AraCom IT Services AG

#4 FACH ARTIKEL

WWW.ARACOM.DE

SUPERVISED LEARNING

AUGUST 13, 2020

„Der Mensch lehrt die Maschine. Das Ergebnis: Der Schüler übertrifft seinen Meister!“

Supervised Learning ist ein Verfahren, welches einem Computer ermöglicht mit Hilfe von gekennzeichneten (= gelabelten) Datensätzen Gesetzmäßigkeiten zu erlernen, zu trainieren und nachzubilden. Der Mensch fungiert als Lehrer und korrigiert gegebenenfalls Fehler manuell. Dadurch können Computer automatisiert neue Daten vorhandenen Daten zuordnen.

Supervised Learning oder auch überwachendes Lernen ist ein Teilgebiet des Machine Learning. Neben Reinforcement Learning und Unsupervised Learning stellt es eines der drei grundlegenden Paradigmen des maschinellen Lernens dar. Das Verfahren basiert auf gelabelte Datensätze, womit ein Supervised Learning Algorithmus trainiert wird. Der verwendete Algorithmus verbessert sich durch die wiederholte Anwendung des gleichen Prozesses auf bereits gewonnene Zwischenwerte. Zweck der Verwendung des Supervised Learnings ist die Automatisierung der Zuordnung von Daten zu vorhandenen Kategorien oder Klassen. Dies ist besonders sinnvoll bei zu verarbeitenden großen Datensätzen und kann viel Zeit ersparen, da das manuelle Bewerten durch einen Menschen einen hohen Zeitaufwand verursacht. Ist der Algorithmus einmal gut trainiert, übertrifft ein Algorithmus langfristig in vielen Fällen die Performance des Menschen.

INHALTSVERZEICHNIS

1. FUNKTIONSWEISE DES SUPERVISED LEARNING

Der Ursprung des Begriffs ergibt sich aus der Tatsache, dass der Lernprozess des Algorithmus vom Menschen überwacht wird. Der Anwender bestimmt einen geeigneten Datensatz, welcher von ihm gekennzeichnet wird und womit der Algorithmus anschließend trainiert wird.

Dieser Datensatz enthält bestenfalls eine große Menge an Key-Value Paare {(x1,y1),…,(xn, yn)}, welche die Menge der Trainingssets darstellen. Da der Anwender selbst diesen Datensatz bestimmt, sammelt und bereitstellt, sind ihm alle Trainingsdaten bekannt.

Der Lernalgorithmus leitet sich mit Hilfe der Input- und Output-Parameter (= KPIs oder Key-Value-Paare) eine Funktion ab, welche Element eines Hypothesenraums ist. Diese Funktion wird mit Methoden der mathematischen Statistik geprüft. Die Hypothese stellt eine Abbildung dar, die jedem Eingabewert einen Ausgabewert, welcher das Überwachungssignal darstellt, zuordnet. Ob der zugeordnete Output-Wert richtig ist, erfährt der Supervised Learning Algorithmus durch das Eingreifen des Menschen, indem dieser angibt, ob die ausgegebene Vorhersage richtig oder falsch ist. Schlussfolgernd sollte die Funktion des Algorithmus so approximiert sein, dass nach dem Training für bisher unbekannte Input-Werte (möglichst) genaue Output-Werte vorhergesagt werden können. [1]

Supervised Learning anhand eines Beispiels

Angenommen man hat drei verschiedene Obstsorten und es ist Aufgabe, die unterschiedlichen Obstsorten, bestehend aus Bananen, Wassermelonen und Erdbeeren zu klassifizieren.

Um ein Modell zu formulieren, müssen die einzigartigen Merkmale der Obstsorten gesammelt werden. Relevante Merkmale zur Kategorisierung des Obstes sind beispielsweise die Farbe, die Form und die Größe. Ist das Obst „rund", die Farbe „Grün" und die Größe „groß" handelt es sich um eine Wassermelone. Das Sammeln dieser Eigenschaften wird mit jeder weiteren Obstsorte fortgeführt und man erhält folgende Tabelle:

NO.	GRÖSSE	FARBE	FORM	OBST (Y-Wert)
1	groß	Grün	rund	Wassermelone
2	groß	Gelb	lang und gebogen	Banane
3	klein	Rot	herzförmig	Erdbeere

Abbildung: Obst Merkmal-Tabelle Supervised Learning

Mit den vorhandenen Merkmalen ist es nun möglich einen Algorithmus zu trainieren. Die Größe, Farbe und Formen stellen die Input-Werte dar und die diskreten Werte „Wassermelone", „Banane" und „Erdbeere" die Output-Werte.

