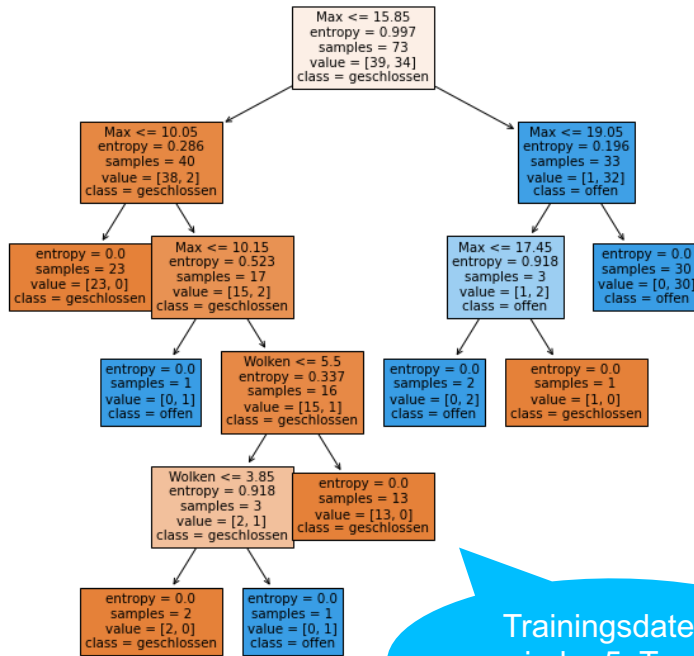


Training und Test

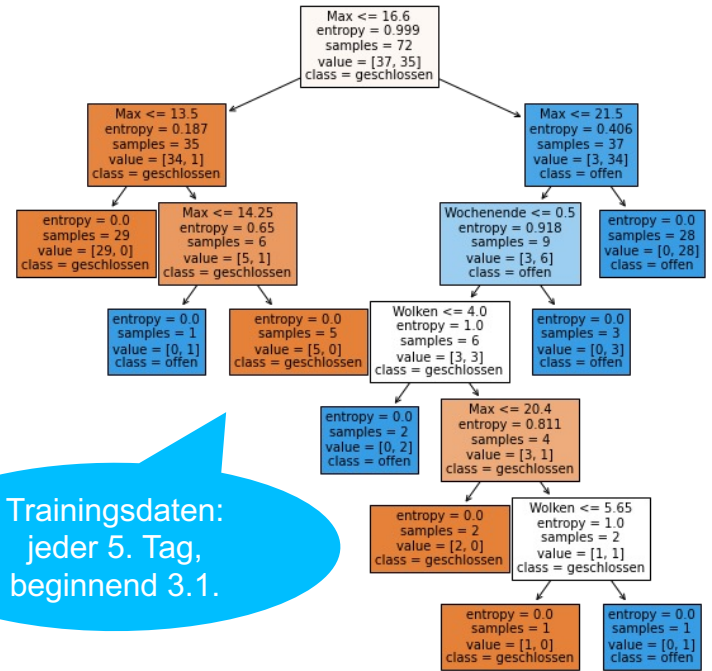
Wie man gute Modelle erstellt

Overfitting

Beispiel (siehe Training von Klassifikationsbäumen)



Trainingsdaten:
jeder 5. Tag,
beginnend 1.1.



Trainingsdaten:
jeder 5. Tag,
beginnend 3.1.

Evaluierung über Kreuz

Accuracy	Baum 1	Baum 2
Auf Daten 1	100 %	89%
Auf Daten 2	86 %	100 %
Precision (offen)	Baum 1	Baum 2
Auf Daten 1	100 %	90%
Auf Daten 2	80 %	100 %
Recall (offen)	Baum 1	Baum 2
Auf Daten 1	100 %	85%
Auf Daten 2	94 %	100 %

- Jeweils nur perfekt auf den Daten, die beim Training „gesehen“ wurden
- Zeigt begrenzte Generalisierbarkeit
- Auf den gegenseitigen Trainingsdaten jeweils ähnlich gut
- Typisches Muster für Overfitting

Overfitting erkennt man daran, dass die Performanz auf Testdaten deutlich unter der Performanz auf den Trainingsdaten liegt.

