Trainingsdaten

… und ihre Rolle beim Maschinellen Lernen

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

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

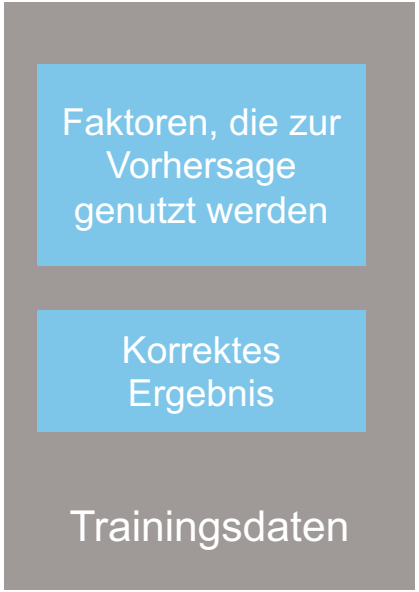
Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Zwei Beispiele für Maschinelles Lernen

Regressionsproblem (Eiscremespiel)

- Historische Daten zu
  - Temperatur
  - Wochentag
  - Bewölkung

- Eisabsatzzahlen



Klassifikationsproblem (Eisstand öffnen?)

- Historische Daten zu
  - Temperatur
  - Wochentag
  - Bewölkung

- Öffnung oder Schließung des Eisstands

Überwachtes Lernen

Beim überwachten Lernen braucht man als Trainingsdaten Beispieldaten, die sowohl die Faktoren beinhalten, die für die Vorhersage genutzt werden können, als auch das korrekte Ergebnis.

Woher kommen Trainingsdaten? – Beispiele bisher

- Wetter: aus Datenbank des deutschen Wetterdienstes
- Wochentag: aus einem Kalender
- Absatzzahlen: hat Yvonne abends von Hand eingetragen
  - Könnten auch digitalisiert vorliegen, wenn Kasse entsprechend ausgerüstet
- Öffnung/Schließung war ursprünglich nicht vorhanden
  - Musste nachträglich „**annotiert**“ werden
  - In unserem Fall automatisch: geschlossen, wenn Absatz < 100

Die nachträgliche Ergänzung von Merkmalen in Daten nennt sich Annotation.

Häufig handelt es sich dabei um das Merkmal, das gelernt werden soll.

Woher kommen Trainingsdaten – generell

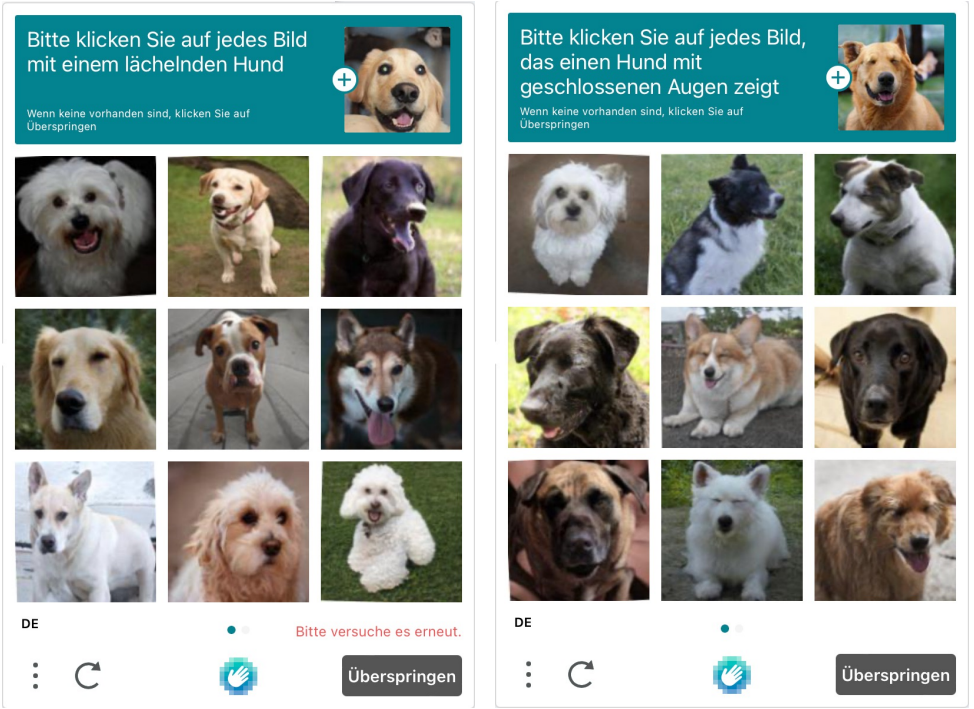
- Aus offenen Datenquellen (Open Street Maps, Wikipedia, Deutscher Wetterdienst, …)
- Von kommerziellen Anbietern
- Aus digitalisierten Prozessen in Unternehmen (z.B. Daten über Kunden durch Transaktionssysteme, Kundenkarten, Customer Relationship Management; Daten über Lagerbestände durch elektronisch erfasste Bestellungen/Anlieferungen; …)
- Von Sensoren aller Art
- Durch Befragung von Kunden/Zielgruppen
- Aus Kundenverhalten auf Webseiten (Tracking)
- Aus Social Media

