

ML - Machine Learning

Grundbegriffe von ML und LifeCycle von ML-Projekten

VideoConnect Inh. Josef Batki

Abstract – In diesem Aufsatz beschäftigen wir uns mit dem AI-Teilgebiet „Machine Learning“. Die wichtigsten Begriffe und Zusammenhänge werden erörtert und die Methoden bei der Entwicklung eines einfachen Modells diskutiert.

Darüber hinaus wird auch der Versuch unternommen allen interessierten Nicht-Datascientists einige Real-Life-Probleme des Geschäftslebens, die mit ML gelöst werden können, zu zeigen.

Im weiteren Textverlauf verwenden wir die internationalen (sprich: englische) Bezeichnungen mit Angabe der deutschen Entsprechung.

Warum benötigen wir ML?

Das exponentielle Wachstum digital erfasster und gespeicherter Daten verlangt nach Tools (unterstützenden Softwarewerkzeugen) für die automatische Analyse dieser riesigen Informationsmengen.

Das Ziel ist dabei das Verstehen der Funktionsweise sehr komplizierter Systeme und Prozesse - die diese Daten in rauen Mengen erzeugen - um zukünftige Abläufe besser vorhersagen zu können.

Der ökonomische Wert ergibt sich durch die aus den Daten gewonnenen Erkenntnissen.

Positionierung von Machine Learning (ML):

ML ist eine echte Teilmenge des Wissensbereiches AI und schließt selbst Deep Learning (DL), ebenfalls als echte Teilmenge, in sich ein. (s. Abb. 1)

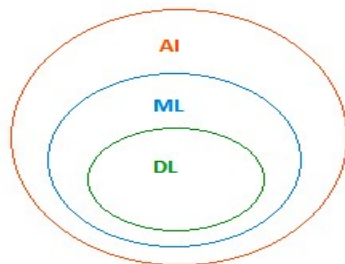


Abb. 1

Anfänglich war Artificial Intelligence (AI) nur ein Sammelbegriff für Computerprogramme/Algorithmen, die etwas machten, das früher nur durch menschliche Denkleistung hätte erbracht werden können. Dazu gehörten z.B. Schachcomputer oder Checkers-Player.

Beim maschinellen Lernen (ML) wird Computern beigebracht, wie sie bestimmte Probleme lösen können, ohne dass sie dafür explizit (*hard wired*) programmiert werden mussten.

Der Mensch lernt durch seine biologische Sensorik (Sinnesorgane) die Umwelt kennen, macht seine Erfahrungen, versteht die Zusammenhänge und erkennt Gefahren. Dieses Wissen kann er später vorteilhaft nutzen um besser überleben zu können.

Computersysteme stützen sich auf eine wohldefinierte Menge von Regeln, die **Algorithmen** genannt werden. Das angestrebte Ziel von ML ist entweder eine Vorhersage (**Prediction**) in der Form eines numerischen oder logischen Output-Wertes bzw. einer vordefinierten Kategorie, oder **Clustering** (Gruppierung von Daten nach bestimmten Kriterien).

Bei einer Vorhersage wird eine Eingangsdatenmenge (**set of input variables**), die zuvor mit geeigneten Labels markiert worden sind, auf die Ausgangsdaten (**output data**) in eindeutiger Weise abgebildet (**Supervised Learning**).

Wenn wir z.B. die Daten von Fahrzeugeigenschaften (Baujahr, Km-Stand, Extras, etc.) eines bestimmten Autotypen als Input nehmen und dazu die jeweiligen Verkaufspreise zuordnen, können wir dieses Modell, nach entsprechendem Training, für die Vorhersage des Preises von einem konkreten Auto verwenden.

Bei Clustering werden ähnliche Daten (z.B. bei Gesichtserkennung oder Beurteilung von Röntgenbildern) in Gruppen zusammengefasst und ihre Übereinstimmung durch eine Korrelationszahl (**Correlation Coefficient**) angegeben. Auf diese Weise erfahren wir wie stark die Beziehung bzw. Ähnlichkeit zwischen den Elementen dieser Datenmenge ist. Diese Vorgehensweise gehört zur Methode des „**Unsupervised Learning**“

Das Supervised Learning kann in **Regression** und **Classification** unterteilt werden.

Die **Regression** liefert eine Vorhersage in numerischer Form. Ein Beispiel dafür ist die Vorhersage zukünftiger Umsatzwerte. Trainingsdaten sind hier die Umsatzzahlen vergangener Zeiten.

Wenn bei der Vorhersage die Ausgangsvariable Teil einer vordefinierten Kategorie-Menge ist, sprechen wir von Klassifikation (**Classification**).

Auf diese Weise kann z.B. eine bestimmte Krankheit anhand der Input-Menge „Symptome“ mit Hilfe eines entsprechend vortrainierten Modells diagnostiziert werden.

Als Nächstes schauen wir uns an mit welchen Strategien ML-Modelle aufgebaut werden können.