Fallstudie 27: Maschinelles Lernen

# Was versteht man unter maschinellem Lernen?

Unter maschinellem Lernen werden Algorithmen zusammengefasst, welchen unter Verwendung eines Training-Sets (Lerndaten), also bestehenden Erfahrungswerten, neue Informationen generieren und diese anschließend verallgemeinern können. Dafür wird als Ergebnis des Lernalgorithmus ein Modell erzeugt, welches alle gewonnen und erzeugten Informationen umfasst und dazu genutzt werden kann diese auf weitere Daten anzuwenden und Vorhersagen (bspw. Klassifizierung) über diese zu treffen. Hierbei ist die Menge und Qualität der im Training-Set enthaltenen Daten von großer Bedeutung. Der Algorithmus lernt also mit zunehmender Erfahrung hinzu und kann somit seine Leistung bzw. Genauigkeit der Vorhersagen verbessern.



Abbildung 1 - Typischer Ablauf eines Maschinellen Lernalgorithmus

Diese Algorithmen werden häufig für Klassifizierungs oder Clusterung Aufgaben verwendet. Maschinelle Lernalgorithmen sind hierfür besonders gut geeignet, da sie in der Lage sind neue Muster innerhalb der Lerndaten zu erkennen und diese dann auf unbekannte Daten anzuwenden.

Das besondere an diesen Algorithmen ist, dass sie nicht speziell darauf programmiert werden müssen eine spezifische Aufgabe auszuführen oder ein spezifisches Training-Set auszuwerten. Sie lernen stattdessen selbst dazu und passen sich dementsprechend an die aktuelle Aufgabe an.

## Kriterien zum sinnvollen Einsatz von maschinellem Lernen

Auch wenn auf maschinellem Lernen basierende Lösungsansätze oftmals deutlich bessere Ergebnisse erzielen als vergleichbare Vorgehensweißen sollten und können sie nicht blind auf alle Probleme angewendet werden. Die Problemstellung sollte auf eine Reihe von Kriterien überprüft werden bevor maschinelles Lernen darauf angewendet wird.

### Messbarkeit der Performance

Die Performance bzw. der Erfolg der Berechnungen muss auf eine sinnvolle Art messbar sein. Ziel des Algorithmus ist es eine definierte Erfolgsmetrik zu optimieren. Wenn der Output nicht klar evaluiert werden kann wird der Algorithmus sich auch nicht verbessern können.

### Genügend Trainingsdaten

Bevor Aussagen über neue Daten getroffen werden können muss zuerst ein dementsprechendes Modell erzeugt werden. Um ein aussagekräftiges Modell zu generieren wird eine gewisse Menge an Trainingsdaten benötigt. Grundsätzlich lässt sich sagen, dass umso mehr Daten zur Verfügung stehen, umso besser das Modell wird.

### Qualität der Trainingsdaten

Eine große Menge an Daten ist jedoch nicht ausreichend um ein qualitativ hochwertiges Modell zu garantieren. Die Trainingsdaten sollten möglichst unterschiedlich voneinander sein und ein möglichst breites Spektrum der möglichen Ausprägungen darstellen. Wenn sich alle Datensätze sehr ähneln wird das Modell nur mit wenig Erfolg auf Datensätze mit anderen Ausprägungen angewendet werden können.

### Klar definierte Problemstellung

Zuletzt ist es noch wichtig zu beachten, dass das Problem an sich gut durch maschinelle Lernansätze lösbar ist. Diese funktionieren besonders dann gut, wenn Input und Output gut angegeben werden können. Sei X gegeben soll Y ausgegeben werden. Außerdem sind diese Ansätze besonders geeignet wenn Korrelationen in solch großen Datenmengen gesucht werden sollen, dass ein Mensch diese nicht erkennen könnte.

