Machine Learning und Prognosen

Umsetzung mit R im SQLServer 2017 anhand von Taxifahrten in New York

Bachelorthesis

Studiengang *Angewandte Informatik*

Duale Hochschule Baden-Württemberg Mannheim

von

Leonhard Applis

Abgabedatum:	26.09.2018
Matrikelnummer, Kurs:	2086307, TINF15/AI-BI
Ausbildungsfirma:	Atos Information Technology GmbH
Betreuer der Dualen Hochschule:	Prof. Tobi Chosen

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich meine Praxisarbeit mit dem Thema

> *Machine Learning und Prognosen Umsetzung mit R im SQLServer 2017 anhand von Taxifahrten in New York*

selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Fürth, den 5. Juni 2018

_____

Leonhard Applis

Abstract

Englisch Abstract to be done

title:	Machine-Learning and Prognosis
author:	Leonhard Applis
matriculation number:	2086307
class:	TINF15/AI-BI
supervisor DHBW:	???
supervisor Atos:	Jonas Mauer

Kurzfassung

Deutscher Abstract muss gemacht werden

Titel:	Machine Learning und Prognosen
Author:	Leonhard Applis
Matrikelnummer:	2086307
Kurs:	TINF15/AI-BI
Betreuer der Dualen Hochschule:	Prof. Tobi Chosen
Betreuer der Firma:	Jonas Mauer

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

DBMS	Database Management System
DHBW	Duale Hochschule Baden-Württemberg
SQL	Structured Query Language

1 Einleitung

Hier steht eine Einleitung, am besten etwas futuristisches wie gut Computer sind. Vielleicht was zu Terminator?

Hasta la vista, baby

1.1 Ziel der Arbeit

1. Grundlagen modernen Machine Learning Algorithmen erläutern
2. Möglichkeiten von Machine Learning im SQL-Server mit R darstellen, in einem Grad das der Leser es reproduzieren kann
3. Ein aussagekräftiges Fallbeispiel, das virtuelle Taxiunternehmen, ausarbeiten und die Algorithmen testen

Es gibt folgende Nicht-Ziele

1. Programmiersprachen (v.A. R) erklären
2. ML-Server mit Python (Aus Umfang eher unrealistisch)
3. ETL-Prozess detailliert schildern, es werden nur Daten dargelegt

1.2 Aufbau der Arbeit

Kapitel 2 dieser Arbeit bildet die Theorie zu modernen Ansätzen des Machine Learnings. Es werden die Algorithmen für lineare Regression, logistische Regression sowie Neuronale Netzwerke detailliert vorgestellt (In Reihenfolge der Nennung). Dieses Kapitel stellt einen rein theoretischen Teil der Arbeit dar, und beinhaltet keine Umsetzung der Algorithmen als Programme.

Darauf aufbauend werden in Kapitel 3 zunächst Grundlagen zu Microsofts SQL-Server 2017 und R geklärt, anschließend liegt der Schwerpunkt des Kapitels auf der Umsetzung von Machine-Learning Algorithmen in R. Innerhalb des Abschnittes 3.3 finden sich allgemeine Programme in T-SQL und R.

Kapitel 4 widmet sich der Umsetzung eines Fallbeispiels eines Taxiunternehmens. Zunächst werden in Abschnitt 4.1 die Ausgangslage der Daten sowie die Ziele des Fallbeispiels exakt definiert.

In Abschnitt 4.2 werden die Stammdaten des Taxiunternehmens und die Wetterdaten in Eigenschaften, Umfang und Bedeutung für Machine Learning dargestellt.

Der Abschnitt 2.3 behandelt die Verwendung der Programme aus 3.3 unter Bezugnahme auf das Fallbeispiel. Abschluss dieses Abschnittes bildet die Erzeugung eines Modells.

Abschnitt 4.4 benutzt das in Abschnitt 2.3 erzeugte Modell, um die in Abschnitt **??** vorgestellten Anforderungen zu bearbeiten. Hier befinden sich die eigentlichen Ergebnisse des Experiments sowie Programme die Prognosen durchführen.

