Wie Maschinen lernen

Maschinelles Lernen wurde in den vergangenen Jahren erfolgreich zur Lösung komplexer Aufgaben, wie Gesichtserkennung oder autonomes Fahren, eingesetzt. Vorgestellt werden die potenziellen Methoden und die Anwendungsfunktionen für die öffentliche Verwaltung.

Autoren

Sebastian Hahn
ist IT-Consultant im Public Sector bei der msg.

Dr. Michael Scholz
ist Senior IT-Consultant im Public Sector bei der msg.

Im öffentlichen Sektor gewinnt maschinelles Lernen (ML) immer mehr an Bedeutung. Immer genauer und nachvollziehbarer sagen ML-basierte Softwaremodelle amtliche Handlungsbedarfe voraus und erhöhen so die Effizienz und Schlagkraft öffentlicher Verwaltungen. Der Grund: Durch gezielte Feedbackdaten erlernen ML-parametrisierte Modelle schnell und schlüssig neue Kontexte. Je nach konkreter Aufgabenstellung kommen in der öffentlichen Verwaltung Methoden der folgenden beiden Kategorien zum Einsatz:

- Supervised (überwachtes) Machine Learning analysiert Eingangs- und Ausgangsvariablen
- Unsupervised (unüberwachtes) Machine Learning nutzt nur Eingangsvariablen

Im überwachten Lernen trainieren Maschinen – per vorgegebener Struktur –, Ausgangsvariablen (Ausgangsdaten) vorherzusagen oder zu erklären. Damit eine Struktur modelliert werden kann, muss in den Trainingsdaten die Beziehung der Ausgangsvariablen zu mindestens einer Eingangsvariable enthalten sein. Zum überwachten Lernen nutzen Behörden vor allem Regression und Klassifikation. Mit diesen zwei Methoden erlernt das System die Parameter eines Modells, bis dieses anhand neuer Eingangsdaten genaue Vorhersagen liefert.

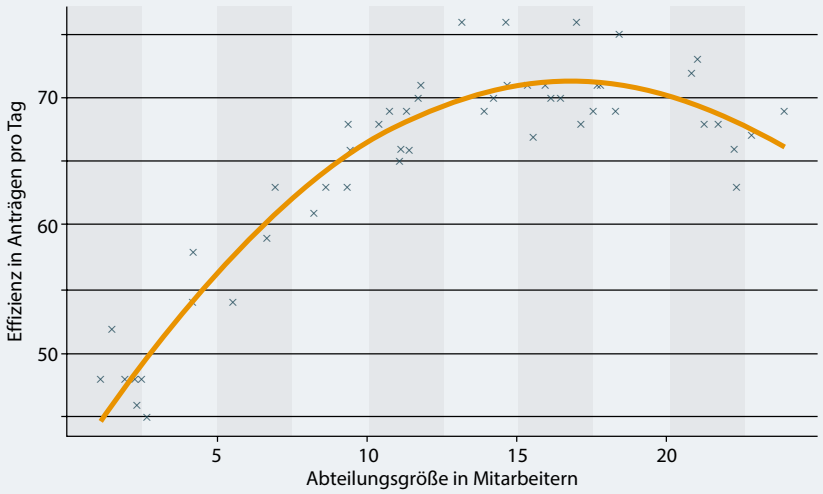
Eine Regression findet und definiert mathematische Zusammenhänge zwischen verschiedenen Eingangs- und einer Ausgangsvariable. Notwendige Parameter erlernt das Modell im Training. Behördlich eignet sich Regression für viele Zwecke. Sie prognostiziert, wie viele Anträge in einem Zeitraum auf ein Amt „zurollen“. Außerdem erklärt sie die Entwicklung von Kennzahlen und hilft dabei, Maßnahmen zu optimieren.

