GML

Künstliche Neuronale Netze als Fortsetzung der Ökonometrie

Prof. Dr. Frank Lehrbass

Copyright

Modul- / Veranstaltungsgliederung

1	**Einführung**
	Lernziele
	Literatur
	Begriffe
	Big Data Beispiel
2	**Regressionsmodelle**
	Wiederholung Wissenschaftliche Methoden – Quantitative Datenanalyse
	Fallstudie Toyota Gebrauchtwagenpreise Ansatz
	Ansatz Laborexperiment Feature Selection
	Regressionsmodelle
	Bootstrap
	Schrittweise Regression
	Regularisierung
	Regressionsbäume
	Visualisierung
	Resumee
3	**Übung Weiterführung Toyota**

Modul- / Veranstaltungsgliederung

4	**Künstliche Neuronale Netze**
	Multi-Layer-Perceptron
	Historie *PRAXISBEISPIEL Futures Trading*
	Multi-Layer-Perceptron als nicht-lineare Regression
	Universal Approximation Theorem
	Gradient Descent
	Backpropagation & SGD
	Globale Modellgüte
	Sensitivitäten
	Architektur & Feature Selection
	Güte der Schätzer
	Diagnostik
	Resumee
5	**Common Problems: Overfitting and Underfitting**

1 Einführung

Pädagogische Einordnung dieses Skripts

- Dieses und die nachfolgenden Skripte dienen als Grundlage **seminaristischen Unterrichts in Präsenz.** Deshalb befinden sich Fragen an die Teilnehmenden (Tn) auf den Folien, die einen **Dialog und dadurch eine Aktivierung der Tn** bewirken sollen.
- Die Folien ersetzen nicht die verwendete Literatur, sondern vermitteln durch praktische Übungen wesentliche Erkenntnisse. Im Eigenstudium wird die Vertiefung / Nachbereitung der in Präsenz vermittelten Inhalte mit Hilfe der angegebenen Literatur geleistet.
- Der Präsenzuntericht ist für die Fachphase im Umfang von 30 UE geplant.Das Vorgehensmodell wurde seit 2018 für Data Science Inhalte praktisch erprobt.
- Folgende Dateien gehören zu diesem Foliensatz mit Prefix **GML_06_01 und 2**:

Datensätze	R Skripte	Literatur als pdf
Case Toyota (revised).csv	Case Toyota (revised)	01_01 RJournal_2010-1_Guenther+Fritsch
	Feature Selection	01_01 v85i11
		01_01 Convergent training.pdf

Modulziele

Die Studierenden können

- die wesentlichen **Methoden & Begriffe** der Business Data Science kennen,
- die **Anwendung der Methoden** beherrschen, d. h.
  - datengestützte Evidenzen über geschäftskritische Faktoren herstellen,
  - mit modernen Methoden des Machine Learning und
  - im Rahmen des Cross Industry Standard Process for Data Mining (CRISP-DM) arbeiten,
- **neuartigen Fragestellungen (Projektarbeit) begegnen** durch
  - die Planung eines Datenanalyse Projektes,
  - die Durchführung desselben,
- **Bewusstsein** über die **besonderen Erfordernisse von Big Data** haben.

Einordnung der Literatur und Benennung der Fachgebiete

Hinweis zur Einordnung

- Für Folien, die mit Ökonometrie überschrieben sind, konsultiert man bei der Nachbereitung die nachfolgend gelisteten “Econometrics” Bücher.
- Folien mit der Überschrift „Machine Learning“ beziehen sich auf die Bücher in rot/schwarz.
- **Fettgedruckte Buchtitel** sind besonders empfehlenswert.
- Grau hinterlegte Buchtitel sind Pflichtlektüre.

Für unsere Zwecke sind auch ältere Auflagen ausreichend. Diese können günstig, z.B. über Abebooks beschafft werden.
Im **Online Campus (OC)** stehen einige Bücher für FOM Studis als softcopy bereit.

Literatur

- Beck, M. W. (2108) NeuralNetTools: Visualization and Analysis Tools for Neural Networks, Journal of Statistical Software July, Volume 85, Issue 11.
- Davidson, R., MacKinnon, J. G. (2009) Econometric Theory and Methods
- Dorschel, J. (Hrsg.) (2015) Praxishandbuch Big Data
- Ghatak, A. (2019) Deep Learning with R
- Goodfellow, I., Bengio, Y., Courville, A. (2016) Deep Learning
- Hastie, T., Tibshirani, R., Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction
- Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2021) An Introduction to Statistical Learning with Applications in R
- Günther, F., Fritsch, S. (2010) neuralnet: Training of Neural Networks, The R Journal Vol. 2/1, June
- Kamath, U., Liu, J., Whitaker, J. (2019) Deep Learning for NLP and Speech Recognition
- Lehrbass, F. (2021): „Deep Learning Diagnostics – How to avoid being fooled by TensorFlow, PyTorch, or MXNet with the Help of Modern Econometrics“, Schriftenreihe des Instituts für Empirie & Statistik der FOM Hochschule 2021
- Lehrbass, F. (2021): „Deviations from Covered Interest Rate Parity: The case of British Pound Sterling versus Euro“ (mit T. S. Schuster), The Journal of Financial Data Science, Volume 3, Issue 1, 2021
- Maddala, G. S., Lahiri, K. (2009), Introduction to Econometrics
- Master, T. (1993) Practical Neural Network Recipes in C++

Literatur

- Ritz, C., Streibig, J. C. (2008) Nonlinear Regression with R
- Schlittgen, R. (2013) Regressionsanalysen mit R
- Mayerl, J., Urban, D. (2010) Binär-logistische Regressionsanalyse
- Seiter, M. (2017) Business Analytics. Effektive Nutzung fortschrittlicher Algorithmen in der Unternehmenssteuerung
- Schröder, M. (Hrsg., 2002) Finanzmarkt-Ökonometrie
- Shalev-Shwartz, S., Ben-David, S. (2014) Understanding Machine Learning: From Theory to Algorithms
- Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python.
- Taddy, M. (2019) Business Data Science
- Verbeek, M. (2012) A Guide to Modern Econometrics
- White, H. (1989): Some Asymptotic Results for Learning in Single Hidden Layer Feedforward Network Models, in: Journal of the American Statistical Association, 84, 1003-1013

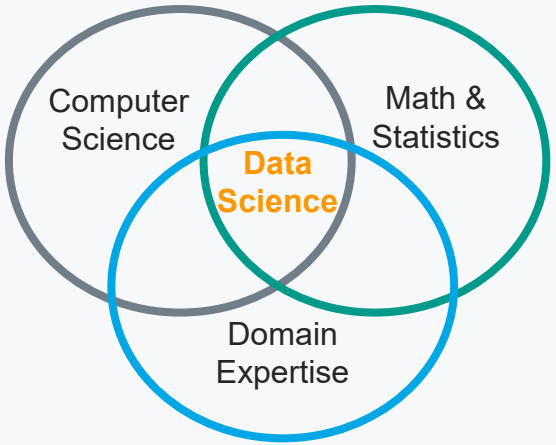
Begriffe

Business

In welchem „Business“ besitzen Sie „Domain Expertise“ (Fachwissen & Erfahrung)?

Data Science

Ist eine Schnittmenge: Data Science is a powerful combination of various disciplines.



Computer Science Skills

- Programming
- Big data technologies

Math & Statistics Knowledge

- Machine Learning
- Ensemble Models
- Anomaly Detection

Domain Expertise

- Business knowledge
- Expert Systems
- User testing

Was bringen Sie bei den ersten beiden „Sets“ aus dem Bachelor mit?

Quelle: Grafik https://insidebigdata.com/2017/07/27/defining-data-science-landscape/, Zugriff 9. Nov. 2021

*Lübke/Krol (2022) Empirisch-quantitative Abschlußarbeiten, in: Boßow-Thies/Krol (Hrsg.) Quant Forschung im Masterarbeiten, S 505.

ATOM = **A**ccept uncertainty, Be **T**houghtful eg know your assumptions and the **context**, Be **O**pen for criticism and transparent, **learn from** your **mistakes**, **M**odest

Auch die ATOM-Empfehlungen unterstreichen „die Bedeutung der Kombination von Methoden- und Fachwissen“*

Begriffe

Business Analytics ist die Kompetenz, verfügbare unternehmensinterne und –externe Daten so zu analysieren, dass konkrete betriebswirtschaftliche Frage/- Problemstellungen evidenzbasiert gelöst werden können, um für Unternehmen (dauerhafte) Wettbewerbsvorteile generieren zu können.

[…] Business Analytics kann daher als angewandte Data Science in der Domäne Betriebswirtschaft verstanden werden.

Wie wird diese Tätigkeit in Ihrem Unternehmen genannt?

Wie heißt die damit befasste Einheit?

Noch prägnanter in Shmueli et al., 2019, Data Mining for Business Analytics: „Business Analytics is the practice and art of bringing quantitative data to bear on decision making“ (P. 3)

Quelle: In Anlehnung an Seiter, M. Business Analytics: Effektive Nutzung fortschrittlicher Algorithmen in der Unternehmenssteuerung (German Edition) (S.1 & 20). 2017, Vahlen. Kindle-Version.

Begriffe

Big Data

„The characteristics of Big Data are commonly referred to as the four Vs:

Volume of Big Data

The volume of data refers to the size of the data sets that need to be analyzed and processed, which are now frequently larger than terabytes and petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities. In other words, this means that the data sets in Big Data are too large to process with a regular laptop or desktop processor. An example of a high-volume data set would be all credit card transactions on a day within Europe."

Quelle: https://www.bigdataframework.org/four-vs-of-big-data/, Zugriff 9. Nov. 2021

Begriffe

Big Data

Velocity of Big Data

“Velocity refers to the speed with which data is generated. High velocity data is generated with such a pace that it requires distinct (distributed) processing techniques. An example of a data that is generated with high velocity would be Twitter messages Facebook posts.

Variety of Big Data

Variety makes Big Data really big. Big Data comes from a great variety of sources and generally is one out of three types: structured, semi structured and unstructured data. The variety in data types frequently requires distinct processing capabilities and specialist algorithms. An example of high variety data sets would be the CCTV audio and video files that are generated at various locations in a city.”

Quelle: https://www.bigdataframework.org/four-vs-of-big-data/, Zugriff 9. Nov. 2021

Begriffe

Big Data

Veracity of Big Data

“Veracity refers to the quality of the data that is being analyzed. High veracity data has many records that are valuable to analyze and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data. The non-valuable in these data sets is referred to as noise. An example of a high veracity data set would be data from a medical experiment or trial.

Data that is high volume, high velocity and high variety must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. Because of these characteristics of the data, the knowledge domain that deals with the storage, processing, and analysis of these data sets has been labeled Big Data.”

Quelle: https://www.bigdataframework.org/four-vs-of-big-data/, Zugriff 9. Nov. 2021

Big Data Beispiel: Mieten in Düsseldorf

Mieten in Düsseldorf

Der Immoscout Datensatz

„Die Basis für die traditionellen Kennwerte beinhaltet Inserate von Mietimmobilien, welche durch die Immobilienseite Immoscout24.de angeboten worden sind. Bereitgestellt wird dieser Datensatz durch die Webseite Kaggle (www.kaggle.com). Kaggle ist eine Internetseite, welche Wettbewerbe für Data Science Herausforderungen anbietet.

Mithilfe einer Scraping-Methode, also einer Methode womit die Informationen direkt von der Internetseite extrahiert werden, wurden die Daten in den Zeiträumen 22.09.2018, 10.05.2019 sowie 08.10.2019 zusammengetragen und in einer kommaseparierten Datei aufgearbeitet. Diese Datei beinhaltet 268.850 Datensätze mit jeweils neunundvierzig Spalten und erstreckt sich inhaltlich bundesweit. Den größten Anteil der Spalten machen dreiunddreißig Stück aus, welche als Zeichenketten (englisch: Strings) formatiert sind. Zehn Stück sind numerisch formatiert und die letzten sechs Spalten haben boolesche Werte, welche die traditionellen Kennwerte repräsentieren.“

Quelle: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3835668, Zugriff 9. Nov. 2021

Big Data Beispiel: Mieten in Düsseldorf

Mieten in Düsseldorf

Raumbezogene Daten von Google

„Das Application Programming Interface (API) für die Google Places Daten (im Folgenden: Google Places API) ist eine Schnittstelle mittels welcher Abfragen an die Google Server gestellt werden können, um geografische Informationen rund um eine Geolokation, bspw. eine Adresse bestehend aus Straße und Hausnummer, zu erhalten. Dieser Dienst wird von Google angeboten und beinhaltet sämtliche Informationen zu einem jeweiligen Ort, welche bspw. ebenfalls in der Navigations-App Google Maps verwendet werden. […]

Diese API wurde schließlich zur Geokodierung von Adressen einzelner Mietobjekte .. sowie zur Abfrage von in der Nähe der jeweiligen Mietobjekte befindlichen POI .. genutzt.“

Quelle: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3835668, Zugriff 9. Nov. 2021

Big Data Beispiel: Mieten in Düsseldorf

Mieten in Düsseldorf

Abfrage von POI in der Nähe

„Nachdem die meisten Adressen korrekt einem Längen- sowie Breitengrad zugeordnet werden konnten, wurden für diese Mietobjekte die in der Nähe befindlichen POI abgefragt. Hierzu wurde eine Datenbank aus sämtlichen Abfragen der Google Places API erstellt. Eine Abfrage beinhaltete hierbei die jeweiligen Koordinaten des Mietobjektes, die Angabe eines Radius von 1.000 Metern in welchem die Google Places API nach entsprechenden POI suchen sollte sowie einen Suchbegriff nach welchem die Ergebnisse gefiltert bzw. ausgewählt werden sollten. Als Suchbegriffe wurden hierbei jeweils die folgenden drei Begriffe übergeben: Restaurant, Bankautomat, Haltestelle.

Dementsprechend wurden sämtliche Restaurants, Bankautomaten sowie Haltestellen in einem Umkreis von 1.000 Metern von einem jeweiligen Mietobjekt abgefragt.“

Quelle: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3835668, Zugriff 9. Nov. 2021

Big Data Beispiel: Mieten in Düsseldorf

Mieten in Düsseldorf

Zuordnung weiterer Daten

„Den einzelnen Mietobjekten werden im nächsten Schritt die entsprechenden nicht-traditionellen Daten als weitere Spalten im Datensatz hinzugefügt, indem mittels eines Algorithmus für jedes Mietobjekt sämtliche POI in einem Radius von 1.000m aus der Datenbank entnommen und hierauf unterschiedliche Aggregationsmethoden angewendet werden.“

„Die Generierung solcher Kennwerte erfolgt durch Nutzer bestimmter Internetseiten. Seiten wie Yelp, Tripadvisor oder Google steuern solche Daten bei.“

Welche Elemente von Big Data weist der Datensatz auf?

Quelle: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3835668, Zugriff 9. Nov. 2021

Begriffe

Business Analytics

„Die **Kompetenz**, verfügbare **Daten** so **zu analysieren**, dass betriebswirtschaftliche **Probleme evidenzbasiert gelöst** werden können, ist für Unternehmen eine zentrale Quelle von dauerhaften Wettbewerbsvorteilen."

„Business Analytics als komplexe Kompetenz erfordert eine **Vielzahl von Teilkompetenzen**

von Datenakquise und -aufbereitung bis Datenanalyse und -visualisierung. Dies begründet auch, warum Bücher für **verschiedene Zielgruppen** existieren. Eine Reihe von Büchern adressiert Spezialisten für die Analyse der Daten – in der Praxis bereits als „**Data Scientist**" bezeichnet […].

Diese Bücher zeichnen sich durch einen **Fokus auf mathematische Algorithmen** der Datenanalyse aus. Speziell für diese Zielgruppe hat sich ein Standardprozess zur Datenanalyse mit der Bezeichnung **CRISP-DM** herausgebildet".

Dieser Kurs hat genau diesen Fokus. Bevor wir den CRISP kennenlernen, schauen wir uns noch kurz die „umgebenden" Prozesse an.

Quelle: Seiter, M. Business Analytics: Effektive Nutzung fortschrittlicher Algorithmen in der Unternehmenssteuerung (German Edition) (S.14). 2019, Vahlen. Kindle-Version.

Begriffe

Business Analytics

„Der Aufbau „Datensammlung – Datenauswertung – Ergebnispräsentation“ ist für Führungskräfte eine untaugliche Unterteilung.

Die Fragen von Führungskräften lauten vielmehr:

- Für welche betriebswirtschaftlichen Probleme sollen die knappen Business Analytics-Ressourcen eingesetzt werden (vgl. Kap. 2)?
- Welche Ressourcen, insbesondere Daten, IT und Personal, müssen für die Lösung eingesetzt werden (vgl. Kap. 3)?
- Mit welchen Algorithmen werden Evidenzen zur Lösung der betriebswirtschaftlichen Probleme gewonnen (vgl. Kap. 4)?
- Wie sind Roh-Evidenzen aufzubereiten, damit diese durch die Führungskraft optimal zur Lösung der Problemlösung eingesetzt werden können (vgl. Kap. 5)?

Diese Fragen lassen erkennen, dass Datenerfassung und -analyse auch für Führungskräfte eine Rolle spielen. Allerdings nicht eine solch dominante Rolle wie für Data Scientists“.