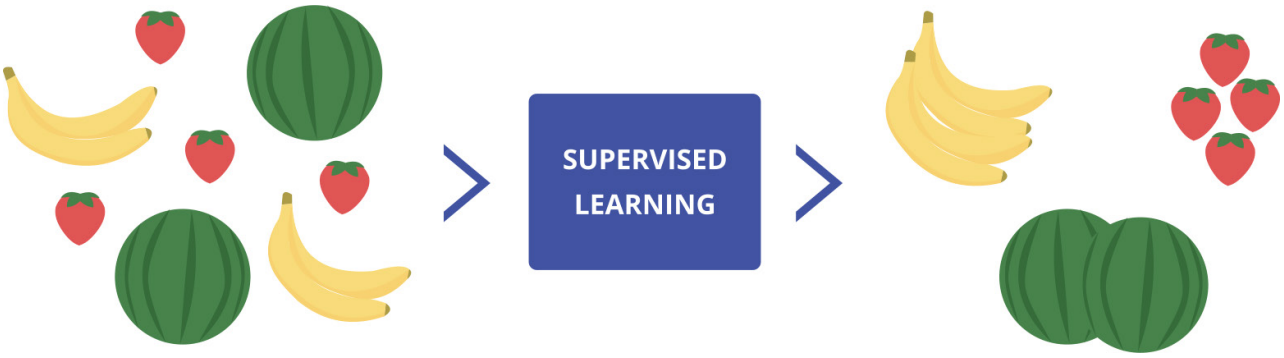

Abbildung: Obstsalat vorher | sortiertes Obst nachher

2. Verlustfunktion des Supervised Learning

Als Methode zur Bewertung eines Algorithmus wird im maschinellen Lernen eine Verlustfunktion verwendet. Sie zeigt die Performance des Algorithmus bei der Modellierung der Daten. Eine solche Supervised Learning Verlustfunktion oder auch „loss function" kann wie folgt aussehen:

$$loss\,(D) = \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} loss\,(X_i, Y_i)$$

Abbildung: Verlustfunktion Supervised Learning [2]

Bei Supervised Learning gibt es für jedes Xi aus einem Datensatz X einen entsprechenden vom Menschen bzw. Anwender gekennzeichneten Wert Yi aus einem Datensatz Y. Damit hätte man einen Satz von N gekennzeichneten Trainingsdaten D. Weichen die Vorhersagen sehr von den Ergebnissen ab, würde das Ergebnis der Verlustfunktion einen hohen Wert ausgeben und zeigen, dass das Modell schlecht abgeschnitten hat. Im Umkehrschluss bedeutet dies, dass ein Modell gut abschneidet, wenn das Ergebnis der Verlustfunktion einen niedrigen Wert annimmt. Nach Optimierungen kann der Supervised Learning Algorithmus die Abweichungen reduzieren und das Ergebnis der Verlustfunktion würde sinken. [2]

Hinweis: Es gibt nicht die eine „wahre" Verlustfunktion für die Gesamtheit aller Supervised Learning Algorithmen. Viele unterschiedliche Faktoren spielen bei der Wahl einer Verlustfunktion eine Rolle. Ein Faktor wird im nächsten Abschnitt näher beleuchtet.

3. Supervised Learning Kategorien: Lernprobleme

Die Wahl des passenden Supervised Learning Algorithmus und der damit einhergehenden Verlustfunktion hängt von den zu lösenden Problemen ab – diese lassen sich im Supervised Machine Learning in folgende Lernprobleme unterteilen: Regressions- und Klassifikationsprobleme.

Regressionsproblem

Die Ausgabevariable (Y-Wert) eines Regressionsmodells nimmt numerische Werte an. Ergänzend

wird sie als abhängige Variable bezeichnet und die Inputvariablen stellen die unabhängigen Variablen dar. Der Algorithmus soll in diesem Fall erlernen, Prognosen (Y-Werte) für Inputvariablen (X-Werte) auszugeben, welche möglichst nah am tatsächlichen Wert liegen.

KLASSIFIKATIONSPROBLEM

Bei der Ausgabevariable (Y-Wert) von Klassifikationsmodellen handelt es sich um einen diskreten Wert. Der Output (Y-Wert) kann in den häufigsten Fällen nur eine geringe Anzahl diskreter Werte annehmen. In diesem Fall wird – mit Hilfe mehrerer erklärenden Variablen – bestimmt, um welche Kategorie oder Klasse es sich bei den Input-Objekten handelt. [3]

Beispiel Obstsalat: Die Form, Größe und Farbe sind in diesem Fall die erklärenden Variablen und „Wassermelone", „Banane" und „Erdbeere" die diskreten Werte.