Ein Anwendungsbeispiel: Abteilungen mit zu vielen Mitarbeiterinnen und Mitarbeitern sind wegen des höheren Verwaltungsaufwands oft ineffizienter als Abteilungen mit weniger Mitarbeitenden. Hat eine Abteilung jedoch zu wenig Mitarbeiterinnen und Mitarbeiter, kann sie die täg-

Kompakt

- Die Methoden des überwachten und des unüberwachten Lernens bieten für Anwendungen der öffentlichen Verwaltung viele Einsatzmöglichkeiten.
- Ämter brauchen vielfältige Klassifikationsmethoden.
- Clustering beschreibt Datenobjekte durch wenige Variablen. Dadurch reduziert sich die Datenmenge.
- Um textuelle Daten zu ergründen, gehen Ämter immer mehr thematisch vor und setzen beim Ermitteln der Segmente auf Topic-Modeling.

Abbildung 1: Zusammenhang zwischen Abteilungsgröße und Effizienz



Quelle: Eigene Darstellung

lich anfallenden Aufgaben nicht mehr abarbeiten. Dank Regression lässt sich aus realen Daten von Abteilungen und deren gemessener Effizienz schnell der Zusammenhang zwischen Abteilungsgröße und Effizienz erkennen und das Optimum bestimmen (siehe Abbildung 1).

Klassifikation erkennt Betrug

Klassifikation dient der Zuordnung vom Datenobjekt zu den im Voraus festgelegten Klassen. Eine Zuordnung der Datenobjekte kann – ähnlich zur Regression – per Zusammenhang zwischen diversen Eingangs- und der Ausgangsvariable (Klasse) erfolgen. Alternativ kann auch ein Modell erlernt werden, das die Grenzen zwischen den beiden Klassen lernt. Wie treffend eine Zuordnung von Datenobjekten zu Klassen ist und wie nachvollziehbar diese Zuordnung anhand der Eingangsvariablen erklärt werden kann, hängt vom gewählten Ansatz ab.

Ämter müssen Antrags- und Steuerbetrug aufdecken, Dokumente und E-Mails nach Priorität, Thema oder Umgangston sortieren etc. Sie brauchen also vielfältige Klassifikationsmethoden. So führen vor allem falsch erklärte Einkommen oder andere Fehlangaben zu unzulässiger Genehmigung und Zahlung von Wohngeld oder zu unzureichender Besteuerung. Eine gute Klassifikation filtert hier – dank präziser Vorhersage – Anträge heraus, die genauer zu prüfen sind. Hierfür empfehlen sich Methoden, die genau herleiten, warum ein Antrag genauer zu prüfen ist, etwa anhand sinnvoll per Trainingsdaten errechneter Baumstrukturen wie im Wohngeld-Beispiel (siehe Abbildung 2).

Clustering schafft fehlende Kategorien

Unüberwachtes Lernen braucht weder Struktur noch Ausgangsvariable und trainiert das System komplett per Eingangsvariablen – Ziele sind Exploration von Daten oder Vorbereitung weiterer Datenanalysen. Clustering – als eine Methode des unüberwachten Lernens – teilt eine Menge von numerischen Daten in Segmente (Cluster) auf. Um jedoch textuelle Daten zu ergründen, gehen Ämter immer mehr thematisch vor und setzen beim Ermitteln der Segmente auf Topic-Modeling.

Fehlen anfangs vordefinierte Klassen, ist Klassifikation meist unmöglich. Ähnliche Datenobjekte sind also oft zu einer Gruppe (Cluster) zu bündeln. Hierfür – das ist wichtig – prüft jede Clustermethode auf andere Weise, ob sich zwei Datenobjekte derart ähneln, dass sie zum selben Cluster gehören. Zunächst lassen sich etwa je Datenobjekt ein Cluster erstellen und die jeweils ähnlichsten Cluster dann bündeln. Alternativ lassen sich Objekte auch anhand ihrer Distanz zu anderen Cluster-Einträgen zuordnen oder sind mit weiteren Methoden kombinierbar.