Der Zertifikatsleiter ist zuversichtlich, dass auch ein Data Scientist diese Fragen beantworten kann. Zudem ist er überzeugt, dass ein Data Scientist erfolgreicher arbeiten kann, wenn er sich der Punkte 1-3 bewusst ist, z B als Projektleiter. Beim letzten Punkt hingegen lohnt sich definitiv die Hinzunahme von Grafikdesignern o.ä. Qualifikationen.

Quelle: Seiter, M. Business Analytics: Effektive Nutzung fortschrittlicher Algorithmen in der Unternehmenssteuerung (German Edition) (S.15). 2019, Vahlen. Kindle-Version.

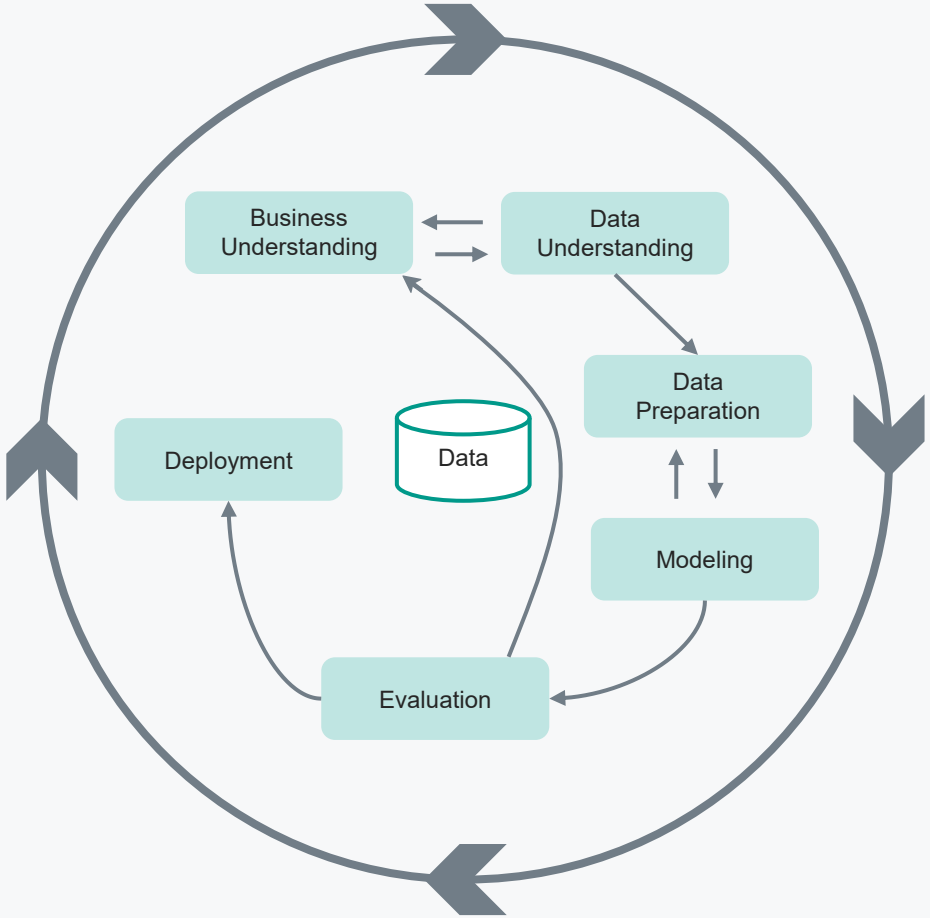
Begriffe

Cross Industry Standard Process for Data Mining (CRISP)

Wir werden diesen Prozess mehrfach durchlaufen. Deshalb nur eine kurze Behandlung. Details finden sich bspw. in der Quelle*.

„For business AI to succeed, you need to combine .. Machine Learning and Big Data with people who know the rules of the game in their business domain“ (Taddy, 2019, 311)

Wo geht Domänenwissen im CRISP ein?
Was müssen Sie beherrschen um Teil der Lösung zu werden / sein?



Quelle: Dorschel, 2015, S. 69

2 Regressionsmodelle

Explanatory Modeling

Goal: Explain relationship between predictors (explanatory variables) and target

- Familiar use of regression in data analysis
- **Model Goal: Fit the data well and understand the contribution of explanatory variables to the model**
- “goodness-of-fit”: **R^2**, residual analysis, **p-values**

Welche Beispiele haben wir in dieser Vorlesung schon behandelt?

Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 162.

Predictive Modeling

Goal: Predict target values in other data where we have predictor values, but not target values

- Classic data mining context
- **Model Goal: Optimize predictive accuracy**
- Train model on training data
- **Assess performance on validation (hold-out) data, e.g. cross validation**
- Explaining role of predictors is not primary purpose (but useful)

Wir werden hierzu den CRISP Kündigungsverhalten bei BGB 489 machen.

Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 162.

Fallstudie Toyota

Zuallererst müssen wir uns also fragen**: "Ist der Fokus die Erklärung oder die Vorhersage?“
Anders als im Lehrbuch von Shmueli et al (2019) formulieren wir unser Erkenntnisziel als

Erklärung der Gebrauchtwagenpreise von Toyota Corollas

Data: Prices of 1000 used Toyota Corollas, with their specification information

Für wen ist dies eine kommerziell relevante Fragestellung?

Welchen Mehrwert / Extra-Service bietet Ihnen ein Fachhändler beim Kauf eines Neuwagens?

Für welche Art Banken ist dies ebenfalls von Interesse?

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 162.
**Lübke/Krol (2022) Empirisch-quantitative Abschlußarbeiten, in: Boßow-Thies/Krol (Hrsg.) Quant Forschung im Masterarbeiten, S 503.

Fallstudie Toyota

Verfügbare Daten

Price in Euros
Age in months as of 8/04
KM (kilometers)
Fuel Type (diesel, petrol, CNG)
HP (horsepower)
Metallic color (1=yes, 0=no)
Automatic transmission (1=yes, 0=no)
CC (cylinder volume)
Doors
Quarterly_Tax (road tax)
Weight (in kg)

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 162.

Fallstudie Toyota

Verfügbare Daten

Price	Age	KM	Fuel_Type	HP	Metallic	Automatic	cc	Doors	Quarterly_Tax	Weight
13500	23	46986	Diesel	90	1	0	2000	3	210	1165
13750	23	72937	Diesel	90	1	0	2000	3	210	1165
13950	24	41711	Diesel	90	1	0	2000	3	210	1165
14950	26	48000	Diesel	90	0	0	2000	3	210	1165
13750	30	38500	Diesel	90	0	0	2000	3	210	1170
12950	32	61000	Diesel	90	0	0	2000	3	210	1170
16900	27	94612	Diesel	90	1	0	2000	3	210	1245
18600	30	75889	Diesel	90	1	0	2000	3	210	1245
21500	27	19700	Petrol	192	0	0	1800	3	100	1185
12950	23	71138	Diesel	69	0	0	1900	3	185	1105
20950	25	31461	Petrol	192	0	0	1800	3	100	1185

Fuel type is categorical (in R - a factor variable), must be transformed into binary variables. R's lm function does this automatically.

Diesel (1=yes, 0=no)

Petrol (1=yes, 0=no)

None needed* for "CNG" (if diesel and petrol are both 0, the car must be CNG)

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 162.

Fallstudie Toyota

Verfügbare Daten / Data Understanding

Price in Euros
Age in months as of 8/04
KM (kilometers)
Fuel Type (diesel, petrol, CNG)
HP (horsepower)
Metallic color (1=yes, 0=no)
Automatic transmission (1=yes, 0=no)
CC (cylinder volume)
Doors
Quarterly_Tax (road tax)
Weight (in kg)

Die meisten kennen Autos…

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 162.

Fallstudie Toyota

Data Preparation

Für diese Wiederholungsübung können wir auf die Modellvarianten verzichten.
Ebenso entfernen wir konstante und unvollständige Regressoren.

```
summary(Daten)

##        Id                                                    Model
##  Min.   :   1.0   TOYOTA Corolla 1.6 16V HATCHB LINEA TERRA 2/3-Doors:109
##  1st Qu.: 361.8   TOYOTA Corolla 1.3 16V HATCHB LINEA TERRA 2/3-Doors: 84
##  Median : 721.5   TOYOTA Corolla 1.6 16V LIFTB LINEA LUNA 4/5-Doors  : 80
##  Mean   : 721.6   TOYOTA Corolla 1.6 16V LIFTB LINEA TERRA 4/5-Doors : 71
##  3rd Qu.:1081.2   TOYOTA Corolla 1.4 16V VVT I HATCHB TERRA 2/3-Doors: 54
##  Max.   :1442.0   TOYOTA Corolla 1.6 16V SEDAN LINEA TERRA 4/5-Doors : 43
##                   (Other)                                           :995
```

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 162.

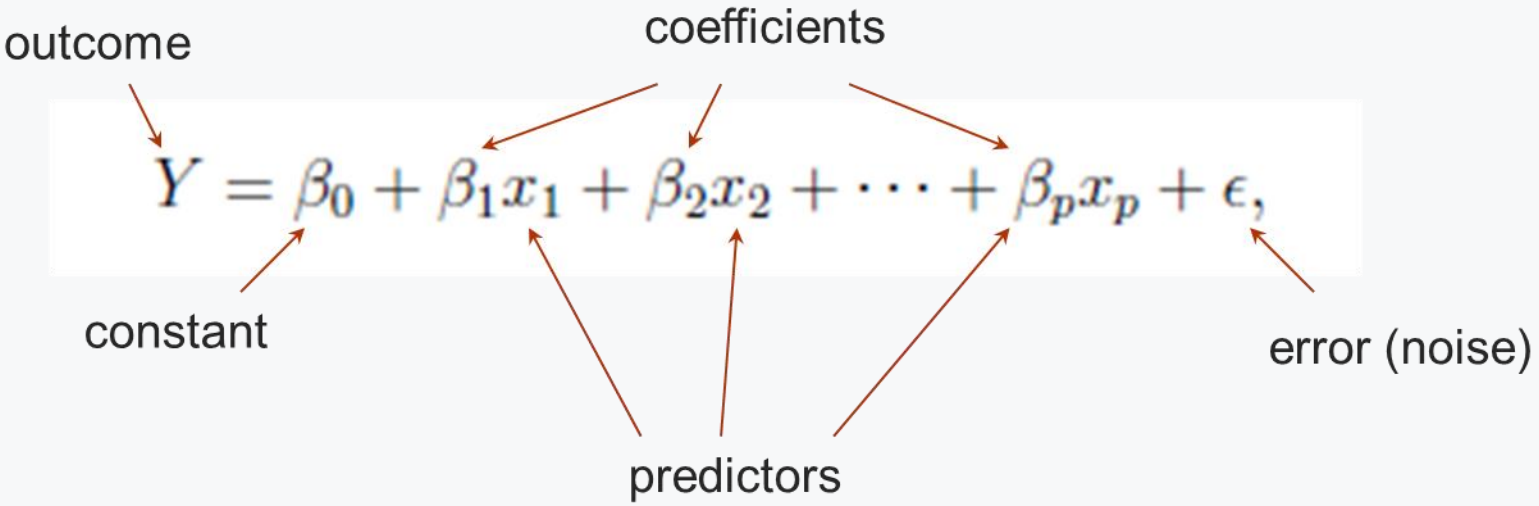
Fallstudie Toyota

Modeling

Wir unterstellen einen linearen Zusammenhang.

Was ist Y in unserem Fall?

Price		Age_08_04	
Min.	: 4350	Min.	: 1.00
1st Qu.	: 8450	1st Qu.	:44.00
Median	: 9900	Median	:61.00
Mean	:10731	Mean	:55.95
3rd Qu.	:11950	3rd Qu.	:70.00
Max.	:32500	Max.	:80.00



*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 162.

Fallstudie Toyota

Modeling*

Wir unterstellen einen linearen Zusammenhang und schätzen mit KQ:

```
reg = lm(Price~., data = insample)
summary(reg, digits = 4)

##
## Call:
## lm(formula = Price ~ ., data = insample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7063.5  -731.2   -15.9   698.4  5267.9
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.622e+02  1.909e+03  -0.452 0.651598
## Age_08_04    -1.194e+02  4.773e+00 -25.013  < 2e-16 ***
## KM           -1.924e-02  1.586e-03 -12.133  < 2e-16 ***
```

Es gibt eine nettere Darstellung auf der nächsten Folie.

*Quelle: Ausführung Case Toyota.R

Fallstudie Toyota

Modeling*

	Price		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-862.20	-4608.24 – 2883.83	0.652
Age_08_04	-119.39	-128.76 – -110.02	**<0.001**
KM	-0.02	-0.02 – -0.02	**<0.001**
Fuel_Type [Diesel]	674.24	-62.18 – 1410.67	0.073
Fuel_Type [Petrol]	1952.50	1187.61 – 2717.40	**<0.001**
. . .			
Parking_Assistant	-685.38	-1871.74 – 500.98	0.257
Tow_Bar	-50.38	-232.09 – 131.34	0.587
Observations	1000		
R^2 / R^2 adjusted	0.908 / 0.904		

Ceteris paribus: Wieviel Wertverlust erleide ich mit jedem einzelnen Jahr?

Könnte der Achsenabschnitt auch Null sein?

Wie lauten die Konfidenzintervalle für den ß-Parametervektor

*Quelle: Ausführung Case Toyota.R

Fallstudie Toyota

Modeling*

	Price		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-862.20	-4608.24 – 2883.83	0.652
Age_08_04	-119.39	-128.76 – -110.02	**<0.001**
KM	-0.02	-0.02 – -0.02	**<0.001**
Fuel_Type [Diesel]	674.24	-62.18 – 1410.67	0.073
Fuel_Type [Petrol]	1952.50	1187.61 – 2717.40	**<0.001**
...			
Parking_Assistant	-685.38	-1871.74 – 500.98	0.257
Tow_Bar	-50.38	-232.09 – 131.34	0.587
Observations	1000		
R^2 / R^2 adjusted	0.908 / 0.904		

Wieviel Prozent der Varianz des Preises erklärt das Modell?

Was müssen wir noch tun?

*Quelle: Ausführung Case Toyota.R

Wiederholung – wahres und geschätztes Modell

- Es ist sehr wichtig, zwischen Residuen und Störtermen zu unterscheiden!
- Wir illustrieren dies an der **einfachen** Regression, die wir - wie gehabt - mit KQ/OLS schätzen:

(w) Das **wahre** Modell sei: $y=a+bx+e$

(g) Das **geschätzte** Modell sei: **(a0)** $y=\alpha+\beta x+u$

- Der Störterm ist die Zufallsvariable e und das Residuum ist u
- Wir haben n Beobachtungen und indizieren diese mit i = 1, …, n.

Welche der beiden Gleichungen generiert die Daten der Grundgesamtheit (population)?

Welche liegt hinter/steuert die Wirklichkeit?

Welche bedient sich/basiert auf der Stichprobe (sample)?

Wiederholung – Satz von Gauss Markov

Es gilt der folgende Satz von Gauss und Markov (Verbeek, 2012, 15-17)

Wenn für die N Störterme e_i die folgenden Gauss-Markov Bedingungen gelten:

- (a1) $E[e_i]$=0, i=1,…,n (Im Mittel heben sich die Störungen weg, ignorierte Faktoren heben sich ggs auf)
- (a2) e und x sind unabhängig (Exogenität von x: Nur y durch Modell erklärt, x von ausserhalb)
- (a3) $Var[e_i]=\sigma^2$ (Homoskedastizität)
- (a4) $Cov[e_i, e_j]$=0, $i \neq j$ (Keine Autokorrelation im Störterm)
- (a5) Im Falle einer **multiplen** lin. Regression: Keine überflüssigen x!

Dann sind die **Schätzer** α, β die

- **B**esten (Sie haben die kleinste Varianz …)
- **L**inearen (unter allen linearen … Schätzern)
- **U**nverzerrten Schätzer (Erwartungstreue) (E[ß]=b und E[α]=a, d.h. wir schätzen im Mittel richtig)

für die wahren Parameter a, b. Das nennt man auch die **BLUE Eigenschaft der Kleinste Quadrate Schätzer** α, β.

Wiederholung – Theoretische Ergänzung

Zudem gilt bei „big data“ (=large samples, Details bei Verbeek, 2012, 34 & 138-141):

Wenn für die N Störterme e_i die folgenden Bedingungen gelten:

- (a1) $E[e_i]$=0, i=1,…,n (Im Mittel heben sich Störungen weg, ignorierte Faktoren heben sich ggs auf)
- (a2) e und x sind unabhängig (Exogenität von x: Nur y durch Modell erklärt, x von außerhalb)
- (a3) $Var[e_i]$ variiert, d.h. Heteroskedastizität liegt vor
- (a4) Autokorrelation im Störterm geht nach „wenigen“ Lags auf Null
- (a5) Im Falle einer **multiplen** Regression: Keine lineare Abhängigkeit in x

Wird für die Störterme bei beiden Sätzen eine bestimmte Verteilung unterstellt?

Dann sind die **Schätzer** α, β

- **Konsistente** Schätzer (d.h. ß konvergiert in WS gg b, analog α gg a; **formal** lim $_{n->inf}$ P(|b-ß$_n$|>d) = 0 für d > 0 (beliebig klein!)
- Die t-Werte können auf Basis der Heteroscedasticity & Autocorrelation Consistent Standard Errors (**HAC SE**) als **normalverteilt** interpretiert werden

Hinweis: Best gilt nicht mehr, weil es bessere Modellansätze gibt, z B Cochrane … GARCH ….

