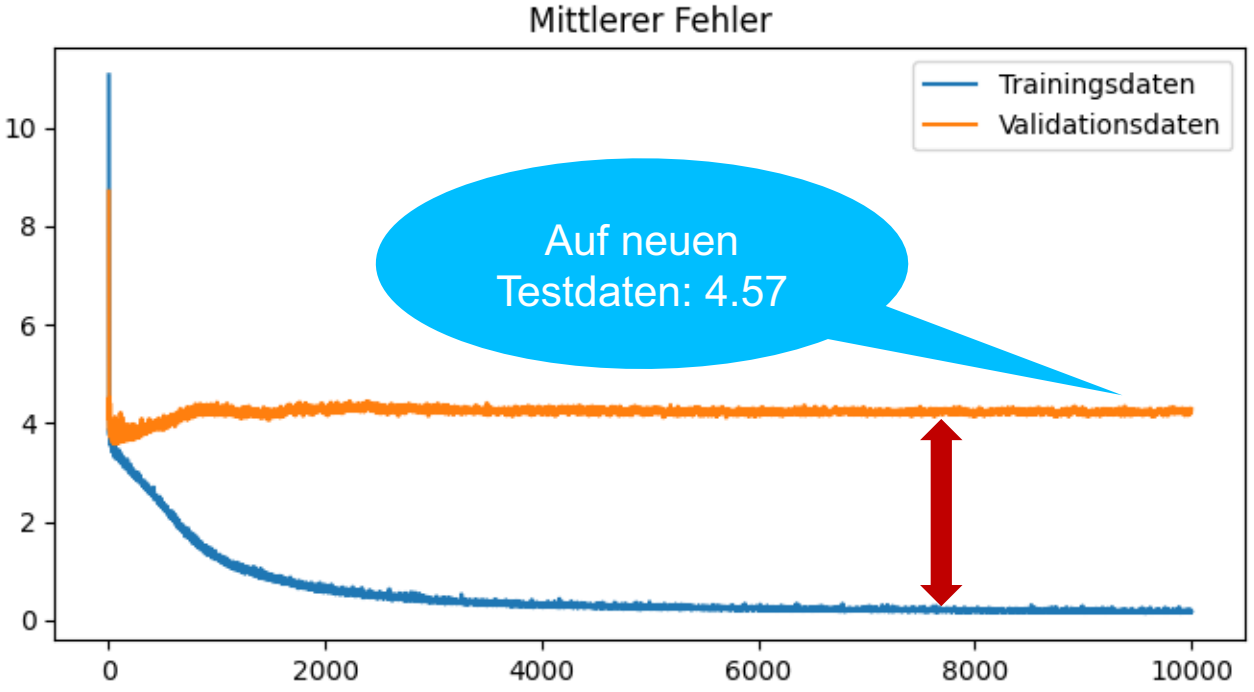
Overfitting



Ein Feed Forward Network

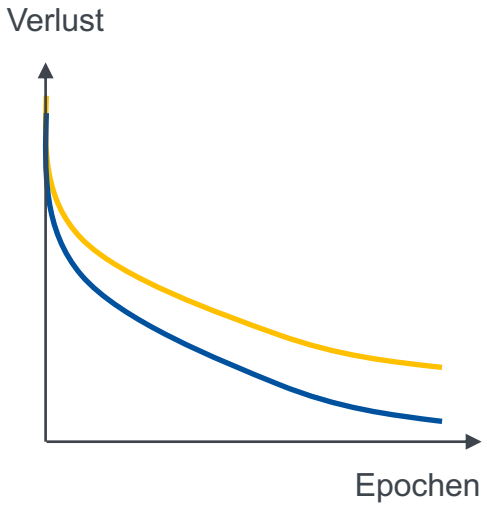
- Fast 240 000 Parameter
- Nach 10 000 Epochen:
- Mittlerer Fehler: 0.15
- Mittlerer Fehler auf Validierungsdaten: 4.28

