Einsatz von Künstlichen Neuronalen Netzen in der Datenanalyse

Masterarbeit

zur Erlangung des akademischen Grades „Master of Science (M. Sc.)“ im Studiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Bock Nico Malte Alexander

[redacted] [redacted]

Prüfer: Prof. Dr. M. H. Breitner

Hannover, Mittwoch den 30. September 2020

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

Lernen ist anstrengend: wie praktisch wäre es dann, wenn eine Maschine das übernehmen könnte. Dieser Wunsch treibt Menschen voran - nicht erst seit es Computer gibt. Gleichzeitig hat die Technologie auf diesem Gebiet im vergangenen Jahrzehnt einen großen Fortschritt erlebt, sodass dies erstmals auch für komplexere Probleme möglich scheint. Glaubt man der These von Psychologen, dass Menschen primär aus Erfahrungen lernen, so stellt sich einem die Frage, was Erfahrungen überhaupt sind. Wird das doch relativ Blackbox-artige Erscheinen von Erfahrungen von außen betrachtet, so lassen sich hauptsächlich Sinneseindrücke als veränderliche Variablen unabhängig des genetischen Inputs identifizieren. Wenn man als Mensch versucht eigene Erinnerungen und Erfahrungen Revue passieren zu lassen, so sind diese keine Summe von Sinneswahrnehmungen, sondern lassen sich eher als ein Gesamteindruck darstellen, in dem wichtige Punkte herausstechen und unwichtige verblassen. Basierend darauf entsteht die Idee, dass Erfahrungen eine Art verarbeitete Aggregation von Sinneseindrücken sind. Überträgt man dies als Analogie auf einen Computer mit Sensoren als Sinne, so lässt sich Lernen als eine große Datenmenge verstehen, die zu Erfahrungen aggregiert und so auf das Wesentliche reduziert wird. Der klassische Weg so etwas mathematisch umzusetzen sind Regressionen; also Verfahren, bei den man eine Funktion bildet, die ein bekanntes Set an Punkten möglichst gut beschreibt. Solche Regressionen sind lange bekannt und und in ihrer Anwendung erprobt. Gleichzeitig ist die Komplexität und Anzahl der Operationen für ein Datenset meist überschaubar und lässt sich als Mensch manuell oder mit normalen Prozessoren durchführen.

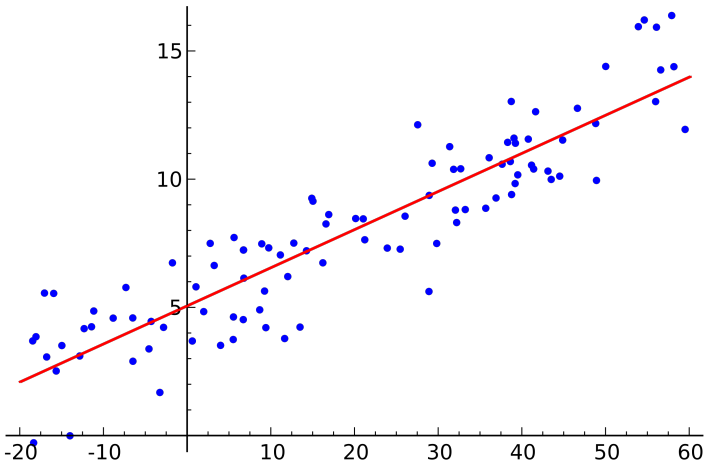

Abbildung 1 Regression Quelle: (Selbst erstellt)