Wiederholung - Zusammenfassung

Voraussetzung	Voraussetzungs-verletzung	Konsequenzen	Was bleibt?	Was könnte man tun?
Linearität der Parameter	Nichtlinearität	Verzerrung der Schätzwerte		Andere funktionale Form
Vollständigkeit des Modells	Unvollständigkeit	Verzerrung der Schätzwerte		Mehr Regressoren
Homoskedastizität der Störterme	Heteroskedastizität	▪ Ineffizienz: Es gibt bessere Schätzfunktion! ▪ SE falsch!	LU // Konsistenz	▪ HAC SE nutzen ▪ Heteroskedastizität explizit modellieren z B GARCH…
Unabhängigkeit der Störgrößen	Autokorrelation	▪ Ineffizienz: Es gibt bessere Schätzfunktion! ▪ SE falsch!	LU // Konsistenz	▪ HAC SE nutzen ▪ Autokorrelation explizit modellieren z B Cochrane…
Keine lineare Abhängigkeit zwischen den unabhängigen Variablen	Multikollinearität	Präzision der Schätzwerte geringer als nötig, d.h. SE korrekt aber zu hoch	BLU	Reduktion der Anzahl der Regressoren gemäß ökon. Vorverständnis oder Zielkriterium
Normalverteilung der Störgrößen	nicht normalverteilt	Ungültigkeit der Signifikanztests, wenn Stichprobe klein ist. Sonst Heilung durch ZGWS.	BLU	Stipo vergrößern. Wenn Stipo groß, kann man ML nutzen und dann mit anders verteiltem Störterm schätzen (s.o. z B GARCH).

Wiederholung - Diagnostik

Modeling

Diagnostik

Die Überprüfung der Annahmen der linearen Regression geschieht durch die Diagnostik. Mit Blick auf den stochastischen Teil der Gleichung (w) können wir die geforderten Eigenschaften des Störterms nur **indirekt** über die Analyse der Residuen überprüfen. Man nennt die Diagnostik deswegen auch Residualanalyse.

In einer Regression von Längsschnittdaten y auf Längsschnittdaten x kann dabei die **serielle/zeitliche/Autokorrelation** der Residuen ein Problem darstellen.

Wiederholung - Diagnostik

Modeling

Diagnostik

In einer Regression von Querschnittdaten y auf x kann dabei die **räumliche Korrelation** der Residuen ein Problem darstellen.

So könnte die Analyse von Bankdaten über mehrere Filialen hinweg höhere Residuen bei einer Filiale mit schlampiger Datenführung aufweisen.

Aus einem anderen Grund könnte zuweilen auch eine Korrelation bei bestimmten Residuenpaaren bestehen. In Schröder (2002, 105) wird deswegen etwa die Korrelation benachbarter Residuen geplottet. Natürlich könnte man dies auch komfortabel mit einem Korrelogramm (in R ACF) leisten.

JEDOCH sollten im Idealfall die Elemente einer Querschnittsstichprobe unabhängig erhoben worden sein. D.h. wenn wir die Reihenfolge im Datensatz ändern, sollte der Gehalt derselbe bleiben. Davon gehen wir im Weiteren aus. Sollte dennoch Autokorrelation vorliegen, gilt:

Um auf der sicheren Seite zu sein, nutzen wir stets HAC SE!

Dies können wir bei großen Stipo getrost tun.

Fallstudie Toyota

Modeling*

Diagnostik

Welche Annahme der linearen Regression testen wir hier?

Wie interpretieren Sie das Ergebnis?

```
##Fehlspezifikation-----------------------------------
resettest(reg)#H0 No spec error

##
##  RESET test
##
## data:  reg
## RESET = 65.109, df1 = 2, df2 = 955, p-value < 2.2e-16
```

*Quelle: Ausführung Case Toyota.R

Fallstudie Toyota

Modeling*

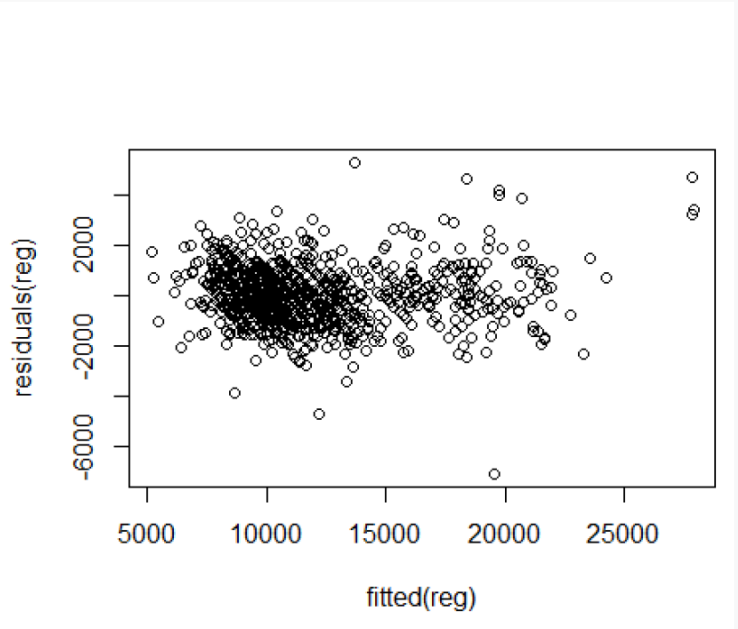
Diagnostik

Welche Annahme der linearen Regression testen wir hier?

Wie interpretieren Sie das Ergebnis?

Variiert die Streubreite der Residuen systematisch?

```
plot(fitted(reg),residuals(reg)) #, type='l')
```



*Quelle: Ausführung Case Toyota.R

Fallstudie Toyota

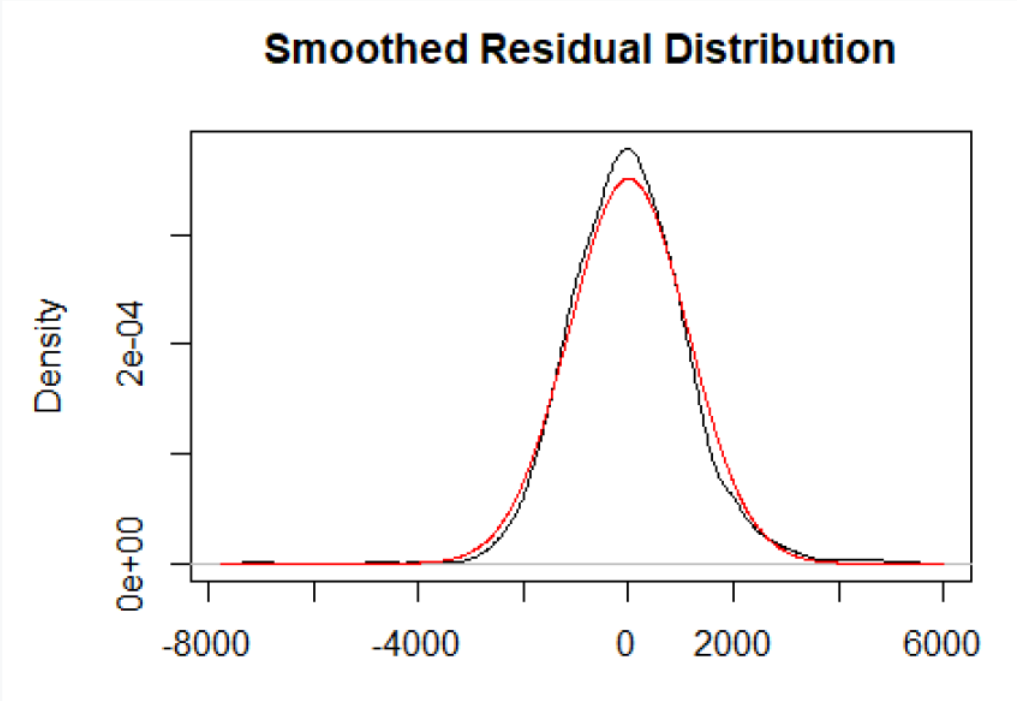
Modeling*

Diagnostik

Was testen wir hier?

Wie interpretieren Sie das Ergebnis?

Ist dies ein Problem für BLUE?



*Quelle: Ausführung Case Toyota.R

Fallstudie Toyota

Modeling*

Ergebnisdarstellung

Welche der beiden Tabellen nutzen Sie für die Präsi ggü. Ihrer Chefin?

I.a.W.: Welchen der beiden Sätze nutzen Sie?

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -8.622e+02  1.909e+03  -0.452 0.651598
## Age_08_04         -1.194e+02  4.773e+00 -25.013  < 2e-16 ***
## KM                -1.924e-02  1.586e-03 -12.133  < 2e-16 ***
## Fuel_TypeDiesel    6.742e+02  3.753e+02   1.797 0.072691 .
## Fuel_TypePetrol    1.953e+03  3.898e+02   5.009 6.50e-07 ***

##                     Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)      -8.6220e+02  5.1066e+03  -0.1688 0.8659582
## Age_08_04        -1.1939e+02  5.8701e+00 -20.3387 < 2.2e-16 ***
## KM               -1.9238e-02  2.1295e-03  -9.0339 < 2.2e-16 ***
## Fuel_TypeDiesel   6.7424e+02  4.6166e+02   1.4605 0.1444828
## Fuel_TypePetrol   1.9525e+03  6.4231e+02   3.0398 0.0024314 **
```

*Quelle: Ausführung Case Toyota.R

Fallstudie Toyota

Modeling*

Ergebnisdarstellung

Ist das schon Ihr Endmodell?

Was benennen Sie als Achsenabschnitt?

```
##                      Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)       -8.6220e+02  5.1066e+03  -0.1688 0.8659582
## Age_08_04         -1.1939e+02  5.8701e+00 -20.3387 < 2.2e-16 ***
## KM                -1.9238e-02  2.1295e-03  -9.0339 < 2.2e-16 ***
## Fuel_TypeDiesel    6.7424e+02  4.6166e+02   1.4605 0.1444828
## Fuel_TypePetrol    1.9525e+03  6.4231e+02   3.0398 0.0024314 **
```

```
vif(reg)[,1]#could reduce X dim ...

##         Age_08_04                KM         Fuel_Type                HP
##          4.700974          2.333795          9.783617          2.427803
##         Met_Color             Color         Automatic                CC
##          1.352826          2.196029          1.176867          1.216384
##             Doors             Gears     Quarterly_Tax            Weight
##          1.328536          1.394619          5.882953          4.063355
##     Mfr_Guarantee   BOVAG_Guarantee Guarantee_Period               ABS
##          1.227178          1.462688          1.696901          1.510669
##         Airbag_1          Airbag_2             Airco    Automatic_airco
##          1.781138          3.065307          1.931547          1.915450
##     Boardcomputer        CD_Player      Central_Lock    Powered_Windows
##          2.480866          1.566257          4.458877          4.537778
##    Power_Steering             Radio         Mistlamps       Sport_Model
##          1.856590         47.912306          2.336991          1.522498
##  Backseat_Divider       Metallic_Rim    Radio_cassette Parking_Assistant
##          3.772178          1.407131         47.736829          1.078769
##          Tow_Bar
##          1.175081
```

*Quelle: Ausführung Case Toyota.R

Fallstudie Toyota

Modellverbesserung: Reducing the Number of Predictors

Nicht nur wg. Ockham‘s Rasiermesser sollten wir versuchen, mit weniger Regessoren auszukommen:

- „It may be expensive … to collect a full complement of predictors for future predictions.
- We may be able to measure fewer predictors more accurately.
- The more predictors, the higher the chance of missing values.
- Regression coefficients are more stable for parsimonious models.
- It can be shown that using predictors that are uncorrelated with the outcome variable increases the variance of predictions.“

Formulieren Sie möglichst pointiert obige Aussagen in Ihren eigenen Worten:
z. B. Vermeide unnötige Kosten des Modellbetriebs durch …

Wir wollen die Suche nach dem Wesentlichen nun in einem **Laborexperiment** durchführen.

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 169.

Ansatz

Wir wollen Daten gemäß eines **Daten-Generierenden-Prozesses (DGP)** erzeugen Y(X) und prüfen, ob unsere Werkzeuge der Ökonometrie diesen DGP aufspüren können.
Es handelt sich um ein Problem des **überwachten Lernens**, da eine Funktion Y von X gesucht wird. Würde man sich nur für X interessieren, läge unüberwachtes Lernen vor.
Diese **Vorübung** ist wichtig und folgt dem Leitgedanken „Business data science can only be learned by doing" (Taddy, 2019, xi). Danach wenden wir die Werkzeuge auf reale Datensätze an (z B Mieten in Düsseldorf).
Big Data lässt sich salopp als 'viele Zeilen und viele Spalten' charakterisieren.
Feature Selection meint die Auswahl der relevanten 'Spalten' / Features / Faktoren / Determinanten / Covariates / Regressoren / Kennwerte / XXX

Welche Termini kennen Sie noch für XXX?

Wie lautete der DGP bei der linearen Regression? Gleichung (w) oder (g)?

Regressionsmodelle

Praxisproblem anhand des Beispiels Feature Selection

Wir nutzen das Beispiel 25.1 aus Shalev-Shwartz & Ben-David (2014). Das Coding heisst **GML_06_02 *feature selection*.R**.
Auch anhand dieses Beispiels wiederholen wir die aus dem Bachelor bekannten Methoden (Module Wiss. Methodik / Ökonometrie o.a.) und führen neue Werkzeuge ein.

```
SEED = 2021

n=50
set.seed(SEED)
x1 = runif(n,-1,1)
z = runif(n,-0.1,0.1)
eps = rnorm(n,0,0.0005)
y=x1^2+eps
x2=y+z
x3=rnorm(n,0,0.0000005)

df1 = data.frame(cbind(y,x1,x2,x3))
```

Welcher Faktor bestimmt y wesentlich und welcher Faktor ist irrelevant?

Nachfolgend werden verschiedene Techniken vorgestellt.

Wie ist der wahre Zusammenhang y(x)?

Beispiel Feature Selection

Eine erste multiple lineare Regression ergibt:

```
(Intercept)  3.986e-02  1.191e-02   3.346  0.00164 **
x1          -2.130e-02  1.217e-02  -1.751  0.08669 .
x2           9.224e-01  2.267e-02  40.684  < 2e-16 ***
x3           2.193e+04  1.299e+04   1.688  0.09813 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05058 on 46 degrees of freedom
Multiple R-squared:  0.9758, Adjusted R-squared:  0.9742
F-statistic: 617.1 on 3 and 46 DF,  p-value: < 2.2e-16
```

Interpretieren Sie!

Was ist noch zu tun?

Beispiel Feature Selection

Die Diagnostik ergibt:

```
resettest(reg)#H0 No spec error

##
##  RESET test
##
## data:  reg
## RESET = 5.4236, df1 = 2, df2 = 44, p-value = 0.007844
```

Interpretieren Sie!

Beispiel Feature Selection

Die Diagnostik ergibt:

```
bptest(reg)#H0: Homoscedasticity!

##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 0.78662, df = 3, p-value = 0.8527
```

Interpretieren Sie!

Beispiel Feature Selection

VIF ergibt:

```
vif(reg)

##       x1       x2       x3
## 1.096692 1.121438 1.065499
```

Interpretieren Sie!

```
(Intercept)  3.986e-02  1.191e-02   3.346  0.00164 **
x1          -2.130e-02  1.217e-02  -1.751  0.08669 .
x2           9.224e-01  2.267e-02  40.684  < 2e-16 ***
x3           2.193e+04  1.299e+04   1.688  0.09813 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05058 on 46 degrees of freedom
Multiple R-squared:  0.9758, Adjusted R-squared:  0.9742
F-statistic: 617.1 on 3 and 46 DF,  p-value: < 2.2e-16
```

Zwischenstand: Weder vif noch p-Werte weisen zur Wahrheit.

Vielleicht sind mehr Daten die Lösung???

Bootstrap

Praxisproblem

Im Beispiel haben wir nur eine Stipo mit einem Umfang von N = 50 Elementen.

In der Praxis hat man zuweilen solch kleine Stipo. Dennoch möchte man ein Gefühl für die Unsicherheit der Schätzergebnisse gewinnen.

Hier hilft Bootstrapping bei dem man sich wie Münchhausen an den eigenen Haaren aus dem Sumpf des Problems zieht. In der englischen Version der Geschichte macht es der Lügenbaron mit Hilfe seiner Schnürsenkel an den Stiefeln. Deshalb: Bootstrapping.
Den wiederholten Vorgang des Resamplings (hier n_boot = 2000) mit Zurücklegen nennt man **Bootstrapping** – die unbekannte Verteilung wird anhand der empirischen Verteilung geschätzt. Hier kennen wir zwar die Verteilung, aber in fortgeschrittenen Modellen nicht mehr.

Bootstrap

Praxisproblem

Im Beispiel haben wir nur eine originale Stipo mit einem Umfang von N = 50 Elementen.
Wie gross ist jede der n_boot = 2000 künstlich generierten Bootstrap-Stichproben zu wählen?
Schlittgen (2013, 58) plädiert für den „gleichen Umfang“, d.h. im „Umfang der Ausgangsstichprobe“ (ibid., 114).

Mehr zum Thema findet sich in Taddy (2019, S. 18-29) und Maddala & Lahiri, (2009, Kap. 16).

Wir verwenden die “direkte Methode” (ibid., 607) bei der 2000 mal 50 Paare y, x gezogen werden und nicht nur auf der rechten Seite der Schätzgleichung neue Störtermrealisationen.

Im Bsp. interessiert uns die Verteilung der Schätzer für die Koeffizienten:

Bootstrap

Beispiel Feature Selection

Wir setzen n_boot = 2000 und erhalten dank R:

```
# Bootstrap regression coefficients
set.seed(SEED)
betas boot = matrix(0,n boot,4)
for (b in 1:n_boot){
  boot.data = df1[sample(nrow(df1), n, replace = TRUE),]
  betas_boot[b,] = coef(lm(y~., data = boot.data))
}
#ESTIMATES and their SE
cbind(matrix((apply(betas_boot, MARGIN = 2,
mean)),4,1),matrix(apply(betas_boot, MARGIN = 2, sd),4,1))

##                [,1]         [,2]
## [1,]  3.983631e-02 1.038659e-02
## [2,] -2.136094e-02 1.294342e-02
## [3,]  9.223407e-01 2.043758e-02
## [4,]  2.182411e+04 1.527289e+04
```

Interpretieren Sie! Wofür stehen die Spalten?

Womit könnte man vergleichen? Und kommen wir zu anderen Schlüssen?

Bootstrap

Feature Selection

Wir vergleichen:

Name	Estimate	Standard Error		
(Intercept)	3.986e-02	1.191e-02	3.346	0.00164 **
x1	-2.130e-02	1.217e-02	-1.751	0.08669 .
x2	9.224e-01	2.267e-02	40.684	< 2e-16 ***
x3	2.193e+04	1.299e+04	1.688	0.09813 .

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05058 on 46 degrees of freedom
Multiple R-squared:  0.9758, Adjusted R-squared:  0.9742
F-statistic: 617.1 on 3 and 46 DF,  p-value: < 2.2e-16
```

Konkret wurde n_boot mal eine neue Stipo mit n=50 Elementen gezogen und darauf die multiple lineare Regression geschätzt.