3.1 VISUALISIERUNG DER SUPERVISED LEARNING KATEGORIEN

Folgende Grafiken visualisieren die Regression und die Klassifikation:

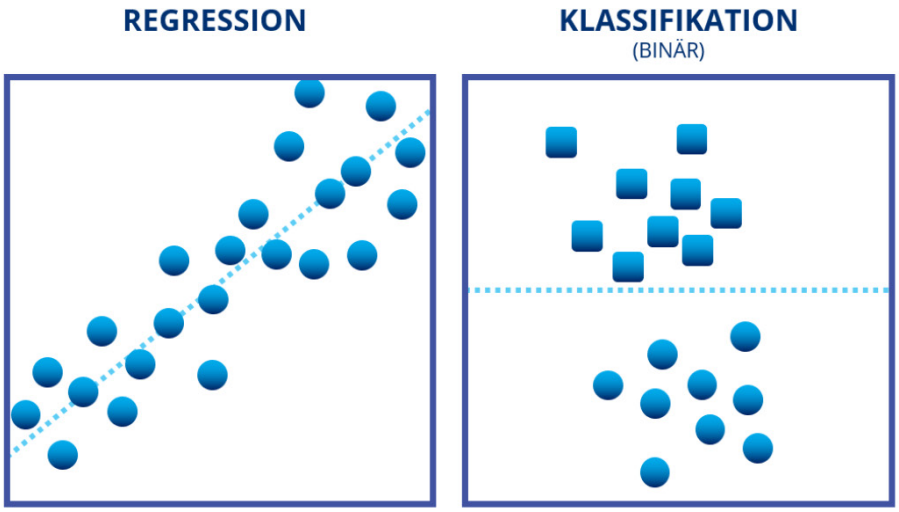

ABBILDUNG: REGRESSION UND KLASSIFIKATION

Das Regressionsmodell zeigt, dass sich das Lernproblem mit einer Geraden durch eine Punktwolke visualisieren lässt. Diese Gerade verläuft dabei besonders nah entlang der Datenwerte. In diesem Falle befinden wir uns in einem zweidimensionalen Raum und es handelt sich um eine lineare Regression. Das Ziel des Modells ist die Ausgabe eines möglichst präzisen numerischen Werts – meistens handelt es sich um die Annäherung an einen tatsächlichen Wert.

Das Klassifikationsmodell zeigt, dass die Datenwerte nur einer Klasse von einer begrenzten Anzahl an Klassen zugehören können – in diesem Fall handelt es sich um ein binäres Klassifikationsmodell, welches lediglich zwei Klassen als Output-Werte ausgeben kann.

3.2 Allgemeine Modell-Problematik

Im Supervised Machine Learning stößt man auf eine große Herausforderung: Die begrenzte Lernfähigkeit auf einen Datensatz. Ein perfektes Modell findet man nicht zwangsläufig auf Anhieb, weshalb auch im Nachgang Optimierungen nötig sind. Folgende Probleme könnten im Zusammenhang mit einem Supervised Learning Modell auftauchen:

Overfitting: Ist ein Modell überangepasst, bedeutet dies, dass das Modell auf die bisherigen Trainingsdaten spezialisiert ist. Das hat zur Folge, dass die Vorhersage des Modells mit einem Testsatz schlechter abschneidet als bei Vorhersagen mit den Trainingsdaten.

Underfitting: Im Umkehrschluss ergibt sich ein Underfitting, wenn das Modell nicht ausreichend an die Trainingsdaten angepasst wird. Das hat zur Folge, dass die Inputvariablen die Outputvariablen nicht hinreichend genug beschreiben und damit wäre eine hohe Verzerrung der Prognosen gegeben.

3.3 Regressions- und Klassifikationsarten

Mit folgenden **Regressionsarten** kann man beim Supervised Learning konfrontiert werden:

- **Lineare Regression:** Es existiert lediglich eine unabhängige Variable, welche zur Prognose der abhängige Variable verwendet wird.
- **Multiple lineare Regression:** Es existieren mehrere unabhängige Variablen, welche zur Prognose verwendet werden.
- **Polynomiale Regression:** In diesem Beispiel verläuft der Graph der Regression wie eine polynomiale Funktion. Damit existiert beim Graph mindestens eine Nullstelle. Die Werte sind somit nicht linear.