Wie funktioniert Annotation?

- Häufig: manuell durch Experten
  - Beurteilung von Fotos oder Untersuchungsergebnissen (Hautkrebs ja/nein, Pflanzenkategorien, Hund vs. Katze, …), Transkription von Sprachdaten, …
- Implizit durch Nutzerverhalten/Resultat
  - Kunde kauft (nicht); Nutzer von Spracherkennung korrigiert Ergebnis (nicht)
  - Maschine läuft (nicht) weiter
  - Überschreiten eines Schwellwerts (Eisabsatz, Überhitzung, ...)
- Durch Crowdsourcing (viele Laien gemeinsam)
  - z.B. auf Plattformen, die Laien für solche Aufgaben bezahlen
  - z.B. Google reCAPTCHA (Kategorisierung von Objekten in Fotos)

„Crowdsourcing“ durch Google reCAPTCHA

- “Unfreiwilliges“ Crowdsourcing, um Dienste in Anspruch nehmen zu können
- Bezahlung für Dienste sozusagen durch Annotation von Daten



Von Menschen annotierte Daten sind besonders wertvoll.

Sie sind potentiell Trainingsdaten für Probleme, die bisher noch eher von Menschen erledigt werden.

Multiple Choice: Überwachtes Lernen

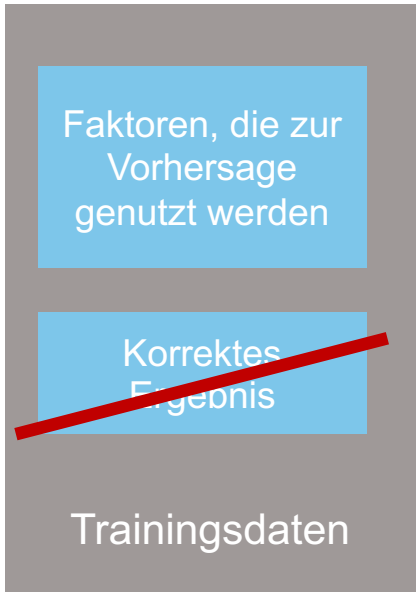

Fill in the blanks:
Trainingsdaten

Fill in the blanks: Annotation

Drag and Drop: Möglichkeiten für Annotation

Nicht überwachtes Lernen

Trainingsdaten für nicht überwachtes Lernen



- Beim nicht überwachten Lernen liegt kein „korrektes“ Ergebnis vor
- Weil noch nicht bekannt ist, was als „korrekt“ gilt
- Es geht darum, automatisch zu ermitteln, ob es in den Trainingsdaten irgendwelche Muster gibt

Beim nicht überwachten Lernen wird dem Lernalgorithmus kein gewünschtes Ergebnis vorgegeben.

Die Trainingsdaten enthalten also kein „korrektes“ Ergebnis.