Name	Estimate	Standard Error
(Intercept)	3.983631e-02	1.038659e-02
x1	2.136094e-02	1.294342e-02
x2	9.223407e-01	2.043758e-02
x3	2.182411e+04	1.527289e+04

Es ergeben sich n_boot (hier 2000) neue Intercepts und Koeffizienten für x1 bis x3. Die Mittelwerte stehen rechts und sind nahezu identisch zur ersten Regression auf einer Stipo mit n=50. Unsere Ergebnisse scheinen robust.

Bootstrap

Beispiel Feature Selection

In unserem Laborexperiment können wir „big data“ selber machen. Eine zweite multiple lineare Regression ergibt:

Wir setzen big_n = n * n_boot = 100000

Interpretieren Sie!

Ist irgendetwas besser geworden?

Oder gar schlimmer?

```
set.seed(SEED)
big_n=100000
big_x1 = runif(big_n,-1,1)
big_z = runif(big_n,-0.1,0.1)
big_eps = rnorm(big_n,0,0.0005)
big_y=big_x1^2+big_eps
big_x2=big_y+big_z
big_x3=rnorm(big_n,0,0.0000005)

big_lm1 = lm(big_y~big_x1+big_x2+big_x3)
```

```
(Intercept)  1.187e-02  2.661e-04   44.582   <2e-16 ***
big_x1       2.239e-04  3.100e-04    0.722   0.4702
big_x2       9.637e-01  5.889e-04 1636.363   <2e-16 ***
big_x3      -8.316e+02  3.595e+02   -2.313   0.0207 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bootstrap

Beispiel Feature Selection

Vergleich beider Regressionen im Detail:

```
(Intercept)  3.986e-02  1.191e-02   3.346  0.00164 **
x1          -2.130e-02  1.217e-02  -1.751  0.08669 .
x2           9.224e-01  2.267e-02  40.684  < 2e-16 ***
x3           2.193e+04  1.299e+04   1.688  0.09813 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Intercept)  1.187e-02  2.661e-04   44.582   <2e-16 ***
big_x1       2.239e-04  3.100e-04    0.722   0.4702
big_x2       9.637e-01  5.889e-04 1636.363   <2e-16 ***
big_x3      -8.316e+02  3.595e+02   -2.313   0.0207 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hat uns Big Data bei der Feature Selection weiter gebracht?
Erkennen wir den wahren DGP nun besser?
Ist der Selektionsalgorithmus „p-Wert kleiner alpha“ hilfreich?

Bootstrapping Einordnung

Beispiel Feature Selection

Nun vergleichen wir die Big Data Regression mit dem Bootstrapping Ansatz:

```
(Intercept)  1.187e-02  2.661e-04   44.582   <2e-16 ***
big_x1       2.239e-04  3.100e-04    0.722   0.4702
big_x2       9.637e-01  5.889e-04 1636.363   <2e-16 ***
big_x3      -8.316e+02  3.595e+02   -2.313   0.0207 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bootstrapping:

```
 Intercept  3.985600e-02 1.541991e-01
        x1 -2.130169e-02 4.686757e-01
        x2  9.223727e-01 1.172105e+00
        x3  2.193349e+04 4.146774e+05
     sigma  5.058215e-02
```

Es wird deutlich, dass beide Ansätze verschieden sind, auch wenn dieselbe Anzahl von Datensätzen eine Rolle spielte:Beim Big Data Ansatz haben wir viele neue X‘e erzeugt und daraus neue Y berechnet. Hingegen haben wir beim Bootstrap die Paare Y, X so belassen und nur Paare für neue Stipo gezogen (Taddy, 2019, 25)

Schrittweise Regression

Feature Selection

Der Selektionsalgorithmus „p-Wert kleiner alpha darf drin bleiben“ hat uns nicht zur Wahrheit gebracht. Weitere Gründe sprechen gegen den Selektionsalgorithmus „p-Wert kleiner alpha“ (Taddy, 2019, 74):

- „When you have multicollinearity … the p-values for all of these variables will be large … even if any one of the variables provides a useful signal on the response“

There is another case:

- „The p-values are based on an overfit model. … If you use these p-values, you will be building a set of candidate models on the foundation of a terrible regression fit“

Schrittweise Regression

Feature Selection

The „general approach – looking at a full model fit and then cutting it down to size - is sometimes called backward **stepwise regression**.“ In R:

```
backward_lm1 = regsubsets(y~x1+x2+x3, data = df1, method=c("backward"))
summary(backward_lm1)
```

```
Selection Algorithm: backward
          x1  x2  x3
1  ( 1 ) " " "*" " "
2  ( 1 ) "*" "*" " "
```

Wegen der Probleme auf der vorigen Folie gilt: „A better solution is … building from simplicity to complexity“

```
forward_lm1 = regsubsets(y~x1+x2+x3, data = df1, method=c("forward"))
summary(forward_lm1)
```

```
Selection Algorithm: forward
          x1  x2  x3
1  ( 1 ) " " "*" " "
2  ( 1 ) "*" "*" " "
```

Hier helfen beide Richtungen. Jedoch stoßen beide Verfahren an Grenzen. Bei 3 Faktoren haben wir 2 hoch 3 Modellvarianten. Generell 2 hoch k bei k Regressoren. Bei 2 hoch 20 also schon mehr als eine Million.

Schrittweise Regression

Feature Selection

Der Grund warum beides Mal bei Auswahl nur eines Faktors x2 gewählt wird, liegt daran, dass die funktionale Form als linear in der Schätzgleichung vorgegeben wurde.

```
Selection Algorithm: backward
         x1  x2  x3
1  ( 1 ) " " "*" " "
2  ( 1 ) "*" "*" " "
```

Erst wenn zwei Faktoren erlaubt sind, wird x1 interessant.

→ Weitere Details zur stepwise regression via help(regsubsets) in R.

Regularisierung

Feature Selection

In Zeiten von Big Data mit vielen Spalten (Faktoren) sind algorithmische Hilfsmittel hilfreich, weil eine visuelle Analyse u. U. zu komplex werden kann.

Der Schätzalgorithmus war bislang die Minimierung der Summe Residuenquadrate. Fügt man dieser Zielfunktion einen passenden Strafterm hinzu, so kann der Schätzalgorithmus incentiviert werden, möglichst wenige Faktoren zu berücksichtigen. Formal unter Zuhilfenahme einer Präsentation* von Taddy (2019):

Min (Summe Residuenquadrate + $\lambda \sum_j c(\beta_j)$

Der Strafterm besteht aus einem Skalar λ mal die Summe über eine Cost Function c() mit den Modellkoeffizienten als Argumenten.

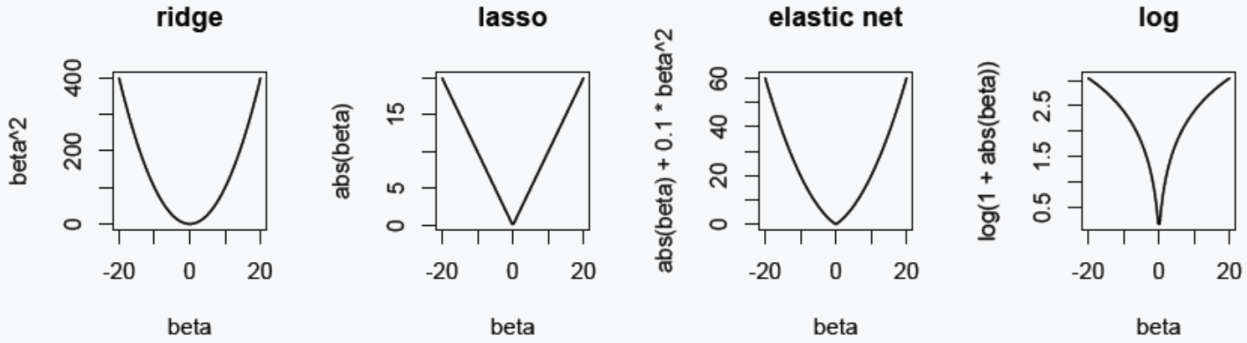
*Quelle: https://github.com/TaddyLab/MBAcourse/tree/master/lectures/03Models.pdf, Zugriff 19. Nov. 2021

Regularisierung

Feature Selection

Mögliche Cost Functions sind unter Zuhilfenahme einer Präsentation* von Taddy (2019):

$\lambda > 0$ is the penalty weight, c is a cost (penalty) function.
$c(\beta)$ will be lowest at $\beta = 0$ and we pay more for $|\beta| > 0$.



Options: ridge β^2, lasso $|\beta|$, elastic net $\alpha\beta^2 + |\beta|$, $\log(1 + |\beta|)$.

„As you will see, the lasso is a fantastic default“ (Taddy, 2019, 79).

*Quelle: https://github.com/TaddyLab/MBAcourse/tree/master/lectures/03Models.pdf, Zugriff 19. Nov. 2021
LASSO = Least Absolute Shrinkage And Selection Operator

Regularisierung

Feature Selection

„As you will see, the lasso is a fantastic default“ (Taddy, 2019, 79).

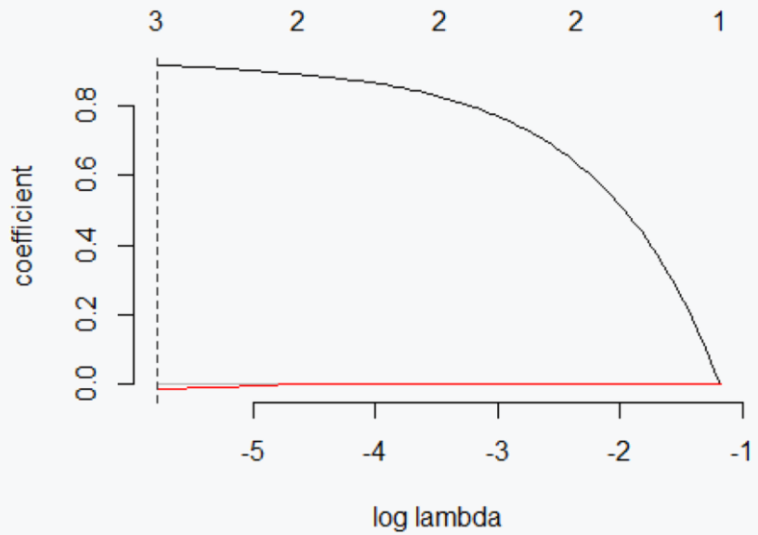
Note that „size (of the covariates) now matters. Snce the ß values are all penalized by the same λ, you need to make sure they live on comparable scales“ (ibid., 85).

```
lasso_lm1 <- gamlr(cbind(x1,x2,x3), y, standardize = T)
```

Von rechts nach links werden die λ kleiner

Entsprechend werden die ß grösser

Default sind 100 Schritte



*Quelle: https://github.com/TaddyLab/MBAcourse/tree/master/lectures/03Models.pdf, Zugriff 19. Nov. 2021

LASSO = Least Absolute Shrinkage And Selection Operator

Regularisierung

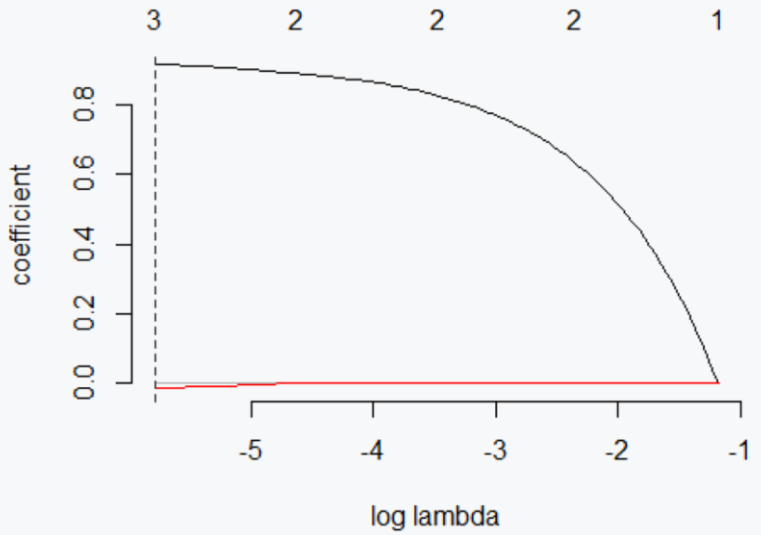
Feature Selection

Von rechts nach links werden die λ kleiner

Entsprechend werden die ß grösser

Welche ß genau?

```
lasso_lm1$beta[,10] #for high lambda

##        x1        x2        x3
## 0.0000000 0.3146378 0.0000000

lasso_lm1$beta[,50]

##        x1        x2        x3
## 0.0000000 0.8256681 0.0000000

lasso_lm1$beta[,100] #for lowest level of lambda

##          x1          x2          x3
## -0.01208422  0.91677043  0.00000000
```



Interpretieren Sie! Auf welchen Faktor kann man im Weiteren verzichten? Und was könnte man noch ändern?

Regularisierung

Feature Selection mit LASSO

Man würde nun ein lineares Regressionsmodell nur auf x1 und x2 schätzen, um Schätzer mit BLUE zu erhalten.
Die LASSO Schätzer sind verzerrt!

```
lasso_lm1$beta[,10] #for high lambda

##        x1        x2        x3
## 0.0000000 0.3146378 0.0000000

lasso_lm1$beta[,50]

##        x1        x2        x3
## 0.0000000 0.8256681 0.0000000

lasso_lm1$beta[,100] #for lowest level of lambda

##          x1          x2          x3
## -0.01208422  0.91677043  0.00000000
```

Regressionsbäume

Feature Selection mit Bäumen

Nachfolgend orientiere ich mich an Schlittgen, R. (2013, Kap. 6.5).

Bislang haben wir versucht eine Gerade durch die Punktewolke zu legen (Hyperebene) – egal ob mit oder ohne Strafterm.

Nun wollen wir es mit einer **stückweise konstanten Funktion** versuchen. Angenommen wir hätten nur y(x), also eine Funktion mit nur einem Argument. Dann könnten ein Intervall / Gebiet x<c definieren und für alle diese x denselben Vorhersagewert des mittleren y nehmen. Für die x>c nehmen wir als Vorhersage das druchschnittliche Niveau von y für eben diese Werte. D.h. wir haben nur zwei Zahlen für die Prognose. Das mittlere y für die x<c und das mittlere y für die x >c.

Ziel ist - wie anfangs - die Residuenquadrate zu minimieren. Ein Strafterm – wie bei LASSO - spielt erstmal keine Rolle.

Wir nutzen den Classification and Regression Tree, kurz CART.

Schlittgen (2013, 180) weist darauf hin, dass Bäume konsistente Schätzfunktionen sind.

Regressionsbäume

Feature Selection mit Bäumen

Ziel ist - wie anfangs - die Residuenquadrate zu minimieren. Als Bezugsgröße messen wir den **Mean Squared Error (MSE)** für den Fall, dass wir stets das **mittlere y** ohne Fallunterscheidung vorhersagen. Diese naive Prognose stellt unsere **Benchmark** dar**.**

Der MSE beträgt dann:

```
sum((y-mean(y))^2)/length(y)

## [1] 0.09708513
```

Nun nutzen wir die Funktion rpart und lassen sie verschiedene c für die x1 bis x3 ausprobieren. Wir legen fest, dass zunächst nur eines der drei xi genommen werden darf!