Overfitting

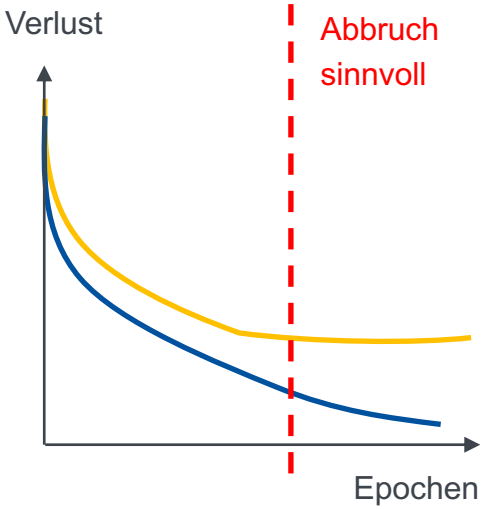


Erinnerung: Grad an Overfitting

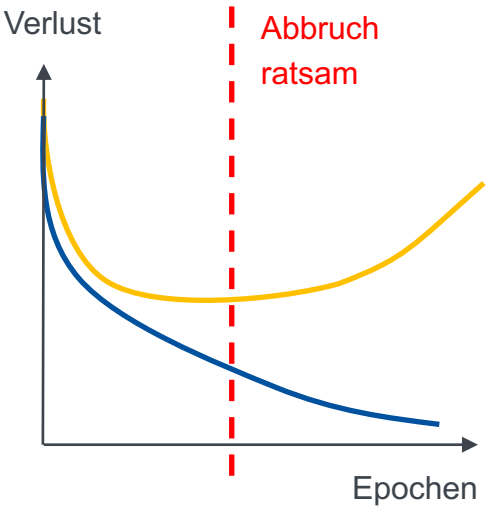
Early Stopping



Modell lernt noch Brauchbares (Verbesserung auf Validierungsdaten)



Modell lernt nichts mehr



Modell wird immer schlechter!

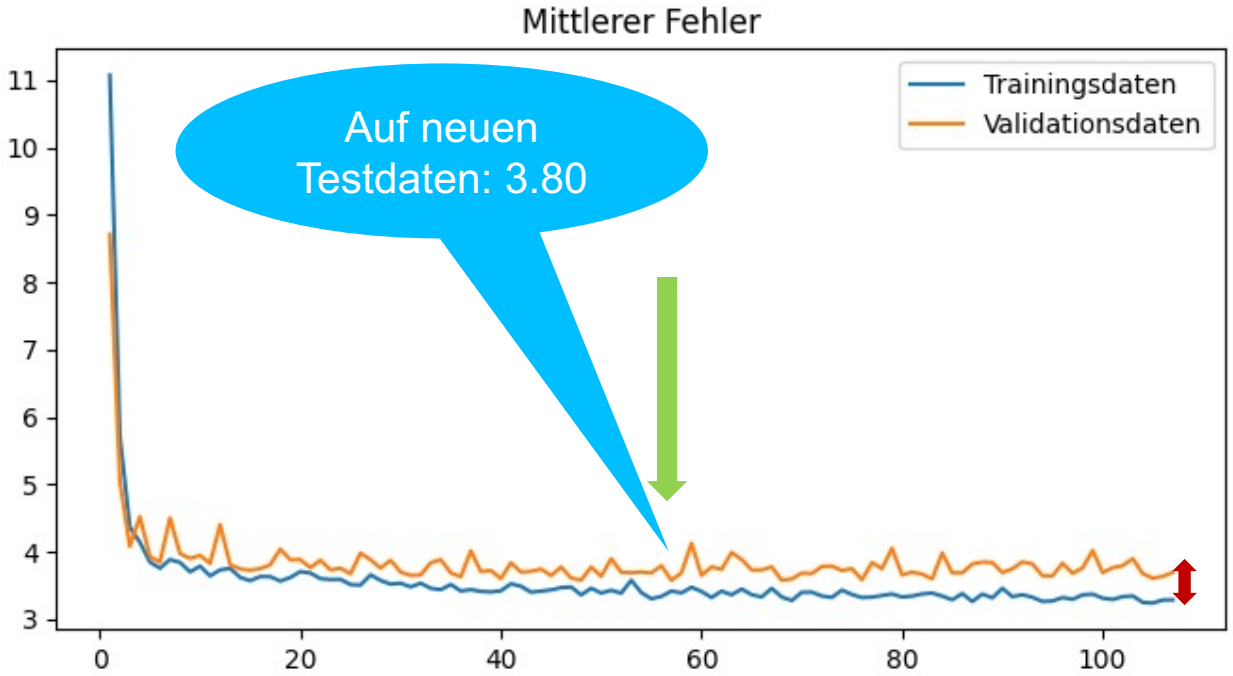
Erinnerung:

Early Stopping ist eine Technik, um Overfitting zu reduzieren.