Dabei berücksichtigen solche Verfahren immer nur einen singulären Einfluss einzelner Variablen, die sich zwar gegenseitig bedingen, gleichzeitig lassen sich komplexe Zusammenhänge mit solchen Verfahren aber weder erkennen noch beschreiben. Eine auf Mcculloch und Pitts 1943 zurückgehende Idee besteht daraus, aus solchen Regressions-

verfahren kleine Einheiten zu Netzwerken aufzubauen und somit auch komplexe Inhalte beschreib- und erkennbar zu machen. Während durchaus einige wissenschaftlich sehr interessante Erfolge mit der Anwendung dieser Algorithmen erzielt werden konnten, zeigten sich in den praktischen Anwendungen einige Limitationen - insbesondere auch in den notwendigen Berechnungen, sodass meist nur sehr kleine Netzwerke unter vertretbarem Ressourcen-Einsatz aufgebaut und angewendet werden konnten. Gegen Ende der 90er Jahre konnte durch inzwischen ganz andere neu verfügbar gewordene Rechenleistungen und einige neue mathematische Verfahren ein neuer Ansatz gefunden werden, sodass seitdem einige, viel diskutierte Forschungsergebnisse publiziert wurden.

2 Statistische Grundlagen

Die Analyse von Daten und die Gewinnung von Erkenntnissen und Einsichten aus diesen Daten wird in der heutigen Zeit und erst recht in der Zukunft in sehr vielen unterschiedlichen Bereichen der Wissenschaft und Wirtschaft immer mehr Relevanz beigemessen. Darüber hinaus werden durch zunehmende Digitalisierung von Prozessen immer mehr Daten gesammelt und stehen so für die Untersuchung zur Verfügung. Gleichzeitig hat sich die Rechen- und Speicherkapazität von Computern in den vergangenen Jahren so massiv verbessert, dass diese in der Regel keine begrenzende Ressource für Untersuchungen mehr darstellen.
Um qualifizierte Aussagen aus Daten lesen, generieren und auch vergleichen zu können, braucht es Standards und vergleichbare Methoden der Untersuchung. Dafür hat sich im Laufe der Geschichte die Statistik als Disziplin entwickelt, die vergleichbar zur heutigen Statistik seit ca. 200 Jahren existiert.
Die Analyse von Daten und die Erforschung von Analysemöglichkeiten ist dabei das Forschungsgebiet der Statistik. So sind auch Künstliche Neuronale Netze als quantitative Methodik ein neueres Forschungs- und Anwendungsgebiet der Statistik. Entsprechend kommen eine Summe grundlegender Begrifflichkeiten und Methoden, die in Künstlichen Neuronalen Netzen genutzt werden, aus der Statistik. Somit ist für ein solides Verständnis der Methodik und Terminologie auch ein entsprechendes statistisches Grundlagenwissen notwendig.
In diesem Kapitel wird in die für diese Arbeit notwendige Terminologie und Methodik eingeführt.
Dieses Kapitel kann dabei bestenfalls eine Ergänzung zu einer fundierten statistischen Vorbildung, nicht aber einen entsprechenden Ersatz darstellen.

4.8 Implikationen

Im Rahmen dieser Arbeit wurde beispielhaft an einem Datensatz untersucht, wie die Datenanalyse mittels Künstlicher Neuronaler Netze stattfinden könnte, um darauf Prognosen herzuleiten. Es waren umfangreiche Vorarbeiten und Vorverarbeitung der Daten notwendig, um diese in ein untersuchbares Format umwandeln zu können. In 4.7.3 bis 4.7.11 wurden verschiedene Versuche und Szenarien zur Datenanalyse mit den E-Scooter Daten durchgeführt. Dabei hat sich gezeigt, dass die Genauigkeit der Modelle von vielen unterschiedlichen Faktoren abzuhängen scheint und sich kaum pauschale Aussagen im Voraus treffen lassen. Grundsätzlich zeigte sich auch, dass nicht gut prognostiziert werden kann, mit welchen Stellschrauben die Qualität eines Modells zu verbessern wäre. Dadurch werden viele unterschiedliche Versuche nötig, um zu testen und auszuprobieren mit welchen Stellschrauben sich die eigenen Modelle verbessern lassen. Dabei scheint es so, dass entsprechende Erfahrung viel dabei helfen kann die richtigen Schritte für die nächsten Versuche zu identifizieren und so entsprechend die Anzahl der nötigen Versuche einzuschränken. Gleichzeitig zeigte sich im Rahmen der Versuche, dass die teilweise doch recht umfangreiche Berechnungszeit für das Training der Modelle einen echten Flaschenhals darstellen, da somit mit jedem Versuch entsprechende Kosten (in Form von Zeit) verbunden sind und dies die Möglichkeiten zum Experimentieren entsprechend einschränkt. Für die Anwendung im Rahmen dieser Arbeit wurden die Berechnungen nur auf einer CPU durchgeführt. Dabei sind die Berechnungen mittels derer Künstliche Neuronale Netze optimal trainiert werden stark parallelisierbar, weshalb es sich anbietet, diese in Umgebungen durchzuführen, die sich entsprechend gut parallelisieren lassen.