Abschluss des Kapitels bildet Abschnitt 4.5 in dem Methoden zum Test eines Modells vorgestellt und durchgeführt werden.

Abschluss der Arbeit bildet in Kapitel 5 ein Fazit über die Qualität der Prognosen unter Berücksichtigung der Komplexität einzelner Teilaufgaben.

1.3 Voraussetzungen an den Leser

Innerhalb dieses Punktes werden die Kenntnisse abgesteckt, die der Leser für das Verständnis der Arbeit benötigt, welche **nicht** im Rahmen dieser Arbeit vorgestellt werden.

- **Mehrdimensionale Algebra:** Im Rahmen dieser Arbeit werden komplexe Algorithmen und Konzepte der mehrdimensionalen Algebra benötigt.

  Schwerpunkte liegen hier v.A. auf dem Lösen von mehrdimensionalen Gleichungen und Matrixoperationen. Der Umfang hierbei entspricht dem Besuch der Vorlesung *Mathematik II*.

Literaturempfehlung Lin.Alg

- **Stochastik:** Zur Bewertung der Algorithmen werden tiefere Kenntnisse der Stochastik und Statistik benötigt. Die benötigten Schwerpunktthemen sind Verteilungsfunktionen, Hypothesentests und Korrelation.

Literaturempfehlung Stochastik/-Statistik

- **R:** Die Programmiersprache R muss dem Leser im Umfang eines Basiskurses bekannt sein. Sie wird im Zuge der Arbeit verwendet, allerdings werden grundlegende Elemente nicht vorgestellt.
- **SQL:** Die Konzepte von SQL und der Dialekt von T-SQL sind in fortgeschrittenen Zügen benötigt. Die Verwendung von R innerhalb des SQL-Servers wird im Zuge der Arbeit vorgestellt.

2 Grundlagen zu Machine-Learning

Hier gebe ich ein Vorwort, wie heftig der Spaß hier wird und warum ich zuerst mit der Theorie Anfange.

2.1 Lineare Regression

Hier im Wesentlichen Stroetmann, das ist denke ich das beste. Was ist das und was macht es, warum ist das erstes Kapitel

2.1.1 Konzept und Ziele linearer Regression

Wofür brauche ich das, was kann ich damit machen, was kann ich damit nicht machen?

2.1.1.1 Beispiel

z.B. Beispiel mit Gerade durch Punkte ziehen, Beispiel sollte für einfache und allgemeine Lineare Regression brauchbar sein

Tabelle aus Werten, damit man später Funktion plotten kann udn mehr Ressourcen hat

2.1.1.2 Arten von Bias

Hier kurzer Text, was ein "Biasïst, und danach Erklärungen der Abweichungen [Pfe]

Bias aus Varianz Grundlegende Abweichungen der Sache aus harten Gründen - etwa Schwankungen im Wetter die einfach auftreten können

Selection Bias Abweichung wenn man seltsame/dumme Stichproben nimmt, z.B. Ernährungsumfrage auf Veganermesse

Confirmation Bias Wenn man (unterbewusst) Werte nimmt, die eine gewisses Schema erfüllen

2.1.2 Einfache Lineare Regression

Hier ist Lineare Regression auf einzelne Werte also
$R^1 -> R^1$

2.1.3 Allgemeine Lineare Regression

Hier ist die komplizierte Regression gemeint, wie wir sie brauchen also
$R^n -> R^m$
mit vielen Vektoren, Matrizen und tollen Dingen

2.1.4 Bewertung der Linearen Regression

Wie berechne ich die statistische Signifikanz meines Linearen Modells?

2.2 Klassifizerung

Hier vielleicht auch Stroetmann, oder etwas leichtgewichtigeres?