Beispiele für **Klassifikationsarten** beim Supervised Learning sind:

- **Binäre Klassifikation:** In diesem Fall werden übergebene Objekte nur zwei unterschiedlichen Klassen zugeordnet.
- **Multiclass-Klassifikation:** Objekte werden in mehr als zwei Kategorien eingruppiert.
- **Logistische Regression/Klassifikation:** Der Ausgabewert (Y-Wert) ist eine binäre kategoriale Variable. Das bedeutet der Wert nimmt entweder die Zahl 0 oder 1 an.

  Am Beispiel von einer Kreditbewilligung:
  1 = der Kredit wird bewilligt
  0 = der Kredit wird nicht bewilligt
- **Naive Bayes-Klassifikation:** Der Naive Bayes-Klassifikator basiert auf dem Bayes'schen Satz. Der Algorithmus ist besonders gut geeignet, wenn die Dimensionalität der Eingabewerte hoch ist. Der Klassifikator basiert darauf, dass eine bedingte Unabhängigkeit der Attributwerte vorhanden ist.

- **Klassifikationsbäume:** Diese Form der Entscheidungsbäume bestehen aus Knoten, Blättern und Zweige. Die Knoten entsprechen Entscheidungskriterien, die Blätter stellen die Entscheidungen dar und die Zweige sind die Merkmale, die zu den Klassen führen.

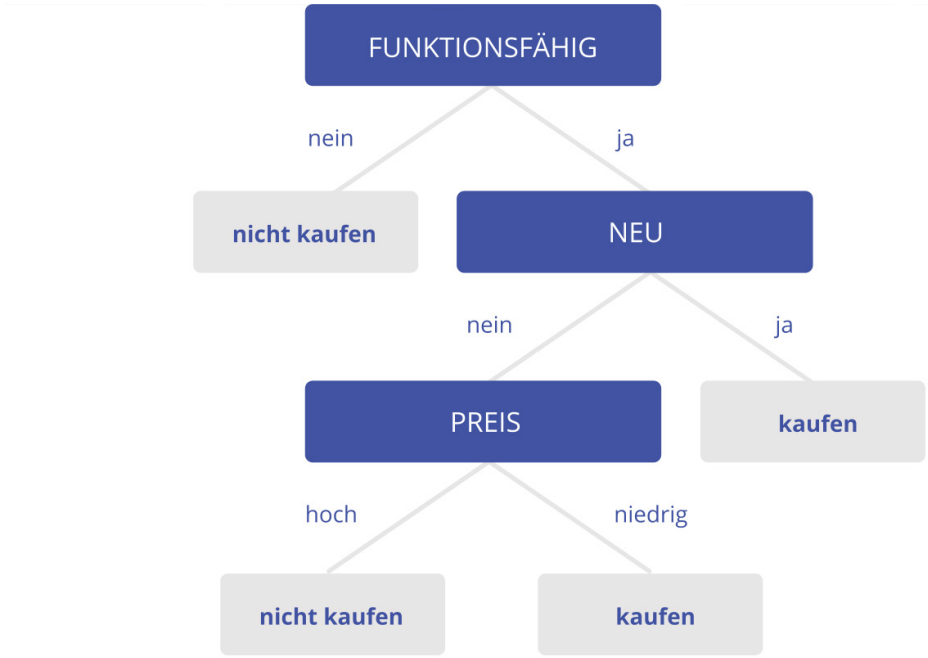

ABBILDUNG: ENTSCHEIDUNGSBAUM - KAUF EINES INSTRUMENTS

3.4 REGRESSIONSALGORITHMUS VS KLASSIFIKATIONSALGORITHMUS ANHAND VON BEISPIELEN

Im Folgenden werden mit Hilfe einfacher Beispiele die Abgrenzung zwischen Regressions- und Klassifikationsalgorithmen verdeutlicht.

REGRESSIONSALGORITHMUS: DER PREIS EINER IMMOBILIE

Mit Hilfe eines Multiple linearen Regressionsmodells kann der numerische Wert, wie der „optimale“ Preis einer Immobilie, bestimmt werden. Sowohl für Käufer als auch Verkäufer könnte diese Information vorteilhaft sein. Angenommen Sie möchten als Privatperson eine Immobilie verkaufen und möchten diese auf einer Onlineplattform anbieten. Welchen Preis wählen Sie? Den Preis könnte ein Supervised Learning Algorithmus für Sie bestimmen. Als Input-Werte könnten beispielsweise die Anzahl von Zimmern, die Wohnfläche, das Alter der Immobilie und auch die Lage innerhalb der Stadt verwendet werden. Daher ist eine Multiple lineare Regression notwendig und die „einfache“ reicht nicht aus. Diese Daten könnte man mit Hilfe bestehender Immobilienangebote sammeln. Daraufhin werden die Daten entsprechend gekennzeichnet,

dem Algorithmus übergeben und der Algorithmus trainiert. Nach dem Training sollte es im Idealfall möglich sein einen „marktüblichen" Preis für jegliche Immobilie auszugeben.