Wichtigste Arten von nicht überwachtem Lernen

- Finden von Gruppen von ähnlichen Datenpunkten
- Finden von Merkmalen, die miteinander in Zusammenhang stehen
- Vereinfachen von Datenmengen

Clusteranalyse (Clustering)	Kundentypen ermitteln, ungewöhnliche Ereignisse finden
Assoziations-regeln	Empfehlungssystem: wer X kauft, kauft auch Y
Dimensionalitäts-reduktion	Visualisierung, Vorbereitung von überwachtem Lernen

Beispiel für Trainingsdaten bei Clustering

Tier	Farbe	Fortbewegung	Pelz	Atmung
Fledermaus	braun	fliegt	Pelz	Lunge
Fliege	schwarz	fliegt	kein Pelz	Tracheen
Amsel	schwarz	fliegt	kein Pelz	Lunge
Bär	braun	geht	Pelz	Lunge
Hase	braun	geht	Pelz	Lunge
Ratte	grau	geht	Pelz	Lunge
Schwein	rosa	geht	kein Pelz	Lunge
Elefant	grau	geht	kein Pelz	Lunge
Wal	grau	schwimmt	kein Pelz	Lunge
Forelle	grau	schwimmt	kein Pelz	Kiemen
Hai	grau	schwimmt	kein Pelz	Kiemen
Aal	grau	schwimmt	kein Pelz	Kiemen

Cluster 2

Cluster 3

Cluster 1

- Sinnvolles Clustering steht und fällt mit der Sinnhaftigkeit der Merkmale, anhand derer geclustert wird
- Hier z.B. Farbe wenig sinnvoll, wenn man hofft, durch Clustering Säugetiere von Fischen trennen zu können

Beispiel für Trainingsdaten bei Clustering

Tier	Farbe	Fortbewegung	Pelz	Atmung
Fledermaus	braun	fliegt	Pelz	Lunge
Fliege	schwarz	fliegt	kein Pelz	Tracheen
Amsel	schwarz	fliegt	kein Pelz	Lunge
Bär	braun	geht	Pelz	Lunge
Hase	braun	geht	Pelz	Lunge
Ratte	grau	geht	Pelz	Lunge
Schwein	rosa	geht	kein Pelz	Lunge
Elefant	grau	geht	kein Pelz	Lunge
Wal	grau	schwimmt	kein Pelz	Lunge
Forelle	grau	schwimmt	kein Pelz	Kiemen
Hai	grau	schwimmt	kein Pelz	Kiemen
Aal	grau	schwimmt	kein Pelz	Kiemen

Cluster 2

Cluster 3

Cluster 1

- Sinnvolles Clustering steht und fällt mit der Sinnhaftigkeit der Merkmale, anhand derer geclustert wird
- Hier z.B. Farbe wenig sinnvoll, wenn man hofft, durch Clustering Säugetiere von Fischen trennen zu können
- Haare wäre hilfreicher als Pelz

Clustering

Vorteile

- Entdeckung von Gruppen mit ähnlichen Merkmalen
- Insbesondere, wenn diese Gruppen vorher nicht bekannt sind und daher kein überwachtes Lernen möglich ist

Nachteile

- Ergebnisse hängen sehr von den in den Trainingsdaten vorhandenen Merkmalen ab
- Bei einigen Algorithmen muss zumindest die Anzahl der gesuchten Cluster vorgegeben werden

Das Zuordnen von Daten zu Gruppen, die Gemeinsamkeiten aufweisen, nennt sich Clustering.

Die gefundenen Gruppen werden als Cluster bezeichnet.

Clustering ist ein typisches Beispiel für nicht überwachtes Lernen.

Multiple Choice: Nicht-überwachtes Lernen

Drag and Drop: Arten von nicht-überwachtem Lernen

Fill in the blanks: Clustering

Reinforcement Learning

Trainingsdaten beim Reinforcement Learning?

- Man fängt „von null“ an
- D.h. es stehen keine Trainingsdaten zur Verfügung
- Statt dessen gibt es Belohnung/Bestrafung für gute/schlechte Ergebnisse
- Das System lernt erst bei der Anwendung

Beim Reinforcement Learning gibt es vorher keine Trainingsdaten.

Die Daten zum Lernen werden erst während des Trainings gesammelt.

Single Choice: Reinforcement-Learning

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

“Trainingsdaten" 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: Lars Kienle, „Fibre optic cable rack“, auf Unsplash, ist lizenziert unter Unsplash-Lizenz. Bildausschnitt verändert.