2.2.1 Konzept und Ziele von Klassifizierung

Hier Beispiel bringen, vllt Binäre Klassifizierung

2.2.2 Definitionen und Notationen

Features

Label & Klassen

Model

Accuracy

Supervised Learning

Unsupervised Learning

2.2.3 Digression: Gradientenanstieg

Erklärung was Stochastic Gradient Ascent ist, kurzes Vorgreifen warum man es braucht

2.2.4 Logistische Regression

2.2.4.1 Aktivierungsfunktion

Hier wird kurz erklärt was es für Aktivierungsfunktionen gibt, der Bezug zur Stochastic/Wahrscheinlichkeit und kurzes Vorgreifen warum man es braucht

Sigmoid Was ist das, was macht die, Eigenschaften beim Ableiten (Siehe Stroetmann)

Bild zur Sigmoidfunktion als Plot, ArcTang und Gauss daneben
Kurze Erklärung warum man nicht die anderen benutzt

ReLU Was ist das?? Wieso ist das SSo super gut"

2.2.4.2 Modell der Logistischen Regression

Wie spielen Aktivierungsfunktion, Optimierung und Lernen innerhalb der Logistischen Regression zusammen bei der Klassifizierung

2.3 Neuronale Netzwerke

Hier Aufbereitung, Anreicherung und Übersetzung von Selby [Sel]!

2.3.1 Modell künstlicher neuronalen Netzen

2.3.1.1 Historisches Konzept

Hier sagen, wie das mit dem Gehirn zusammen hängt, woher kommt der Name und warum spricht man heute von künstlichen NN

2.3.1.2 Aufbau des Modells

Klassisches Bild mit Input-Layers, Hidden Layers, Output Layers

Bezug auf Logistische und Lineare Regression, das diese NN ohne Hidden Layers sind

Zuordnung der Begriffe?

2.3.2 Hidden Layers

Wie wird das Modell der logistischen Regression erweitert?

Deep Learning = Mehr als 1 Hidden Layer

Manchmal: Deep Learning = Unsupervised Learning + viele Hidden Layers

2.3.3 Forward Propagation

Reinschmeißen eines Trainingsbeispiels und messen, wie schlimm es daneben liegt bzw. ob es daneben liegt

2.3.4 Backward Propagation

Nach dem Forward-Prograpagation nachjustieren der Gewichte in Matrix

Optimierungsfunktion und Gewichte

2.3.5 Training

Wie geht Training allgemein

worauf muss man bei Trainingsdaten achten, welche Größenordnungen sind notwendig

Anmerkung zu teilweise Export des Modells

3 SQLServer 2017 und R

In diesem Kapitel geht es dann um die Technische Realisierung der Dinge aus Kapitel 2 und noch erweiterte Grundlagen.

3.1 SQL-Server 2017

3.2 Programmiersprache R

3.3 Machine Learning im SQL-Server 2017

Hier bringe ich jeweils direkte Code-beispiel wie man sie in SQL-Server einfügen kann

Grundlagen sind ja schon geklärt hier, deswegen Verweise ich nur welche Funktion was erwartet und wie das dargestellt ist

Umsetzung in Python Hier sage ich kurz, wieso ich nicht genauer drauf eingehe, aber das alles geht

kurze Infos zu MLServer von Windows

Möglichkeiten in R Hier sage ich, dass und warum ich alles in R mache.

3.3.1 Lineare Regression

Bedingungen an Lineare Regression in R

Auflösen des Algorithmus in R (Programmcode)

Vllt Hintergrundwissen/Parameter wenn möglich mit Erklärung.

3.3.2 Klassifikation

Wie Lin. Regression

3.3.3 Neuronale Netze

3.3.3.1 Von Grund auf

Hier quasi den Code und Beispiel von [Sel] übernehmen und etwas aufbereiten.

3.3.3.2 Mit Package NNet

Ich denke, ich sollte das Package benutzen. Das haben totale Profis geschrieben.