Ungeachtet des Ansatzes lassen sich Cluster, die für Eingangsvariablen entstanden, auch grafisch so darstellen, dass Machine-Learning-Architekten sie beschreiben können. Doch Clustering ergründet nicht nur Daten. Explorativ beschreibt Clustering Datenobjekte durch wenige Variablen. Dadurch reduziert sich die Datenmenge.

Ämter bauen per Clustering also Empfehlungssysteme für Dokumente oder Personen auf, entdecken Anomalien in Datensätzen oder ordnen Regionen spezielle Förder- und Schutzmaßnahmen zu.

Topic-Modeling ordnet Texte in Themensegmente

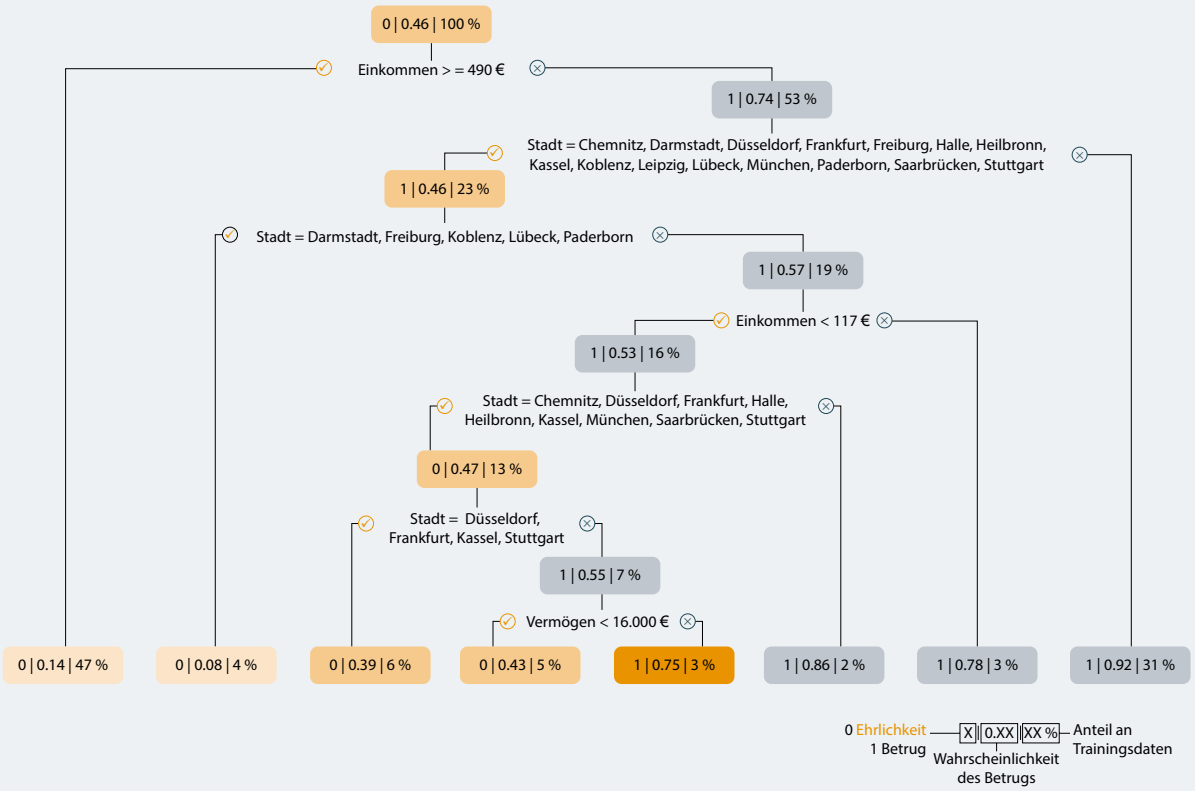
Oft sind Themen (Topics) noch unerkannt. Hier findet Topic-Modeling das Gemeinsame in Texten und leitet daraus Themen ab – erst das Modell erzeugt und definiert die Topics als solche. Topic-Modeling erkennt zwar keine Topics direkt (wie Einzelworte im Text), liest aber eine Häufung von Worten, die wahrscheinlich zum selben Thema zählen, als verdeckte Strukturen und modelliert sie zu Topics.