Regressionsalgorithmus: Zusammenhang zwischen Körpergröße und Alter

Ein linearer Regressionsalgorithmus hingegen könnte die Körpergröße eines beispielsweise 12-jährigen Jungen bestimmen. In diesem Fall stünden zwei Arten von Variablen zur Verfügung: Das Alter und die Körpergröße. Nach dem Training des Supervised Learning Algorithmus mit einer Vielzahl von Datenpaaren sollte, wenn dem Algorithmus ein Alter als Input-Wert übergeben wird, ein Wert ausgegeben werden, welcher die potenzielle Körpergröße darstellt.

Klassifikationsalgorithmus: Bilder kategorisieren mit Supervised Learning

Ein Supervised Learning Algorithmus könnte in der Lage sein, Bilder zu kategorisieren, indem die Trainingsbilder entsprechend gekennzeichnet werden. Dies könnte in einem binären Klassifikationsmodell wie folgt aussehen: Man hat eine große Bildersammlung von Wölfen und Löwen. Diese Bilder werden entweder mit der Kennzeichnung „Wolf" oder „Löwe" an den Algorithmus übergeben. Während des Trainings erlernt der Algorithmus Merkmale, welche den Wolf von einem Löwen unterscheidet. Um die Performance des Algorithmus zu testen, müssen ungekennzeichnete Bilder eines Löwen oder Wolfes übergeben werden. Dann sollte die korrekte Kennzeichnung der Bilder möglich sein – ist dem nicht so, sollte der Anwender manuell Fehler korrigieren und wenn möglich die Trainingsdaten erweitern.

Regressions- oder Klassifikationsalgorithmus: Kreditbewilligung im Bankenwesen

Die Kreditbewilligung hat im Falle der Klassifikation nur zwei mögliche Klassen, denen die Daten zugeordnet werden können – „bewilligen" und „nicht bewilligen". Abhängig von der Kredithöhe und der Bonität wird ein Kredit bewilligt oder eben nicht.

Hingegen bei der Regression wird die Kredithöhe in Abhängigkeit von der Bonität ausgegeben.

In beiden Fällen werden dem Algorithmus – wie gehabt – Input- und dazu passende Output-Werte übergeben und der Supervised Learning Algorithmus mit Hilfe des manuellen Eingreifens eines Menschen trainiert.

3.4 Fazit zu Supervised Learning

Ein Supervised Learning Modell wird durch die Erfahrungen und Datensätze des Menschen optimiert. Je mehr geeignete Datenpaare vorliegen, desto besser kann der Algorithmus trainiert werden, da dieser unter Berücksichtigung bisheriger Erfahrungen (inklusive der manuellen Korrekturen des Menschen) Ausgaben erzeugt. Dennoch kann das Training eines solchen Modells bei größeren Datenmengen einen hohen zeitlichen Aufwand bedeuten – allein die Aufbereitung der Trainingsdaten macht einen erheblichen Anteil aus. Ergänzend wird man mit allgemeinen Problemen von Modellen konfrontiert, wie das potenzielle Overfitting oder Underfitting. Um diesen Problemen entgegenzuwirken, sind mehrere Iterationen und Optimierungen des Modells nötig, was weiteren Zeitaufwand bedeutet. Im Gegensatz zu dem Reinforcement Learning Ansatz ist dieses Verfahren noch weiter entfernt von einer tatsächlichen „künstlichen Intelligenz", da die Anwendung eines solchen Verfahrens noch stärker vom Menschen abhängt – dennoch ist das Supervised Learning populär, da Anwender die Kontrolle über den gesamten Prozess behalten.

Wir bleiben gespannt, welche Entwicklungen das Supervised Learning in den nächsten Jahren erfährt.

QUELLEN:

1. **Mohri A., Rostamizadeh A., Talwalkar A. (2012).** Foundations of Machine Learning.

2. **Jing L., Tian Y. (2019).** Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey.

3. **James G., Witten D., Hastie T., Tibshirani R.** (2019) An Introduction to Statistical Learning.