4 Fallbeispiel: Prognosen eines Taxiunternehmens

Dieses Kapitel dreht sich um das Fallbeispiel der Taxidaten, sowie Ergebnisse

4.1 Ziele und Anforderungen

4.2 Eigenschaften der Daten

Innerhalb dieses Abschnittes werden zunächst die Daten vorgestellt, die dem Fallbeispiel zugrunde liegen.

Die vorgestellten Daten haben bereits einen ETL-Prozess durchlaufen. Dieser besteht im Wesentlichen darin, die CSV-Dateien dahingehend aufzubereiten, das amerikanische Nummerierungen (z.B. Angabe von Dezimalzahlen mit '.' anstelle von ',') auf europäische Normen gebracht werden. Prinzipiell entfällt dieser Schritt für eine rein amerikanische Umgebung.

4.2.1 Taxifahrten

Zunächst werden die Daten der Taxifahrten erläutert.

Diese stammen von der Stadt New York [Gova] und wurde von der *Taxi and Limousine Commission* (Kurz: TLC) bereitgestellt.

Die TLC stellt einen Dachverband mehrerer Taxiunternehmen dar und veröffentlicht die Daten nur - die Erhebung erfolgt in einzelnen, anonymisierten Kleinunternehmen.

Zusätzlich teilen sich die Fahrten in zwei Kategorien auf: *Green* und *Yellow*. Bei

grünen Fahrten handelt es sich um Fahrzeuge mit einer anderen Lizenzierung (vgl. [Giu] Absatz 5ff) und besonderen Auflagen. Im Allgemeinen verhalten sich die Fahrten allerdings gleich, insofern werden lediglich Unterschiede aufgelistet falls diese bestehen.

Attribute und Datentypen

Die folgende Übersicht entspricht der von der NYC bereitgestellten [Govb], die Beschreibung wurde übersetzt und eine Spalte für den Datentyp [1] ergänzt.

[1] Wie sie innerhalb des SQL-Servers bezeichnet werden

Name	Beschreibung	Datentyp
VendorID	Ein Code für das Taxiunternehmen, welches die Daten bereitstellt	smallint
pickup_datetime	Uhrzeit und Datum, wann die Fahrt begann	datetime
dropOff_datetime	Uhrzeit und Datum, wann die Fahrt endete	datetime
Passenger_count	Anzahl der Fahrgäste	smallint
store_and_fwd_flag	Angabe, ob die Fahrt direkt hochgeladen wurde, oder ob die Fahrt zwischengespeichert wurden vor einem Upload	bit
RatecodeID	Ein Code für die Rate, welche für die Taxifahrt bezahlt wurde	smallint
PULocationID	Ein Code für die Zone, in welcher die Fahrt begann	smallint
DOLocationID	Ein Code für die Zone, in welcher die Fahrt endete	smallint
trip_distance	Distanzangabe des Taximeters	real
fare_amount	Der Fahrpreis berechnet aus Zeit und Distanz	
extra	Verschiedene Zuschläge auf den Fahrpreis	real
MTA_tax	Aufschlag, automatisch erhoben bei entsprechender Rate	real
improvement_surcharge	Aufschlag, automatisch erhoben in bestimmten Zonen	real
payment_type	Angabe des Zahlungsmittels als Code	smallint
tip_amount	Höhe des Trinkgeldes	real
tolls_amount	Summierter Betrag von Zuschlägen dieser Fahrt	real
total_amount	Gesamtbetrag der Fahrt **ohne Trinkgeld**	real

Die Daten der Grünen Taxis sind erweitert um einen Code für den *Trip_Type* (Ob eine Fahrt von einem Taxistand begann oder ob die Gäste an der Straße abgeholt

wurden).

Alle Distanz-Angaben entsprechen amerikanischen Meilen (1 mile→1,6 km), alle Währungsangaben Dollar.

Für die Angaben der Codes sind ebenfalls Dictionaries bereitgestellt, diese spielen allerdings für den Machine-Learning-Aspekt dieser Arbeit keine Rolle und sind daher vernachlässigt worden.

Umfang

Aus Ressourcengründen wurde ausschließlich das Jahr 2017 betrachtet.