Beim Early Stopping wird das Training beendet, wenn der Verlust auf den Validierungsdaten nahe legt, dass das Modell sich nicht mehr verbessert.

Mit Early Stopping

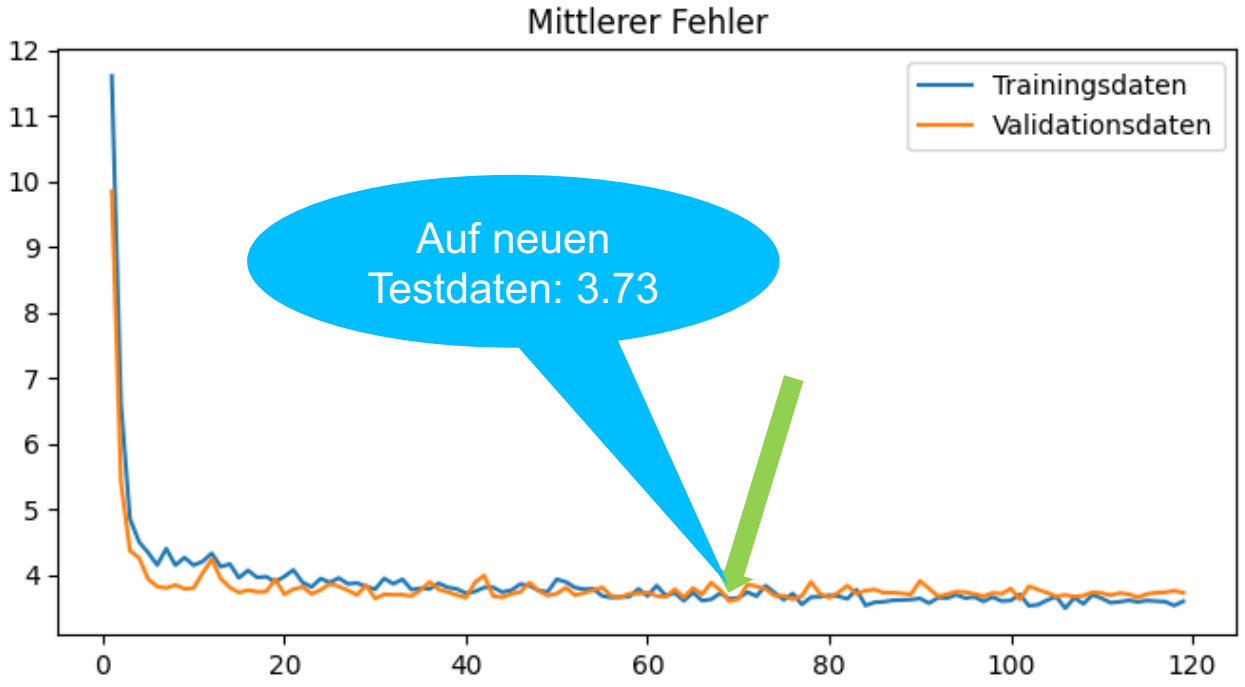
- Abbruch nach 107 Epochen, Zurücksetzen auf Modell aus Epoche 57
- Mittlerer Fehler 3.42 auf Trainingsdaten* 3.58 auf Valid.-daten
- Nach Abschluss: 3.22 auf Trainingsdaten
- Insgesamt noch wenig Overfitting