```
cart1 = rpart(y~x1+x2+x3, data = df1, maxdepth = 1)
summary(cart1)
```

Die Funktion rpart probiert nun alle drei xi durch und berichtet uns jenes, welches den MSE am meisten senkt.

Regressionsbäume

Feature Selection mit Bäumen

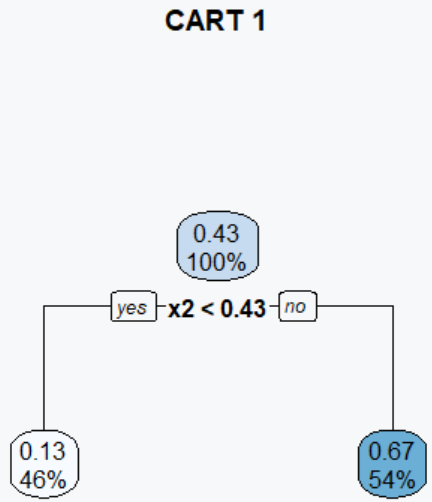
Das Ergebnis sieht so aus:

Wir können die dazu passende Vorhersagefunktion bauen:

```
cart1Forecast <- function(x){
  if (x < 0.4278535) (0.1339119)
  else (0.6749837)
}
```

QS:

```
#cross check first forecast
yx2_pairs = Matrix(c(y,x2),50,2)

m2 = 0
j=0
for (i in (1:50)){
  if(yx2_pairs[i,2]<0.4278535){
    m2 = m2 + yx2_pairs[i,1]
    j=j+1

      }
}
m2=m2/j
m2 #hence, average y is forecasted
```



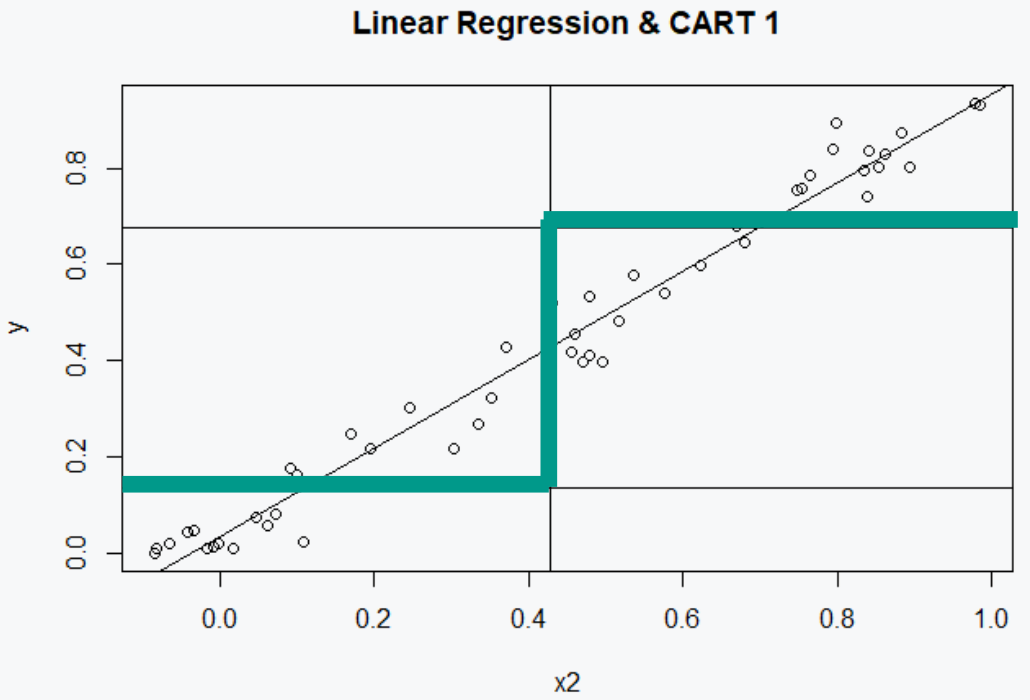
Interpretieren Sie! Formulieren Sie die Vorhersageregel in eigenen Worten?
Wie sieht nun die geschätzte Funktion aus?

Regressionsbäume

Feature Selection mit Bäumen

Wir können die dazu passende Vorhersagefunktion zeichnen:

```
cart1Forecast <- function(x){
  if (x < 0.4278535) (0.1339119)
  else (0.6749837)
}
```



Interpretieren Sie! Wie genau läuft die stückweise konstante Funktion?

Regressionsbäume

Feature Selection mit Bäumen

Wir wollen berechnen, wie sich der MSE verändert hat:

```
y_hat_cart1 = apply(as.matrix(x2), MARGIN = 1, FUN = cart1Forecast)
sum((y-y_hat_cart1)^2)/length(y)#compare this to MSE in cart summary

## [1] 0.02436387

23/50*0.01691391 + 27/50*0.03071012

## [1] 0.02436386
```

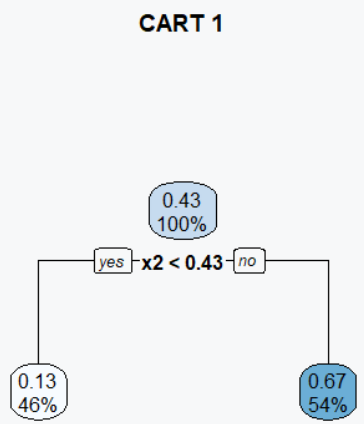
Wie hoch ist der MSE nun?

Regressionsbäume

Feature Selection mit Bäumen

Wir wollen den Output von R verstehen:

```
## Node number 1: 50 observations,    complexity param=0.7490464
##   mean=0.4260907, MSE=0.09708513
##   left son=2 (23 obs) right son=3 (27 obs)
##   Primary splits:
##       x2 < 0.4278535     to the left,  improve=0.7490464, (0 missing)


## Node number 2: 23 observations
##   mean=0.1339119, MSE=0.01691391
##
## Node number 3: 27 observations
##   mean=0.6749837, MSE=0.03071012
```



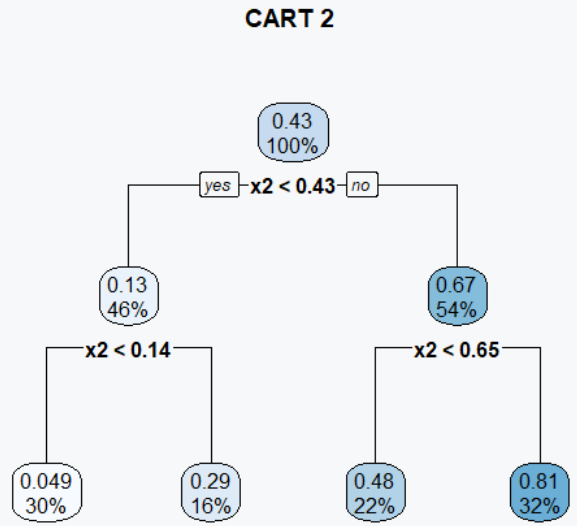
Interpretieren Sie!

Regressionsbäume

Feature Selection

Wir wollen nun einen weiteren Knoten mit davon ausgehenden Ästen zuzulassen. Damit bestünde die Möglichkeit, ein weiteres xi hinzuzunehmen. Es könnte aber auch zu einer weiteren Unterteilung der Intervalle von x2 kommen.

```
round(cor(lm1$fitted.values,y)^2,4)

## [1] 0.9758

round(cor(rpart.predict(cart2,as.data.frame(cbind(x1,x2,x3)),rules=F),y)^2,4)
#improve

## [1] 0.9525
```



Wofür hat sich CART entschieden?

Wie ist die Modellgüte?

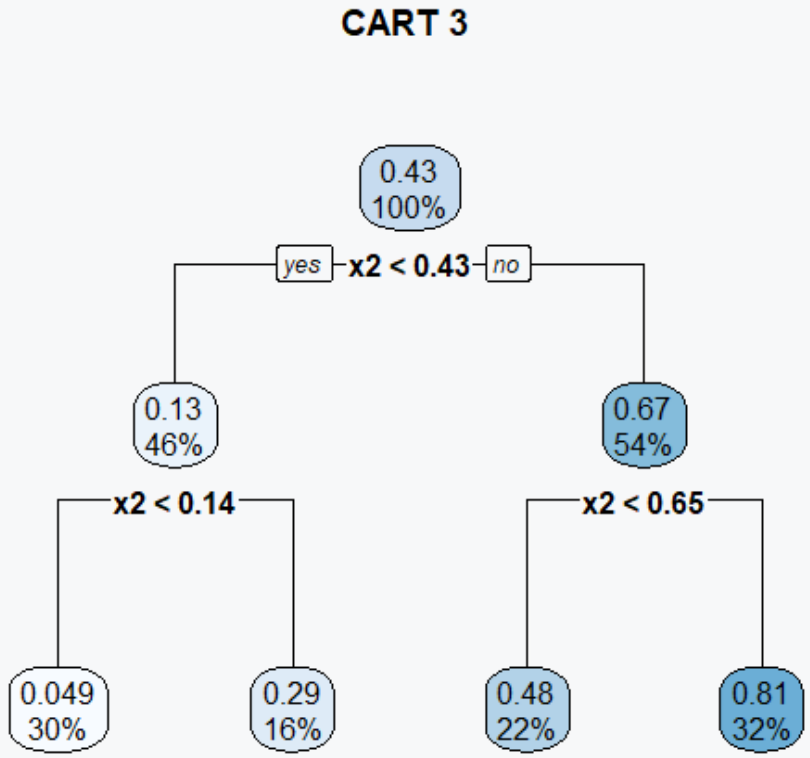
Regressionsbäume

Feature Selection mit Bäumen

Wir wollen nun **noch** einen weiteren Knoten mit davon ausgehenden Ästen zuzulassen.

Wofür hat sich CART entschieden?

Wo stehen wir bzgl. Feature Selection?



Regressionsbäume

CART im Vergleich zur Regression

„Given enough data, trees will fit non-linear .. [functions] and interaction effects **without you having to specify** them in advance.“

Bei der lin. Regression ist die Interaktion auf die Summierung der xi beschränkt.

Wir werden diesen Punkt für Bäume noch vertiefen am konkreten Bsp..

Was mussten wir bei der lin. Regression vor der Schätzung spezifizieren?

„Nonconstant variance is no problem: … you can have completely different error variance in different parts of the input space.“

Was bewirkte dies bei der lin. Regression?

Quelle: Taddy, 2019, 268.

Visualisierung

Feature Selection

Die Auswahl der funktionalen Form wird nun weiter thematisiert.

Dabei überspringen wir die nicht-lineare Regression zunächst.

Bei den Mietpreisen wird sie ihren ersten Auftritt haben.

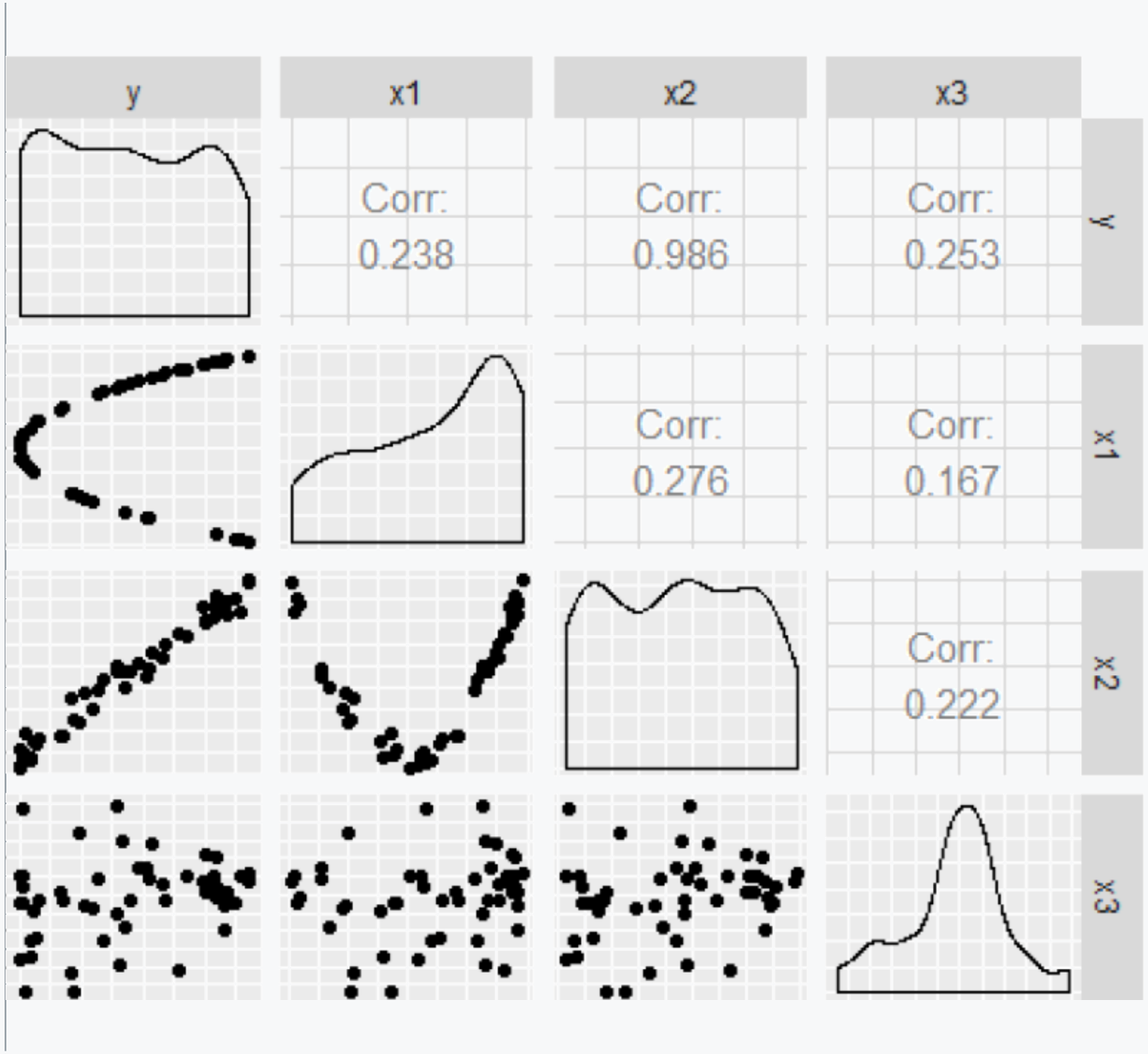
Visualisierung

Beispiel Feature Selection

Weder Bootstrap noch **Big Data** noch **LASSO** noch **CART** haben uns auf den rechten Pfad gebracht.

Eine visuelle Analyse ergibt:

Selbst wenn der RESET die lineare Spezifikation unterstützt, lohnt dennoch eine **visuelle Inspektion, um alternative Spezifikationen auszuloten**.



Resumee über Ökonometrischen Ansatz

Beispiel Feature Selection

Die visuelle Analyse legt eine Variablentransformation nahe.

Das R^2 motiviert weiter unten Ausführungen zum **Overfitting.**

Interpretieren Sie!

Welche Eigenschaften haben die KQ Schätzer?

```
x1sq = x1^2
lm3 = lm(y~x1sq+x2, data = df1)
summary(lm3)