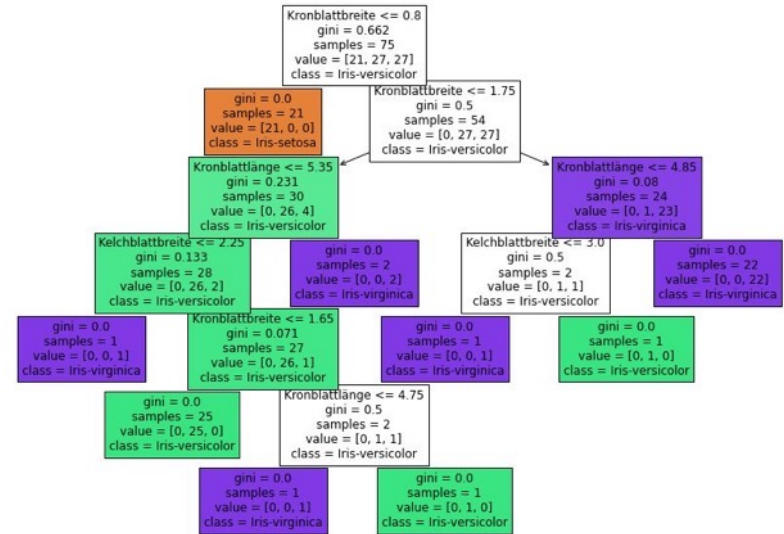
Evaluierung von Modellen

- Evaluierung von Modellen auf den Trainingsdaten nicht aussagekräftig
- Objektive Evaluierung nur auf nicht gesehenen Daten
- Daher immer Testdaten vorhalten
 - Diese sollte man streng genommen vorher nicht mal ansehen!

**Eine objektive Evaluierung von KI-
Modellen ist nur auf vorher nicht
gesehenen Testdaten möglich.**

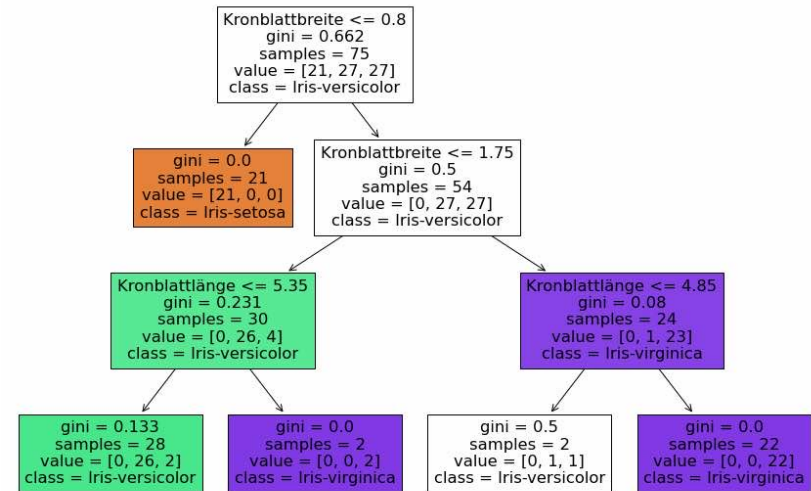
Wie schlimm ist Overfitting?

- Auch ein Modell mit Overfitting kann noch sinnvolle Ergebnisse liefern
- Accuracy
 - Auf Trainingsdaten: 100%
 - Auf Testdaten: 94,7%
- Auch auf Testdaten nicht schlecht...
- Aber:
 - Unnötig komplex
 - Es geht noch besser



Modell ohne Overfitting

- Accuracy
 - Auf Trainingsdaten: 96,0%
 - Auf Testdaten: 97,3%
- Weniger komplex
- Wahrscheinlich trotzdem besser als das vorige Modell mit 94,7%
- (Kleine Unterschiede können Zufall sein)



Overfitting macht die Modelle unnötig komplex und oft auch schlechter als ähnliche Modelle ohne Overfitting.