Topic-Modeling nutzt zudem oft Latent-Dirichlet-Allocation (LDA), um in einer Textsammlung zunächst jedes Wort aller Dokumente einem Zufallstopic zuzuordnen. Ist die Sollmenge vorerst zufälliger Topics erreicht, lässt sich die Aufteilung anhand folgender Fragen verfeinern:

- Wie oft erscheint ein Wort in jedem Topic? Aus dieser Sicht ist ein Wort eher nur einem bestimmten Topic (statt vielen) zugewiesen.
- Wie viele andere Worte desselben Dokuments erscheinen in jedem Topic? Aus dieser Sicht enthält ein Dokument eher ein oder wenige Themen.

Aus beiden Sichtweisen errechnet Topic-Modeling, wie wahrscheinlich ein Wort zu jedem einzelnen Topic passt. Als Li-

Abbildung 2: Kassifikationsbaum Wohngeld



Quelle: Eigene Darstellung

mit dient etwa eine vordefinierte Sollmenge an Durchgängen oder dass sich kaum noch etwas ändert. All dies eröffnet drei eng vernetzte Wege, um Dokumente zu lenken:

- Zuordnung von Worten zu Topics
- Zuordnung zwischen Worten und Dokumenten
- Beziehung zwischen Dokumenten und Topics

Topics helfen dabei, Texte nach Schlagwort, Segment oder Ähnlichkeit zu gruppieren. Behörden können dank Topic-Modeling also große Mengen an Dokumenten kategorisieren und per einfacher Variablen für weitere Analysen (etwa Klassifikation) verwenden.

Machine Learning toppt Blackboxes

Von Amts wegen ist maschinelles Lernen vielseitig einsetzbar. Gelernte Modellparameter lassen sich mit ML-Methoden zwar nicht vorhersagen – um aber Erklärungen oder Vorhersagen abzuleiten, lassen sich die gelernten Modelle – je nach Methode – oft sogar hoch aussagefähig deuten.

Zum Teil übertreffen Methoden, die Modelle interpretierbar machen, sogar Blackbox-Verfahren (zum Beispiel neuronale Netze). Werden sie mehrfach auf dieselben Daten (oder deren Teilmengen) angewandt, steigern sie oft sogar die Genauigkeit der Prognosen. Zur Bilderkennung sind neuronale Netze aber weiterhin hoch flexibel und schätzen Modelle – auch bei Millionen Eingangsvariablen – in akzeptabler Zeit. Denn schon ein 1.024 x 768 Pixel großes Bild hat 786.432 Bildpunkte mit je drei Eingangsvariablen (rot, grün, blau) – also fast 2,4 Millionen Variablen.

Um maschinelles Lernen zu etablieren, sollten Ämter mit Methoden beginnen, die Ergebnisse erklären und etwa Dokumente klassifizieren, Betrug aufdecken oder Prognosen erleichtern. Dann sollten sie die Modellparameter stetig anpassen oder mehrfach auf dieselben Daten oder deren Teilmengen anwenden, um Ergebnisse von sehr hoher Qualität zu erzielen. Erst im dritten Schritt sollten sie Blackbox-Methoden wie neuronale Netze nutzen oder unterschiedliche Methoden zu meist einem Blackbox-Ansatz kombinieren. ■

SpringerProfessional.de

Maschinelles Lernen

Lanquillon, C. (2019): Grundzüge des maschinellen Lernens, in: Blockchain und maschinelles Lernen, Berlin, Heidelberg, S. 89-141, www.springerprofessional.de/link/17442170