*für Trainingsdaten eigentlich: mittlerer Fehler über alle Batches der Epoche

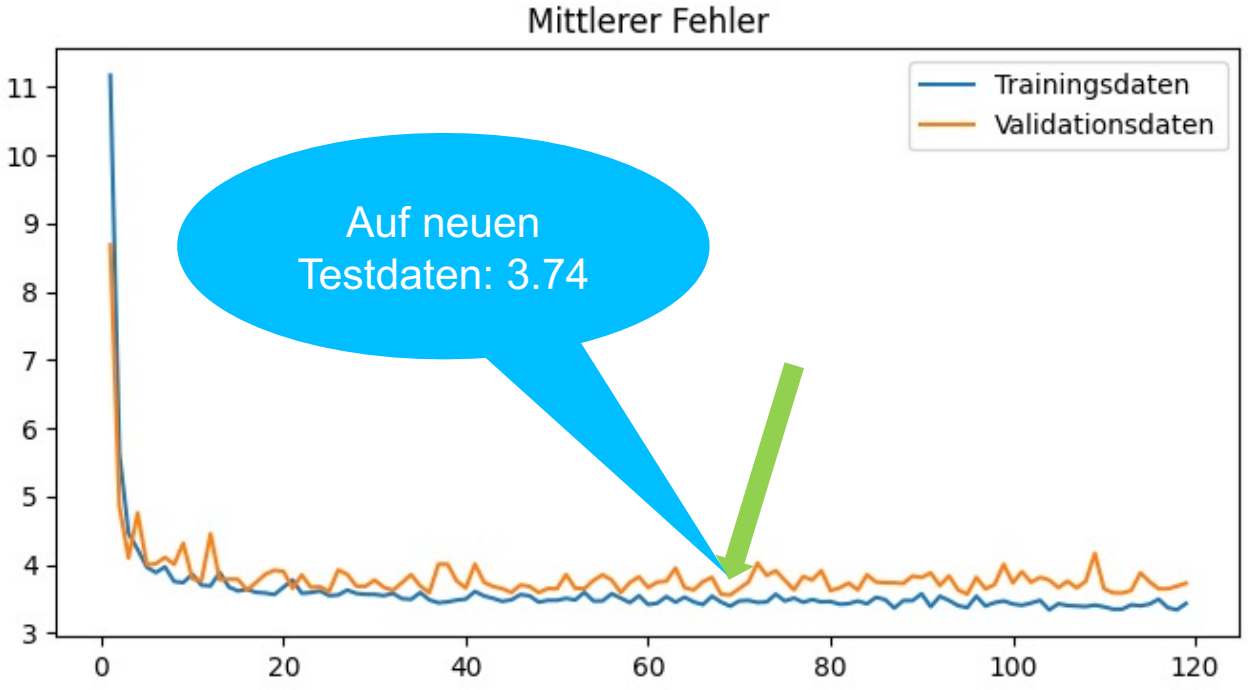
Mit Dropout und Early Stopping

- Mittlerer Fehler
  nach 69 Epochen
  3.63 auf Trainingsdaten
  3.60 auf Valid.-daten
- Dropout erschwert Vorhersage (nur für die Trainingsdaten)
- Ohne Dropout 3.31 auf Trainingsdaten



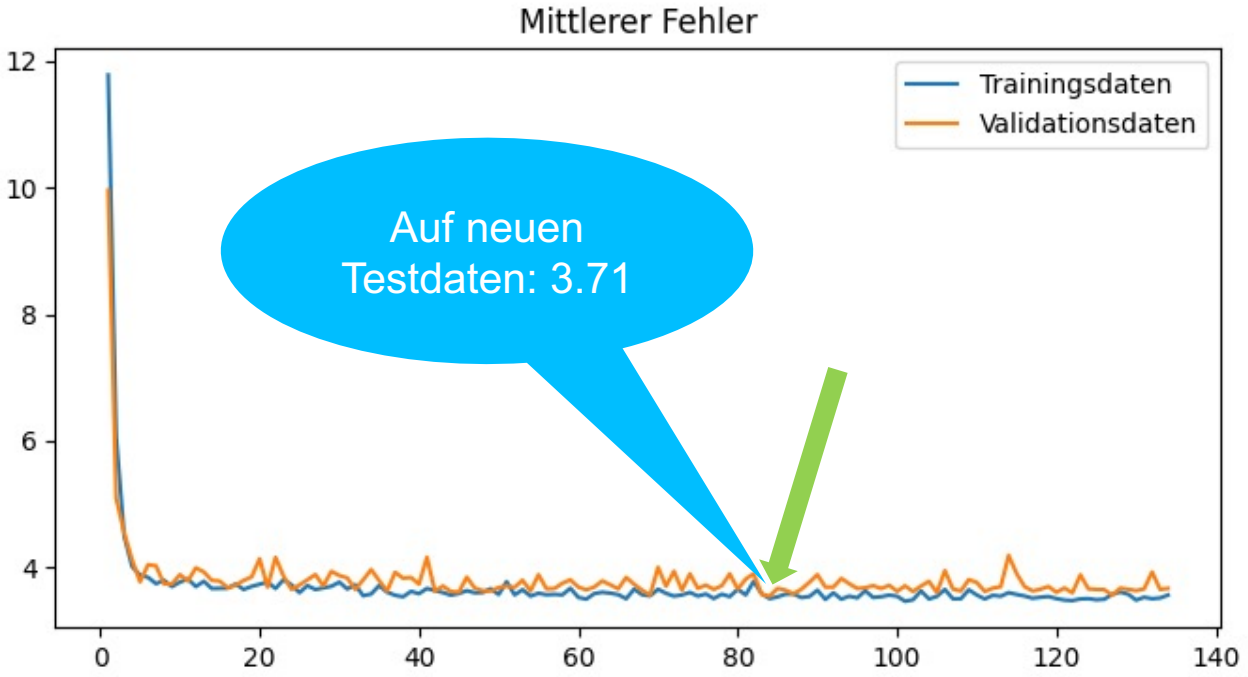
Mit L2-Regularisierung und Early Stopping