Die richtige Aufteilung

Repräsentativität der Daten

- Modelle für Eisverkäufe, aus Notebook
- Modelle trainiert auf
 - Daten von Anfang Januar bis Ende Februar
 - Daten von Ende Februar bis Mitte April
- Ausgewertet auf
 - Daten von Mitte April bis Ende Mai
- Jahreszeitliche Unterschiede in den Verkäufen
- Gewünscht: repräsentativ für Anwendung!

	Jan/Feb-Modell	Feb/April-Modell
Jan/Feb-Daten	19,16	33,47
Feb/April-Daten	63,71	43,87
April/Mai-Daten	110,40	54,60

Feb/April ist ähnlicher zu April/Mai

Ein weiteres Beispiel aus dem Notebook

- Aufteilung der Daten zur Klassifikation von Schwertlilien
- Erster Versuch:
 - Trainingsdaten überwiegend Iris Setosa und Iris Versicolor
 - Testdaten ausschließlich Iris Virginica
- Verteilung der Klassen in Trainings- und Testdaten sehr unterschiedlich
- Weder Trainings- noch Testdaten repräsentativ für Anwendung!

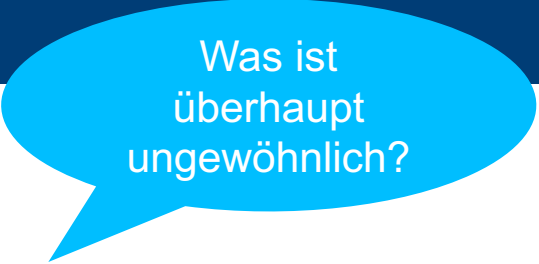
Die Aufteilung in Trainings- und Testdaten sollte so erfolgen, dass die Daten in beiden Mengen ähnlich sind.

Beide sollten möglichst repräsentativ für die bei der Anwendung erwarteten Daten sein.

Der Zufall im Spiel

Aufteilung in Trainings- und Testdaten per Zufall

- Wünschenswert (z.B. kein Effekt von Sortierung)
- Bei wenig Daten aber deutliche Zufallseffekte
 - Zufällig viele ungewöhnliche Daten in den Trainingsdaten -> schlechte Generalisierbarkeit!
 - Zufällig viele ungewöhnliche Daten in den Testdaten -> unzuverlässige Evaluierungsergebnisse
- Lösung: so viele Daten wie möglich
- Üblich: mehr Daten für Training als für Test (75-25 oder 80-20, evtl. auch 90-10)
- Bei Knappheit Kompromiss oder: siehe nächstes Thema



Was ist
überhaupt
ungewöhnlich?

**Mit mehr Daten werden die Modelle
oft besser und die
Evaluierungsergebnisse
zuverlässiger.**

Üblicherweise verwendet man den Großteil der Daten zum Training und nur einen kleineren Teil zum Testen.

**Eine typische Aufteilung
(engl.: Test-Training-Split)
ist 80 % Training, 20% Test.**



**Single Choice: Training und
Test**



**Multiple Choice: Trainings- und
Testdaten**

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung



Universität Stuttgart

Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering



Reutlingen | Tübingen | Zollernalb



Lizenzhinweise

„Training und Test“ 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 \(https://creativecommons.org/licenses/by/4.0/deed.de\)](https://creativecommons.org/licenses/by/4.0/deed.de)

(Namensnennung 4.0 International)

Quellenverzeichnis

Titelfoto: [Venti Views \(https://unsplash.com/@ventiviews\)](https://unsplash.com/@ventiviews), A woman runs up a trail during sunset, auf [Unsplash \(https://unsplash.com/photos/l1EWTM5mFEM\)](https://unsplash.com/photos/l1EWTM5mFEM), ist lizenziert unter [Unsplash-Lizenz \(https://unsplash.com/license\)](https://unsplash.com/license).

Bildausschnitt verändert.