Es gibt **113 Millionen** Einträge für gelbe Fahrten, welche insgesamt knapp **8,1 GB Speicher** benötigen. Zusätzlich wurden Indizes angelegt mit weiteren 9,4 GB Speicher (Um schnelle Anfragen auf Uhrzeiten und Orte zu ermöglichen).

Es gibt **11,7 Millionen** Einträge für grüne Fahrten, mit insgesamt **900 MB Speicherplatz**. Es wurden zusätzlich Indizes mit 1,1 GB Speicher erstellt.

Zusammen gibt es aus dem Jahr 2017 also fast **125 Millionen Einträge** welche insgesamt 19,5 GB Speicher belegen.

Anomalien

Innerhalb der Daten traten einige Ungewöhnlichkeiten auf - zum Beispiel gibt es Fahrten, die von Ort A nach Ort A gingen und keine Strecke zurückgelegt haben. Bei einer genaueren Untersuchung ergab sich allerdings, dass diese Fahrten meist wenige Minuten dauerten und ebenfalls keinen Passagier hatten.

Es ist anzunehmen, das die Taxis an dieser Stelle auf ihre Passagiere gewartet haben. Aufgrund dieser Erkenntnis wurden alle Anomalien in die Machine-Learning Algorithmen übernommen, um die Daten und somit auch die Ergebnisse nicht zu verfälschen.

Es wurden außerdem weitere Anomalien gefunden, welche kurz genannt werden:

- Fahrten mit Negativkosten
- Fahrten außerhalb von 2017
- Fahrten mit extrem hohen Trinkgeld ($\sim$ 100$) oder extrem hohen kosten ($\sim$ 300$)
- Fahrten, die wenige Sekunden gedauert haben
- Fahrten, welche mehrere Stunden gedauert haben und dabei nur kurze Strecken zurücklegen

4.2.2 Wetteraufzeichnungen

In diesem Unterabschnitt werden die Wetterdaten sowie ihr Umfang vorgestellt.

Die Wetterdaten stammen von der *National Oceanic and atmospheric Administration* [NOA] (Kurz: NOAA), welche verschiedene Klimadaten sammelt. Für dieses Fallbeispiel wurden die Wetterdaten der Wetterstation des JFK-Airports für das Jahr 2017 abgefragt.

4.3 Erstellen eines Neuronalen Netzwerkes mit R

Hier packe ich meine Daten in die vorgestellten NN Algorithmen und Trainiere fleißig.

Quasi Umsetzung der vorgestellten Möglichkeiten

4.4 Prognosen mithilfe des Neuronalen Netzes

4.5 Test des Modells

4.5.1 Kriterien

4.5.2 Testfall

4.5.3 Ergebnisse

5 Fazit

Literaturverzeichnis

[Giu] GIUFFO, John: *NYC's New Green Taxis: What You Should Know*. https://www.forbes.com/sites/johngiuffo/2013/09/30/nycs-new-green-taxis-what-you-should-know/#5ca3d25732a2. – Erklärung was es mit den Grünen Taxis auf sich hat

[Gova] GOV, NYC: *TLC Trip Record Data*. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml. – Quelle der Taxidaten, angegebenes Date = Letztes Errata

[Govb] GOV, NYC: *TLC Trip Record Data*. http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf. – Data-Dictionary der gelben Taxidaten, angegebenes Date = Letztes Errata

[NOA] NOAA: *Land-Based Station Data*. https://www.ncdc.noaa.gov/data-access/land-based-station-data. – Quelle der Wetterdaten, angegebenes Date = Letztes Update

[Pfe] PFEIFER, Stella: *B wie Bias*. https://blog.eoda.de/2018/05/08/b-wie-bias/#more-4991

[Sel] SELBY, David: *Building a neural Network from scratch in R*. https://selbydavid.com/2018/01/09/neural-network/. – Schritt für Schritt Erklärung von NN's + Codebeispiele in R ohne Package