##
## Call:
## lm(formula = y ~ x1sq + x2, data = df1)
##
## Residuals:
##        Min          1Q      Median          3Q         Max
## -9.586e-04 -2.997e-04  5.668e-05  2.694e-04  9.874e-04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0001603  0.0001121  -1.430    0.159
## x1sq         0.9997046  0.0012434 803.984   <2e-16 ***
## x2           0.0002441  0.0011596   0.211    0.834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0004507 on 47 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.195e+07 on 2 and 47 DF,  p-value: < 2.2e-16
```

3 Übung Weiterführung Toyota

Übung

Fallstudie Toyota

Nutzen Sie die am Laborexperiment gezeigten Verfahren, um die Anzahl der Regressoren auf weniger als 10 zu reduzieren.

Rechts **ein mögliches** Ergebnis.

Definieren Sie vorab für sich selbst, was Ihnen dabei wichtig ist.

- Erklärung?
- Vorhersage?
- Verfügbarkeit der Regressoren?
- Plausibilität & Kommunizierbarkeit des Modells?

Variable	Forward	Backward	Both	Exhaustive
Age_08_04	✓	✓	✓	✓
KM	✓	✓	✓	✓
HP	✓	✓	✓	✓
Met_Color				
Automatic				
CC	✓	✓	✓	✓
Doors				
Quarterly_Tax	✓	✓	✓	✓
Weight	✓	✓	✓	✓
Fuel_Type Diesel	✓	✓	✓	✓
Fuel_TypePetrol	✓	✓	✓	✓

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 6, p. 178.

4 Künstliche Neuronale Netze

Neuronale Netze

Multi-Layer-Perceptron

Die Auswahl der funktionalen Form wird weiter thematisiert.

Wir beginnen mit einem historischen Rückblick und Erfahrungen der WestLB aus den 1990‘ern mit der Nutzung für das DAX-Futures Intraday Trading.

Neuronale Netze

Historie

„Neural networks have a long history. Work on these types of models dates back to the mid-twentieth century, e.g., including Rosenblatt‘s perceptron. This early work was focused networks as models that could mimic the actual structure of the brain. In the late 1980‘s, advances in algorithms for training neural networks opened the potential for these models to act as general pattern recognition tools rather than as a toy model of the brain. This led to a boom in neural network research, and the methods developed during the 1990‘s are at the foundation of much of deep learning today. However this boom ended in bust. Because of the gap between promised and realized results (and enduring difficulties in training networks on massive datasets) from the late 1990‘s, neutral networks became just one ML method among many.“ (Taddy, 2019, 299)

Neuronale Netze

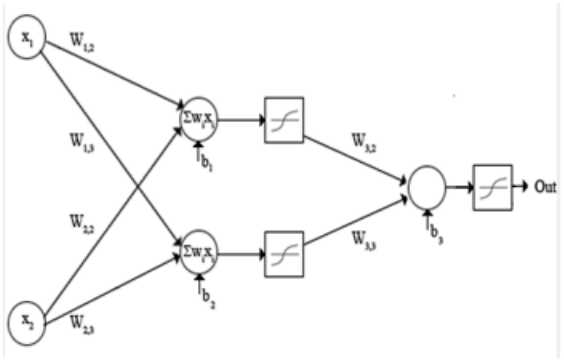
Historie

Folie aus Vortrag**

Resultat*

WestLB

Inputs:
18 Inputs (x_i)
Basierend auf
9 Informationen, die
vor Handelsbeginn
verfügbar sind



Output:
Kauf/Verkaufsents
cheidung

Big Data:
1991-1995: Tägliche Daten, Tick Data für DAX Futures
Start: 236 potentielle Inputs

Source: http://www.eurexchange.com/exchange-de/produkte/idx/dax/DAX--Futures - as of 18 Nov 2015

Literatur:
*Bsp: LP 1996

Wo finden Sie das Problem der Feature Selection wieder?

**Quelle: https://www.fom.de/2018/oktober/fom-professor-erklaert-nutzung-von-big-data-in-der-finanzbranche.html, Zugriff 19. Nov. 2021.
Lit: Lehrbass F und Peter M J (LP 1996) „DAX-Future-Trading mit künstlichen Neuronalen Netzen", ZfgK 4, 4-16

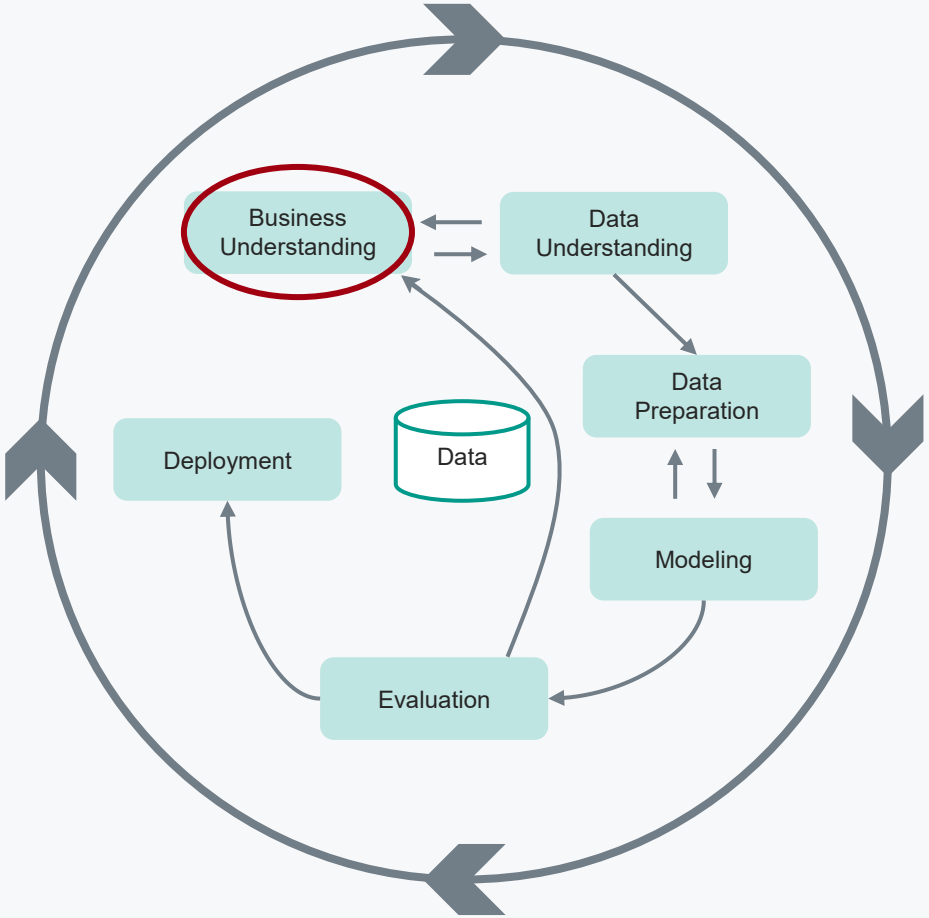
Neuronale Netze

Historie

Folie aus Vortrag**

- Für das **Business Understanding** handelte ich selber DAX-Futures und erkundete bei den Kollegen, auf welche Daten sie schauen und welche Berechnungen sie anstellen.
- Die fachlichen Anforderungen und Ziele waren wie bei einem menschlichen Händler: Gehe Positionen im DAX-Futures mit einem Stop Loss Limit ein und erziele daraus Gewinn.
- Die weiteren Phasen des CRISP-DM werden auf den nächsten Folien skizziert.

Was überrascht Sie?



**Quelle: https://www.fom.de/2018/oktober/fom-professor-erklaert-nutzung-von-big-data-in-der-finanzbranche.html, Zugriff 19. Nov. 2021.

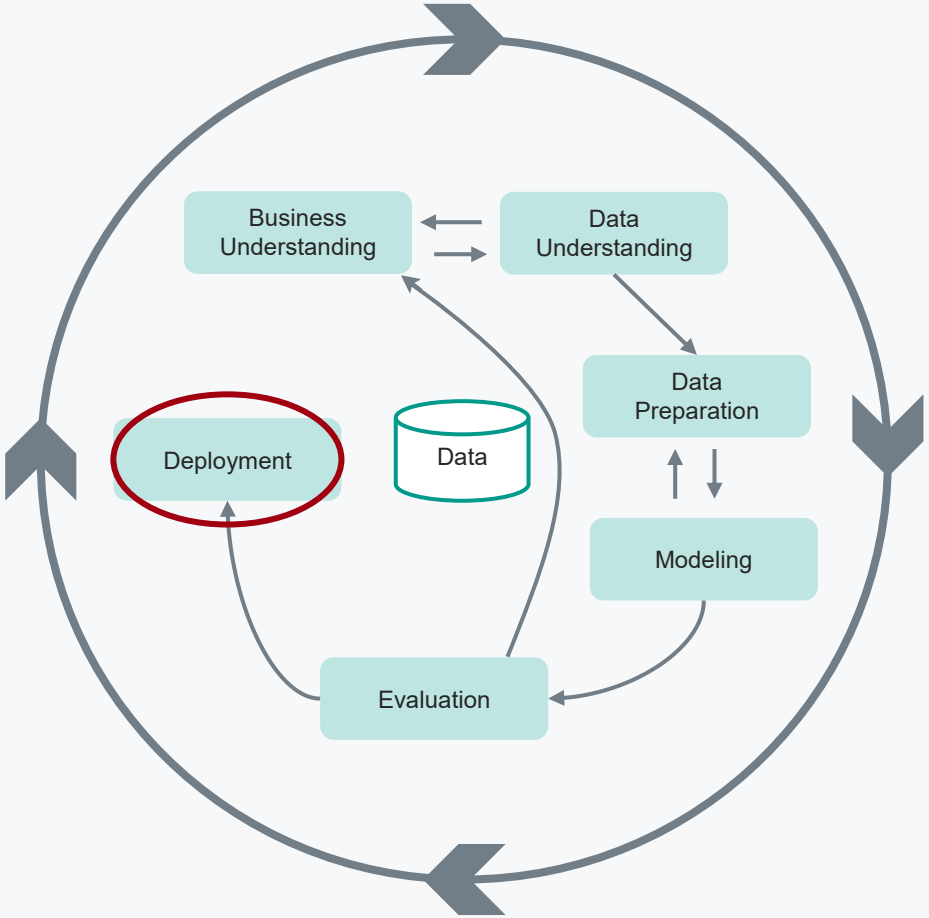
Neuronale Netze

Historie

Folie aus Vortrag**

- Für das **Deployment** waren u.a. folgende Fragen zu beantworten:
- **Wie bestimmen wir den Release-Kandidaten?**
  - Z.Bsp. Bzgl. MLP konkret: Netzarichtektur? → Ockham´s Rasiermesser
- **Wie sieht es aus mit den zentralen Kriterien wissenschaftlicher Modelle:**
- **„Transparency“**
- **„Interpretability“**
- **„Explainability“ (nach R 2019) aus?**
  - Z. Bsp. Lineare Regression von Y auf 18-dim X-Vektor, Einordnung der Koeffizienten nach Vorzeichen und relativer Größ ein bisheriges Marktverständnis
  - Kohonenkarte als Ergänzung („Kohonens selbstorganisierende Karten und der Terminkontrakt auf den DAX“ (mit R. Volmer), WIRTSCHAFTSINFORMATIK, 39, 1997, 339-343)

Was überrascht Sie?



**Quelle: https://www.fom.de/2018/oktober/fom-professor-erklaert-nutzung-von-big-data-in-der-finanzbranche.html, Zugriff 19. Nov. 2021.

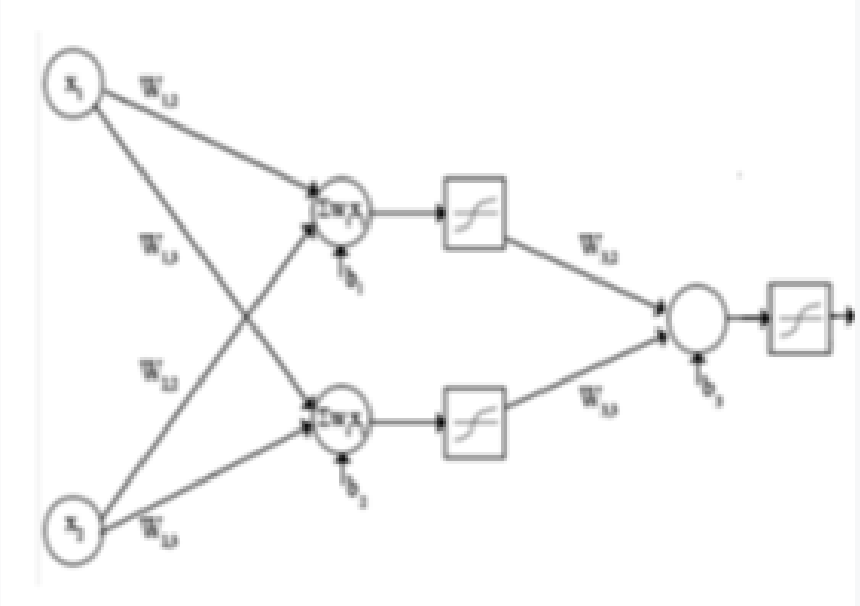
Neuronale Netze

Multi-Layer-Perceptron als nicht-lineare Regression

Wir kommen zurück auf das Single-Layer-Perceptron der WestLB.

In der Grafik gehen zwei Inputs x1 und x2 gewichtet in einen Hidden-Layer mit zwei Neuronen ein. In jedem Neuron wird eine Funktion f(.) ausgeführt.

Die beiden Outputs dieser Neuronen gehen summiert in das Output Neuron ein, welches unsere Vorhersage für y generiert.



y

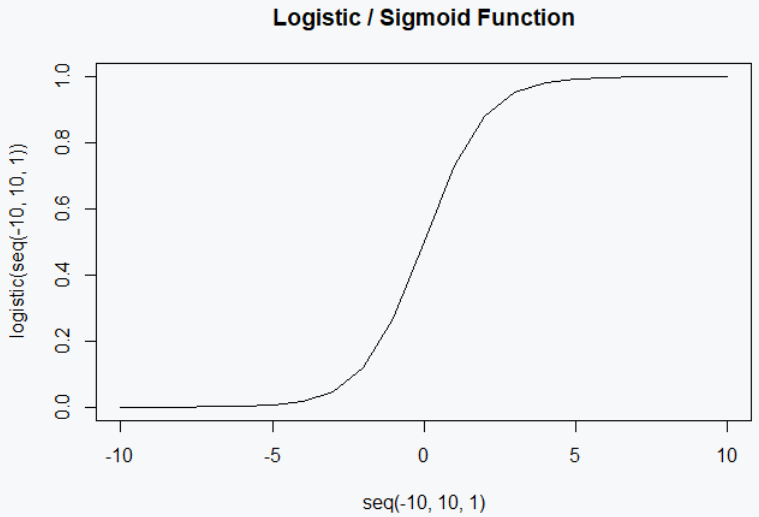
$y = \boldsymbol{f}(w_{h1o}{}^*\,\text{h1} + w_{h2o}{}^*\,\text{h2} + w_{hbo})$ mit $h_i = \boldsymbol{f}(w_{i1}{}^*\,\text{x1} + w_{i2}{}^*\,\text{x2} + w_{ih})$, i=1,2

Neuronale Netze

Multi-Layer-Perceptron

Als Aktivierungsfunktion f(x) wurde verwendet:

```
logistic <- function(x)
{
        1/(1+exp(-x))
}
```



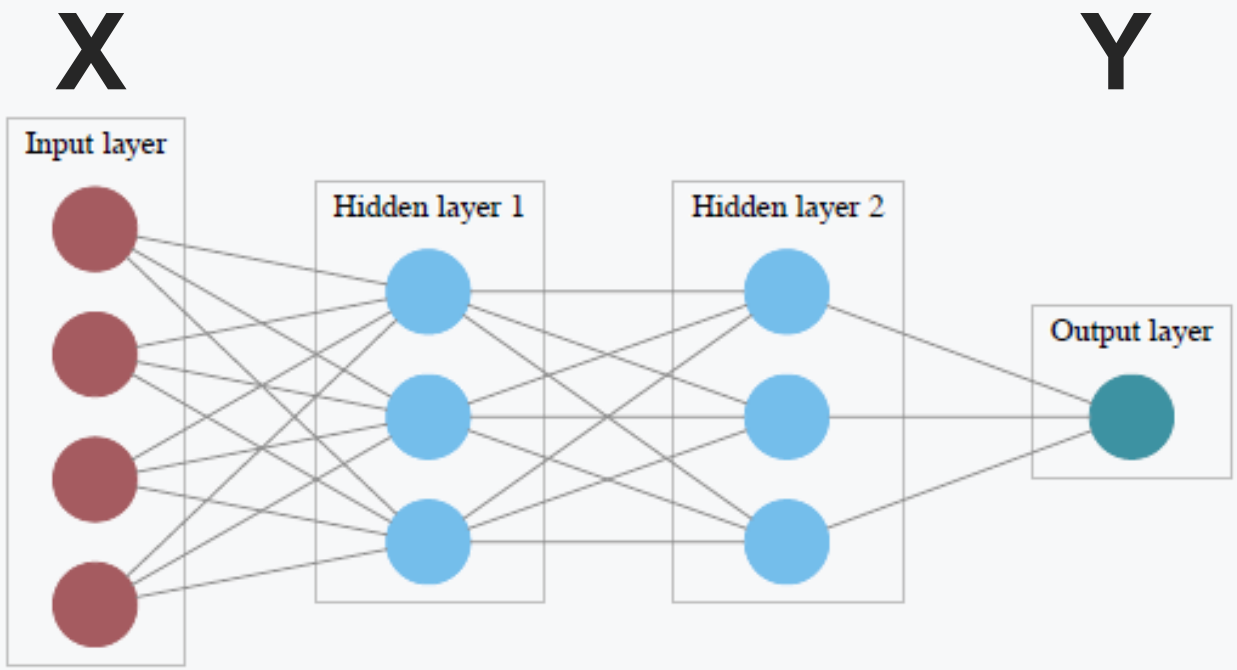
Man beachte, dass diese Funktion konvexe und konkave Bereiche aufweist!

An welches Regressionsmodell aus dem Bachelor erinnert Sie ein Hidden-Neuron Sigmoid / Logistic (x)?

Neuronale Netze

Multi-Layer-Perceptron

Auch eine Architektur mit zwei Hidden Layers ist möglich:



Wir arbeiten b.a.w. mit nur einem Hidden Layer. Begründung folgt.

Wie unterscheiden sich die beiden schematischen Darstellungen?

Brauchen wir so viele Hidden Layer?

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 11, p. 285.

Neuronale Netze

Universal Approximation Theorem

Im **Training** (->Schätzung) werden die **Residuenquadrate** minimiert, d.h. unsere Zielfunktion ist $(Y-f(X,W))^2$ und wird minimiert.

Je nachdem wie die Gewichte W gewählt werden, erhalten wir eine konkave oder konvexe Funktion. Je nachdem wieviele Hidden Neuronen wir erlauben, erhalten wir recht komplexe Funktionen. Zeit für unser erstes Theorem:

Universal Approximation Theorem

Jede stetige Funktion y(x) kann durch ein MLP mit einer Hidden Schicht approximiert werden

(Goodfellow et al, 2016, 192, bewiesen in: Hornik, K., Stinchcombe, M., White, H., 1989, “Multilayer Feedforward Networks are Universal Approximators”, Neural Networks, 2, 359-366).

Ist damit gesagt, wie viel Hiddens benötigen werden?

Und wie einfach diese zu trainieren sind?

Neuronale Netze

Multi-Layer-Perceptron

In unserem Beispiel haben wir drei Inputs, d.h. x1 bis x3. Erstmaliges Training des Netzes wird ausgeführt durch den Befehl:

net0 = neuralnet(modelform, data = df1, hidden = 1, err.fct = "sse", lifesign = "full", act.fct = "logistic", linear.output = F)

Wie viele Hidden Neurons gibt es?

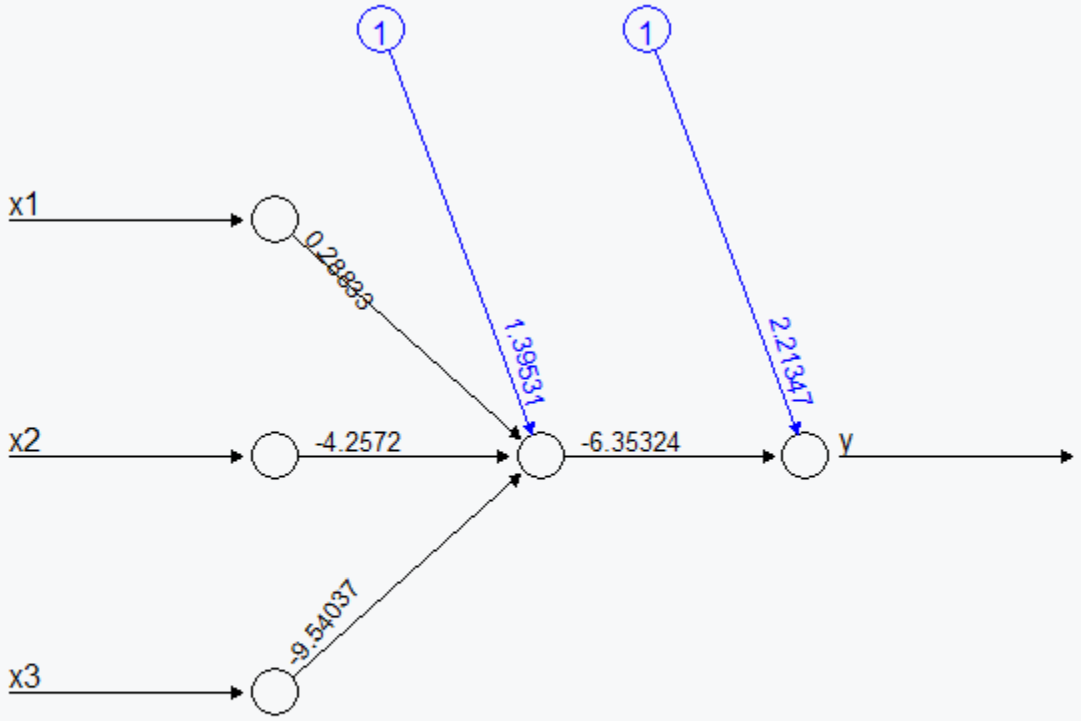
Warum?

Wofür stehen die Argumente? Erkennen Sie Vertrautes wieder?

Neuronale Netze

Multi-Layer-Perceptron

In unserem Beispiel haben wir x1 bis x3. Training des Netzes ergibt:



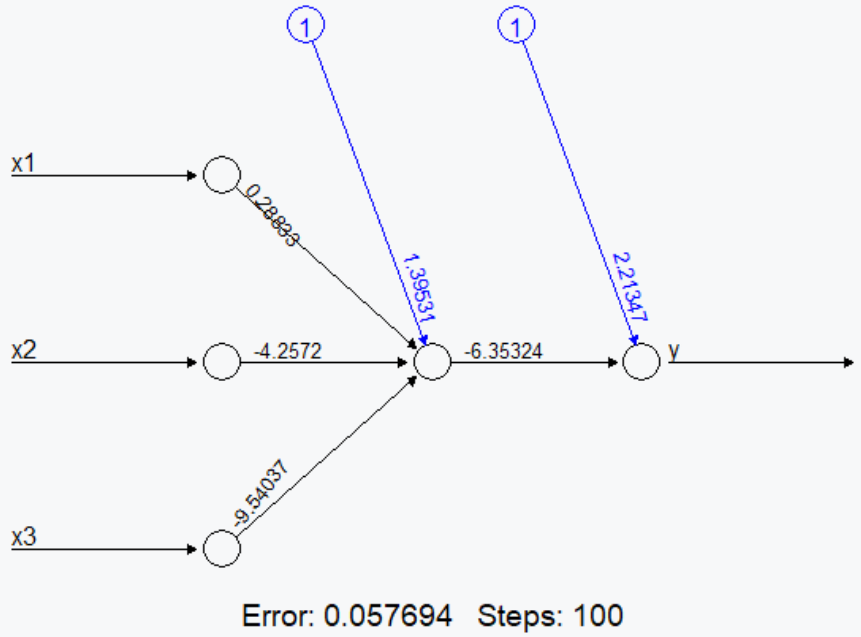
Error: 0.057694 Steps: 100

Neuronale Netze

Multi-Layer-Perceptron

Wir finden die Gewichte aus der Grafik auch in der Gewichtsmatrix von net0.