- Mit L2-Regularisierung
- Mittlerer Fehler
  nach 69 Epochen
  3.39 auf Trainingsdaten
  3.56 auf Valid.-daten
- Nach Abschluss
  3.28 auf Trainingsdaten



Mit L2-Regularisierung und Early Stopping

- Mit L2-Regularisierung
- Aber anderer Parameter für Regularisierung
- Mittlerer Fehler
  nach 84 Epochen
  3.50 auf Trainingsdaten
  3.54 auf Valid.-daten
- Nach Abschluss
  3.48



Zusammenfassung Overfitting bisher

- Kann die Performanz auf neuen Daten erheblich einschränken
- Macht die Modelle unnötig komplex/aufwendig
- Kann durch Techniken wie Early Stopping, Dropout, L2-Regularisierung reduziert werden
  - Mehr dazu später!
- Kann durch Verwendung von Validierungsdaten optimiert werden
  - Diese dürfen dann auf keinen Fall als Testdaten verwendet werden
  - Da ggf. Parameter der finalen Modelle so gewählt wurden, dass auf den Validierungsdaten optimale Ergebnisse erreicht wurden

Die Modelle von eben im Überblick

Wie zu erwarten: konsistent schlechter als auf Validierungsdaten

Zufällig besonders leicht

Hyperparameter	Training	Validierung	Test
Max. 1000 Epochen, Early Stopping mit Patience 50	3.22	3.58	3.80
Max. 1000 Epochen, Early Stopping mit Patience 50, Dropout (mit Parameter 0.2) nach den ersten vier Schichten	3.31	3.60	3.73
Max. 1000 Epochen, Early Stopping mit Patience 50, L2-Regularisierung mit Parameter 0.001	3.28	3.56	3.74
Max. 1000 Epochen, Early Stopping mit Patience 50, L2-Regularisierung mit Parameter 0.01	3.48	3.54	3.71

Je mehr Testdaten, desto zuverlässiger die Schätzung

Validierungsdaten werden häufig dazu verwendet, gute Hyperparameter zu finden.

Allerdings braucht man dann für eine objektive Evaluierung des finalen Modells neue (noch gar nicht gesehene) Testdaten.

Auf ungesehenen Testdaten ist immer eine etwas schlechtere Performanz zu erwarten.

Um Zufallseffekte der Auswahl der Testdaten zu minimieren (d.h. eine zuverlässigere Schätzung der Performanz zu erhalten), sollte man nicht zu wenige Datenpunkte als Testdaten verwenden.

Dies steht natürlich im Widerspruch dazu, möglichst viele Datenpunkte für das Training zu verwenden.

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

Quellenverzeichnis

Titelfoto: Riho Kroll (https://unsplash.com/de/@rihok), Nahaufnahme von Würfeln, auf Unsplash (https://unsplash.com/de/fotos/nahaufnahme-von-wurfeln-m4sGYaHYN5o), lizenziert unter Unsplash-Lizenz (https://unsplash.com/license). Bildausschnitt verändert.