5 Schluss

Im Rahmen dieser Arbeit wurde der Einsatz von Künstlichen Neuronalen Netzen in der Datenanalyse untersucht. Künstliche Neuronale Netze sind mathematische Methoden, die seit ca. 70 Jahren bekannt sind, allerdings in den letzten 10 Jahren einen enormen Schub erlebt haben und seitdem für viele Lösungen eingesetzt werden, die im Allgemeinen als Künstliche Intelligenz bezeichnet werden. Künstliche Neuronale Netze haben dabei ihren Ursprung als eine Art mathematisch abstrakter Modellierung eines natürlichen Nervensystems. Insbesondere in jüngerer Forschung haben sie sich aber auch immer weiter von diesem Vorbild entfernt. So wurden unterschiedliche Wege entwickelt, Künstliche Neuronale Netze zu modellieren, die für bestimmte Problemstellungen jeweils besonders gut geeignet sind.
In dieser Arbeit wurde dafür auf statistische Grundlagen und wichtige Maßzahlen eingegangen. Es wurde die Begrifflichkeit das mashine-learning, als Obergruppe von Künstlichen Neuronalen Netzen erörtert und zu anderen Begrifflichkeiten abgegrenzt. Weiter

wurde in die theoretischen Grundlagen von Künstlichen Neuronalen Netzen eingeführt. Dafür wurden die geschichtliche Entwicklung, sowie die biologische Grundlagen erläutert. Weiter wurde auf aktuelle Methoden und einige Formen Künstlichen Neuronaler Netze eingegangen. Im Rahmen der Recherche zu dem Thema zeigte sich, wie umfangreich das Gesamtthema inzwischen geworden ist. Auch zeigten sich in der Recherche viele Stimmen zur Komplexität der Entwicklung und Anwendung Künstlicher Neuronaler Netze. So sind einfache neuronale Netze dank umfangreicher Programmbibliotheken bereits schnell entwickelt und zeigen auch teils beeindruckende Ergebnisse. Um diese aber weiter optimieren zu können, bedarf es umfangreicher Praxiserfahrung und ein gutes Problemverständnis.
Um das so erarbeitete theoretische Vorwissen anzuwenden und eigene Aussagen über Künstliche Neuronale Netze treffen zu können, wurde eine Case-Study durchgeführt. Im Rahmen dieser Case-Study wurden E-Scooter-Fahrten in Berlin analysiert. Als Grundlage waren Daten über verschiedene Scooterfahrten in Berlin mit Start- und Zielpunkt vorhanden. Berlin wurde dafür in verschiedene Gebiete unterteilt, wobei diese Gebiete abhängig von U-Bahn-Haltestellen so verteilt wurden, dass jeder E-Scooter der nächst liegenden Haltestelle zugeordnet wurde. Darüber hinaus wurden Zeitfenster definiert, innerhalb derer eine Fahrt stattgefunden hat und so aus dem Gebiet und dem Zeitfenster ein gemeinsamer Index gebildet. So entstanden unterschiedliche Indexe, denen ein Fahrtbeginn oder -ende zugewiesen werden konnte. Mittels der so entstandenen Datensätze wurden Künstliche Neuronale Netze modelliert und trainiert und anschließend Ergebnisse ausgegeben.
So zeigten sich in den ersten Versuchen schnell durchaus beeindruckende Ergebnisse, die sich aber für komplexere Szenarien nicht reproduzieren ließen.