# Geschichtliche Entwicklung

# Grundbegriffe

### Ground truth

Die ground truth im Gebiet des maschinellen Lernens stellt das korrekte/gemessene Ergebnis dar. Bei einem Algorithmus zur Handschriftenerkennen von Zahlen ist die ground truth beispielsweise der Wert der zu erkennende Zahl.



Abbildung 2 - Handgeschriebene Ziffer "7"

In dem Beispiel in Abbildung 1 soll die Ziffer „7“ erkannt werde, somit wäre die ground truth hier ebenfalls die 7.

## Überwachtes(supervised) learning:

Überwachtes Lernen wird beispielsweiße für Klassifizierungsprobleme eingesetzt, da hierbei die Klassen bereits vorher bekannt sind und jedem Datensatz genau eine Klasse zugewiesen werden kann.

Beim überwachten lernen erhalten die Datensätze des Training-Sets jeweils ein Label, welches die korrekte Zuordnung des Datensatzes zu einer gegeben Klasse darstellt. Der Datensatz für das Beispiel in Abbildung 1 würde das Label „7“ erhalten, da eben die Ziffer „7“ erkannt werden soll. In den meisten Fällen entspricht das Label der ground truth.

Ziel des Algorithmus ist es, Zusammenhänge und Muster zwischen den restlichen Informationen des Datensatzes und dem Label zu erkennen und daraus ein Modell zu bilden, welches dafür genutzt werden kann um dieses Wissen auf Datensätze zu extrapolieren, welche kein Label enthalten. Dabei ist es wichtig, dass der Algorithmus nicht einfach sämtliche Datensätze „auswendig“ lernt, da er ansonsten bei unbekannten Datensätzen, welche nicht Teil des Training-Sets sind nur wenig erfolgreich ist.

## Unüberwachtes(unsupervised) learning:

Das unüberwachte Lernen hingegen, wird hauptsächlich für Clusteringprobleme eingesetzt bei denen die einzelnen Cluster/Gruppen nicht im Vorfeld bekannt sind. Dies bedeutet, dass hierbei im Gegensatz zum überwachten Lernen kein Label in den Datensätzen enthalten ist. Der Algorithmus muss sich die Gruppen selbst erstellen und die einzelnen Datensätzen selbst dazu zuordnen. Dafür muss der Algorithmus selbst, oft für Menschen nicht offensichtliche, Muster und Zusammenhänge zwischen den einzelnen Datensätzen erkennen. Es gibt also keine „richtige“ Lösung und das Ergebnis des Algorithmus kann im Vorfeld nicht bestimmt bzw. vorhergesagt werden.

Eine weit verbreitete Implementierung dieser Vorgehensweise zum Clustering ist beispielsweise der k-means Algorithmus.

### k-means Algorithmus

# Anwendungsgebiete

Weit verbreitete Umsetzungen dieser „lernenden“ Algorithmen sind beispielsweise Support-Vector-Machines(SVM), künstliche neuronale Netze (KNN) mit dem Untergebiet Deep Learning oder genetische Algorithmen.

## Support Vector Machines

## Künstliche neuronale Netze

### Deep Learning

## Genetische Algorithmen

# Konkreter Einsatz im Betrieb

# Prototypische Konzepte und Implementierung

## Prototyp Supervised

### Tensorflow

Zur Implementierung des Prototyps eignet sich aufgrund von einer umfangreicher Dokumentation, sowie einfacher Handhabung die Bibliothek Tensorflow

Tensorflow ist eine von Google entwickle Bibliothek für maschinelles Lernen mit dem Schwerpunkt auf Sprach- und Bilderkennung durch (tiefe) neuronale Netze. Sie basiert auf Python und C++, bietet jedoch auch eine API für Java, Go und einige weitere Sprachen.

Zusätzlich wird mit Tensorboard eine einfache Visualisierung der erzeugten Zwischen- und Endergebnisse angeboten.

/ANN,

## Prototyp Unsupervised

//K-means , Autoencoder etc….