```
net0$result.matrix[4,1]#Intercept.to.1layhid1

## Intercept.to.1layhid1
##              1.395305

net0$result.matrix[5,1]#x1.to.1layhid1

## x1.to.1layhid1
##      0.2883294

net0$result.matrix[6,1]#x2.to.1layhid1

## x2.to.1layhid1
##      -4.257198

net0$result.matrix[7,1]#x3.to.1layhid1

## x3.to.1layhid1
##      -9.540368
```



Den ausgewiesenen Fehler können wir nachrechnen:

```
round(sum((y-net0_fc)^2)/2,5)

## [1] 0.05769
```

Neuronale Netze

Multi-Layer-Perceptron

Nutzt man die trainierten Gewichte, so lautet die erste Vorhersage des Modells

```
net0_fc[1,1]

## [1] 0.05893039
```

Dies wollen wir nachrechnen und nutzen abermals die Gewichtsmatrix von net0

```
single_hidden_neuron =
logistic(x1[1]*net0$result.matrix[5,1]+x2[1]*net0$result.matrix[6,1]+x3[1]*ne
t0$result.matrix[7,1]+net0$result.matrix[4,1])#a+bx
#from first hidden layer to output neuron
logistic(net0$result.matrix[9,1]*single_hidden_neuron+net0$result.matrix[8,1]
)

## 1layhid1.to.y
##    0.05893039
```

Was wir hier betrieben haben nennt man **Feedforward**, d.h. Ausführung von f(X).

Multi-Layer-Perceptron (MLP)

Gradient Descent

Die Minimierung geschah numerisch. Da das Problem im Regelfall nicht-linear ist, gibt es keine analytische Lösung.

Zudem wissen wir nicht, ob die Funktion f(X,W) konkav oder konvex oder was auch immer ist. Gleichsetzen des Gradienten (Ableitungen nach Wi) mit dem Nullvektor weist somit nicht zwingend den Weg zum Minimum. Es könnte auch ein Maximum sein.

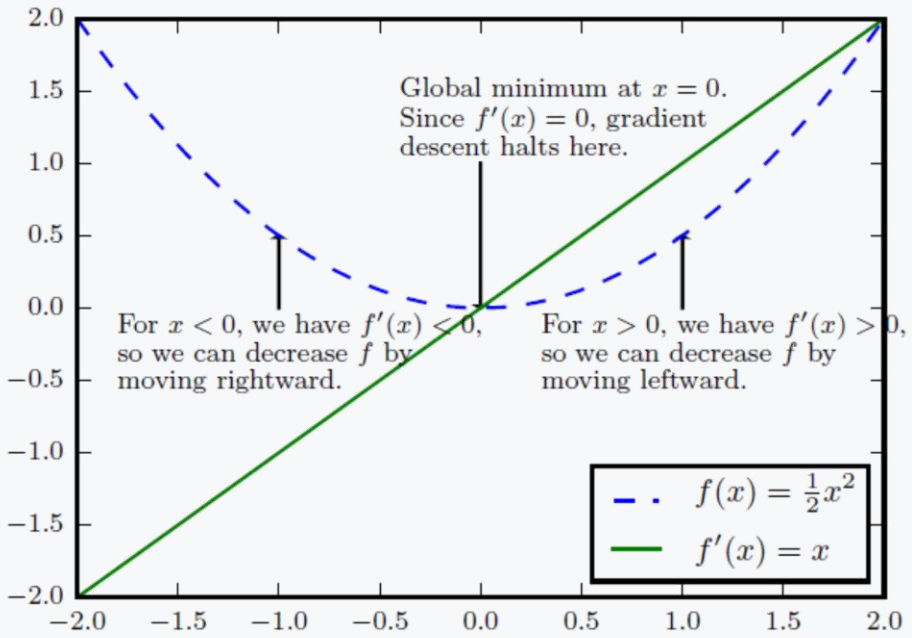
Dennoch hat der Gradient einen informatorischen Wert! Er weist in die Richtung der grössten Steigung im W-Raum. Wenn wir uns also die Gewichte in die Gegenrichtung anpassen, nähern wir uns dem Minimum!

Dies nennt man Gradientenabstieg und wird nachfolgend illustriert.

Multi-Layer-Perceptron (MLP)

Gradient Descent

Der Gradientenabstieg im Bsp. Y = ½ x² (Goodfellow, 2016, 80)



Rechts von x=0 ist der Gradient positiv, also ist die Richtung zur Erhöhung des Funktionswertes x+. Da wir minimieren wollen gehen wir in die entgegengesetzte Richtung also x-.

Neuronale Netze

Multi-Layer-Perceptron (MLP)

Backpropagation

Das MLP ist eine komplexe Funktion von X und W, kurz Funktion f(X,W). Backpropagation ist „a method for calculating gradients on the parameters of a network. In particular, backprop is just an .. implementation of the chain rule from calculus“ (Taddy, 2019, 303).

Was sind die „parameters“?

Wie übersetzt man „chain rule“?

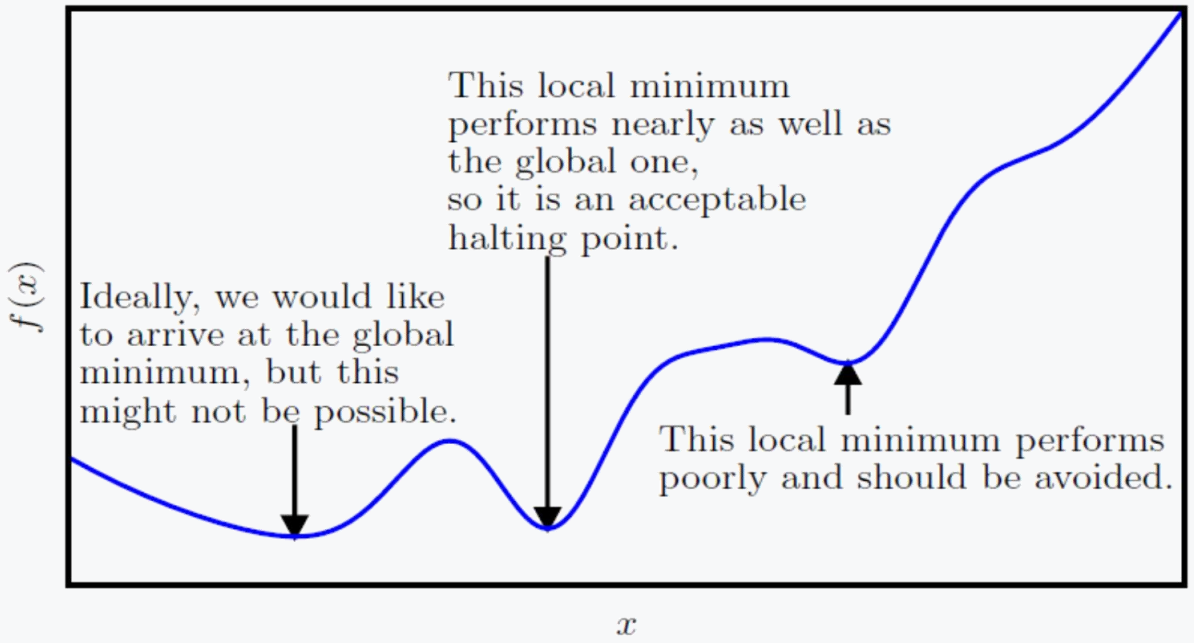
Die Berechnung des Gradienten wächst mit der Datenmenge. Deshalb nutzt man in der Praxis „estimates of those gradients based upon a subset of the data. This is the **SGD*** algorithm“ (Taddy, 2019, 304). Den Subset nennt man „**minibatch**“ (Goodfellow, 2016, 148). Typische Grössen sind 1 bis einige Hundert.

*Stochastic Gradient Descent

Multi-Layer-Perceptron (MLP)

Backpropagation

Die vielen Facetten des Netztrainings werden im MBB** behandelt. Einen Einstieg bietet Goodfellow, 2016, Kap. 1-8). Ein Problem numerischer Verfahren ist etwa das Folgende (ibid., 81):



Wir vertiefen nun das Bsp..

*Stochastic Gradient Descent **Master of Science (M.Sc.) Big Data & Business Analytics

Neuronale Netze

Multi-Layer-Perceptron (MLP)

Netztraining

Von den vielen Facetten des Netztrainings illustrieren wir mögliche Regen für das Trainingsende:

„With sufficient iterations, **neural net can easily overfit the data**

To avoid overfitting:

- Track error in validation data or via cross-validation
- Limit iterations
- Limit complexity of network”

Welches Theorem steckt hinter dem einleitenden englischen Satz?

Im Paket **neuralnet** wird als Default eine weitere „stopping rule“ verwendet:

„The process stops if a pre-specified criterion is fulfilled, e.g. if all absolute partial derivatives of the error function with respect to the weights … are smaller than a given threshold” (Default 0.01, aus Anlage 01_01 RJournal_2010-1_Guenther+Fritsch.pdf).

Dauert das Training zu lange kann man threshold hochsetzen.

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 11, p. 291.

Neuronale Netze

Multi-Layer-Perceptron (MLP)

Netztraining

Im Paket **neuralnet** wird als Default eine weitere **Variante des SGD** verwendet:

„Instead of the magnitude of the partial derivatives only their sign is used to update the weights. […] In order to speed up convergence in shallow areas, the learning rate .. will be increased if the corresponding partial derivative keeps its sign. On the contrary, it will be decreased if the partial derivative of the error function changes its sign since a changing sign indicates that the minimum is missed due to a too large learning rate. Weight backtracking is a technique of undoing the last iteration and adding a smaller value to the weight in the next step. Without the usage of weight backtracking, the algorithm can jump over the minimum several times.” (aus Anlage 01_01 RJournal_2010-1_Guenther+Fritsch.pdf).

Eine Weiterentwicklung des Default Lernalgorithmus “resilient backpropagation with and without weight backtracking = rprop+ findet sich in Aristoklis et al. (2005). Als Datei 01_01 Convergent training.pdf.
Auch diese ist in neuralnet verfügbar.

*Quelle: Shmueli, G., Bruce, P. C., Gedeck, P., Patel, N. R. (2019) Data mining for business analytics: concepts, techniques and applications in Python, Ch. 11, p. 291.

Multi-Layer-Perceptron (MLP)

Globale Modellgüte

Das MLP kann als nicht-lineare Regression verstanden werden. Man kann die R^2 vergleichen:

```
round(cor(lm1$fitted.values,y)^2,4)

## [1] 0.9758

round(cor(net0_fc,y)^2,4)#clear cause

##         [,1]
## [1,] 0.9764
```

Das MLP hat 6 Parameter, die lineare Regression nur 4. Insofern ist der Anstieg im R^2 enttäuschend.

Neuronale Netze

Multi-Layer-Perceptron (MLP)

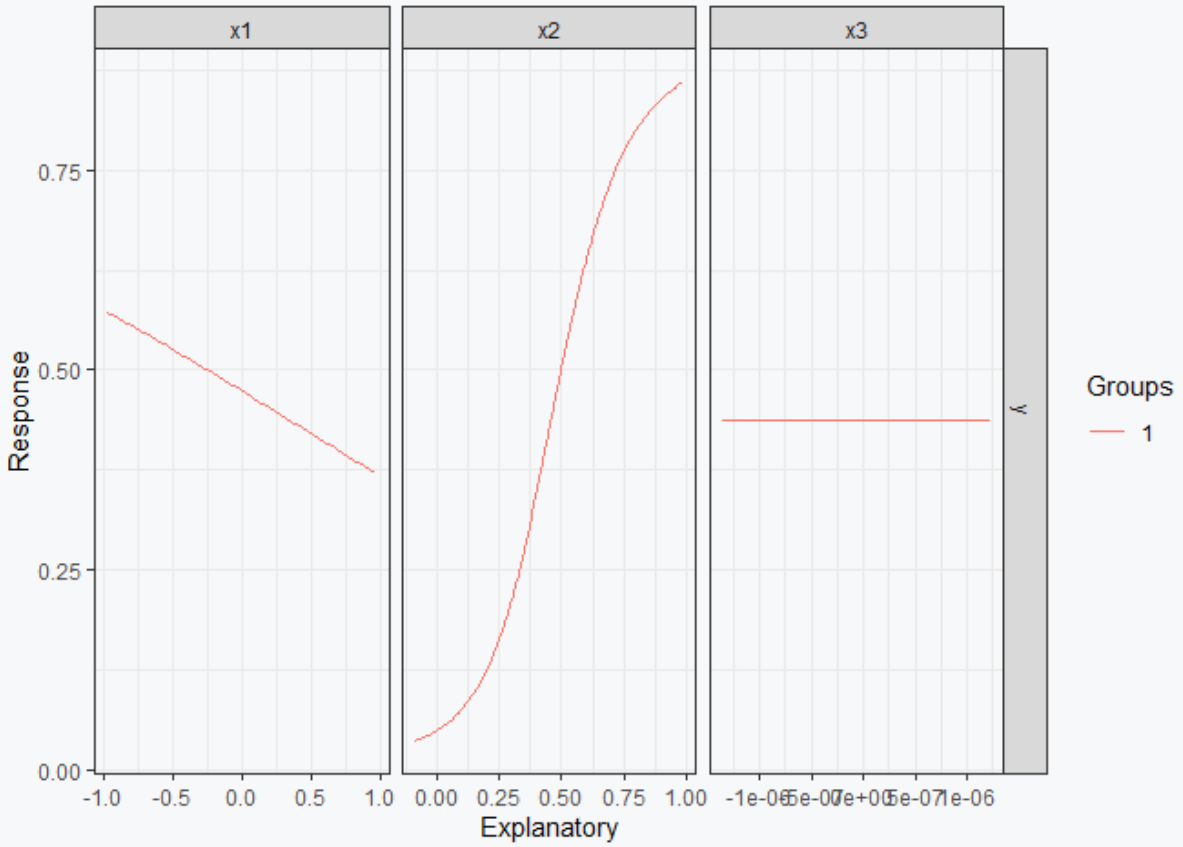
Sensitivitäten

Jedoch wurde das MLP ohne grosse Vorüberlegungen trainiert. Insbesondere wurde nur ein Hidden gewählt, um das Nachrechnen einfach zu halten.

Interessant ist nun, welche funktionale Form gelernt wurde.

Dies verdeutlicht das Lekprofil exemplarisch für Quantilswerte von 50%. D.h. übrige xi werden am Median fixiert.

Offenbar spielt x3 keine Rolle!



Neuronale Netze

Multi-Layer-Perceptron (MLP)

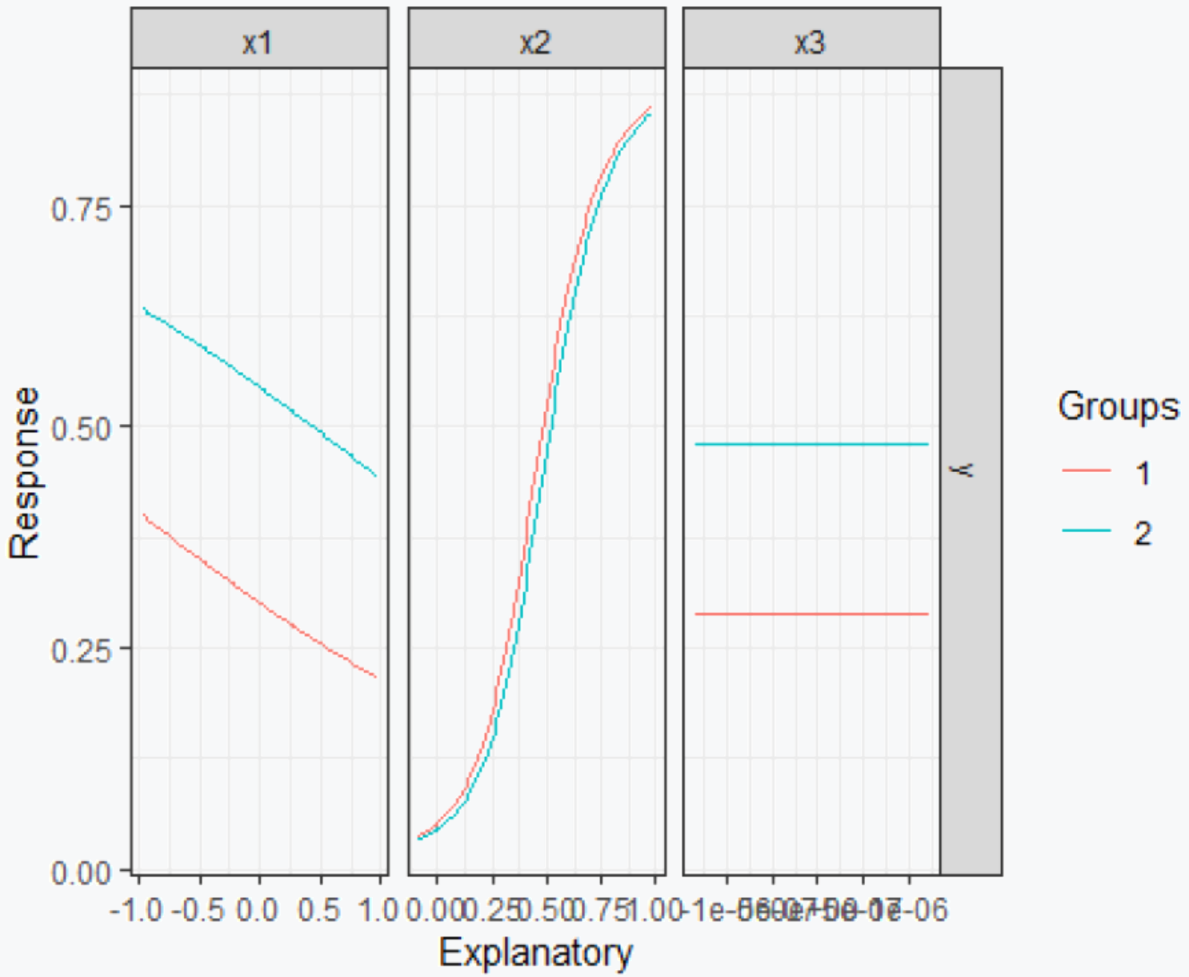
Sensitivitäten

Jedoch wurde das MLP ohne grosse Vorüberlegungen trainiert. Insbesondere wurde nur ein Hidden gewählt, um das Nachrechnen einfach zu halten.

Interessant ist nun, welche funktionale Form gelernt wurde.

Dies verdeutlicht das Lekprofil exemplarisch für Quantilswerte von 40 und 60%.

Offenbar spielt x3 keine Rolle!



Neuronale Netze

Multi-Layer-Perceptron (MLP)

Architektur & Feature Selection

Wir wollen nun die **Anzahl der Hidden Neurons** erhöhen. Eine Praktikerregel (Master, 1993, Kap. 10) ist:

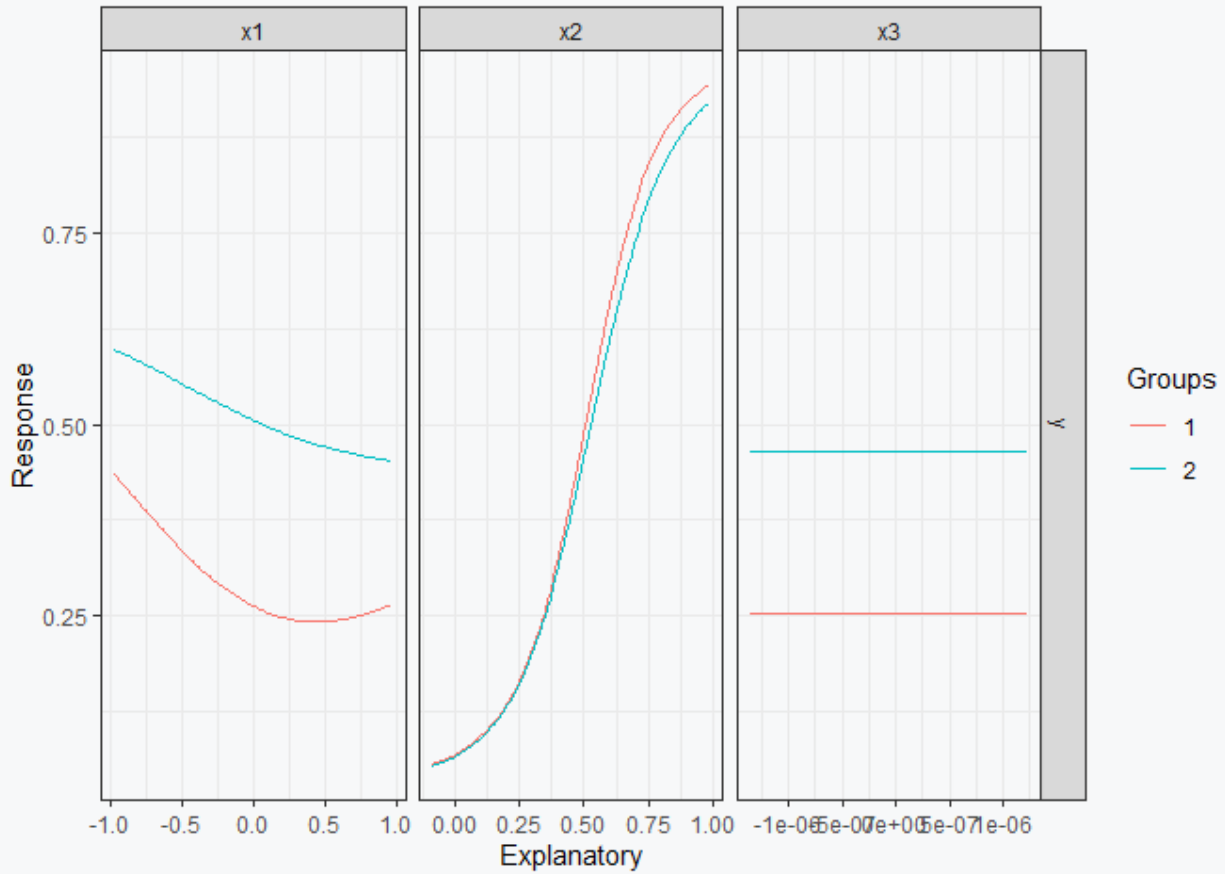
Dimensionen X + Y mal 2/3, also

(3+1) * 2/3 = 2,66 abgerundet („Ockham“) auf 2

Das R^2 steigt erwartungsgemäss auf 0,9777.

Hervorzuheben ist die in x1 erkannte funktionale Form!!!

Dank ausreichend reicher Architektur konnte erstmals x1 als quadratisch wirksam erkannt werden!



Beim MLP finden Spezifikation und Schätzung simultan statt. Dies unterstützt die Feature Selection.

Neuronale Netze

Multi-Layer-Perceptron (MLP)

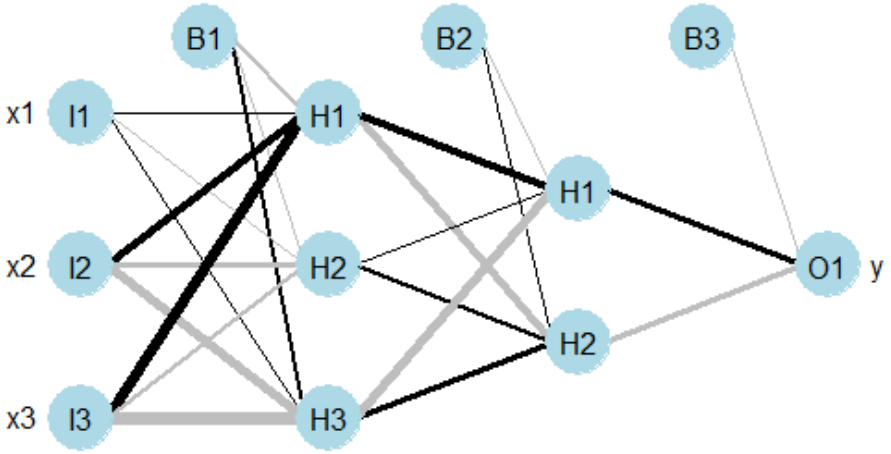
Architektur

Wir wollen nun die Anzahl der Hidden und Schichten erhöhen.

Eine Praktikerregel (Master, 1993, Kap. 10) empfiehlt einen pyramidalen Aufbau (hier 3-2-1).

Das R^2 fällt überraschenderweise auf 0,9715.

Zwei Schichten sind nach dem Theorem unnötig und werden erst bei Unstetigkeiten benötigt.



Multi-Layer-Perceptron (MLP)

Güte der Schätzer

Es folgt ein weiteres Theorem. Angenommen der DGP ist

$$y_t = f(x_t, W) + u_t$$

Der Störterm erfülle zudem die Gauss-Markov Annahmen, d.h. uiv mit endlicher Varianz und Erwartungswert Null.
Die Gewichtsmatrix W enthalte keine irrelevanten Elemente und es gebe keine überflüssigen Inputs.
Dann sind die trainierten Gewichte konsistente Schätzer der wahren Matrix W.

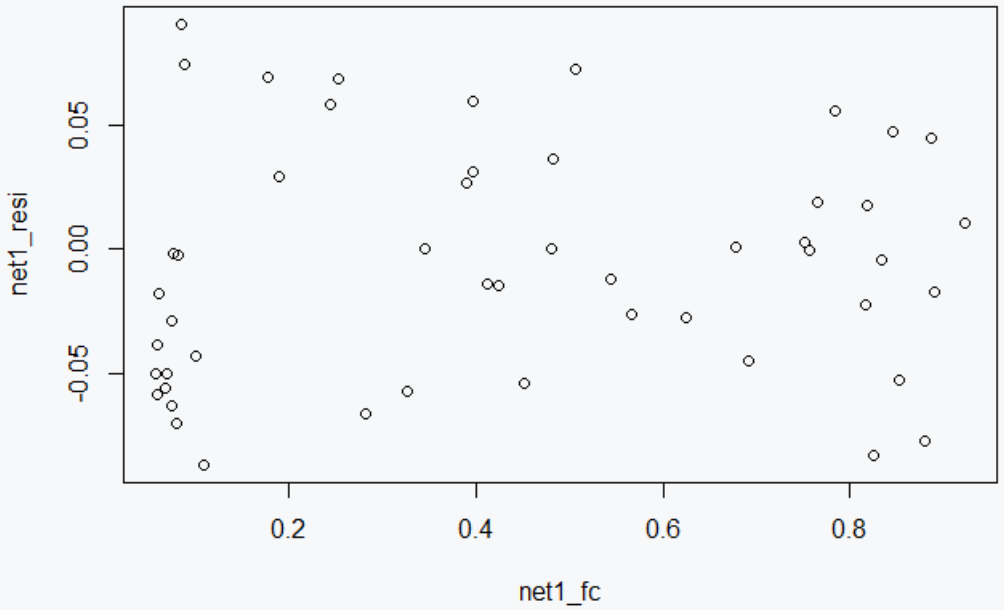
D.h. für einen wachsenden Stipo Umfang konvergiert $(W)\hat{}_n$ gegen die wahre Matrix mit Wahrscheinlichkeit eins (White, 1989).

Was müssen wir leisten, ehe wir dies nutzen dürfen?

Neuronale Netze

Multi-Layer-Perceptron (MLP)

Diagnostik

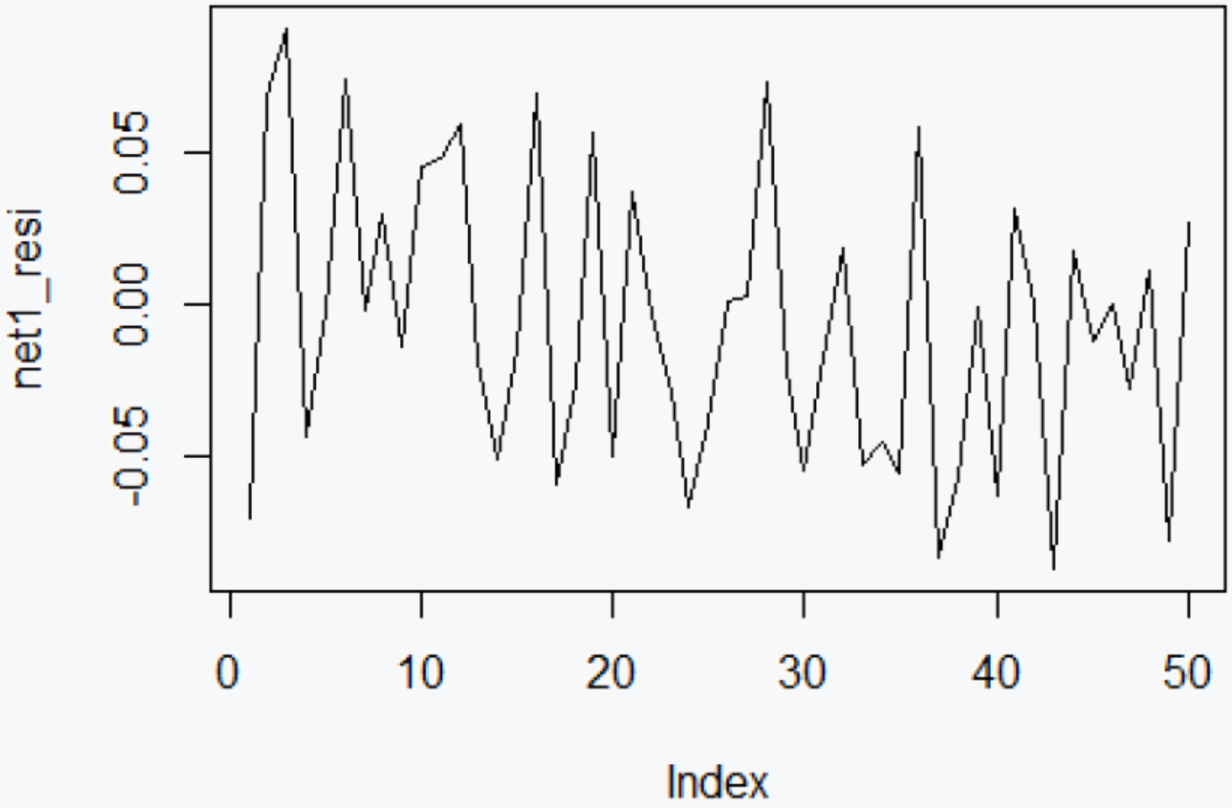


Interpretieren Sie!

Neuronale Netze

Multi-Layer-Perceptron (MLP)

Diagnostik



Interpretieren Sie!

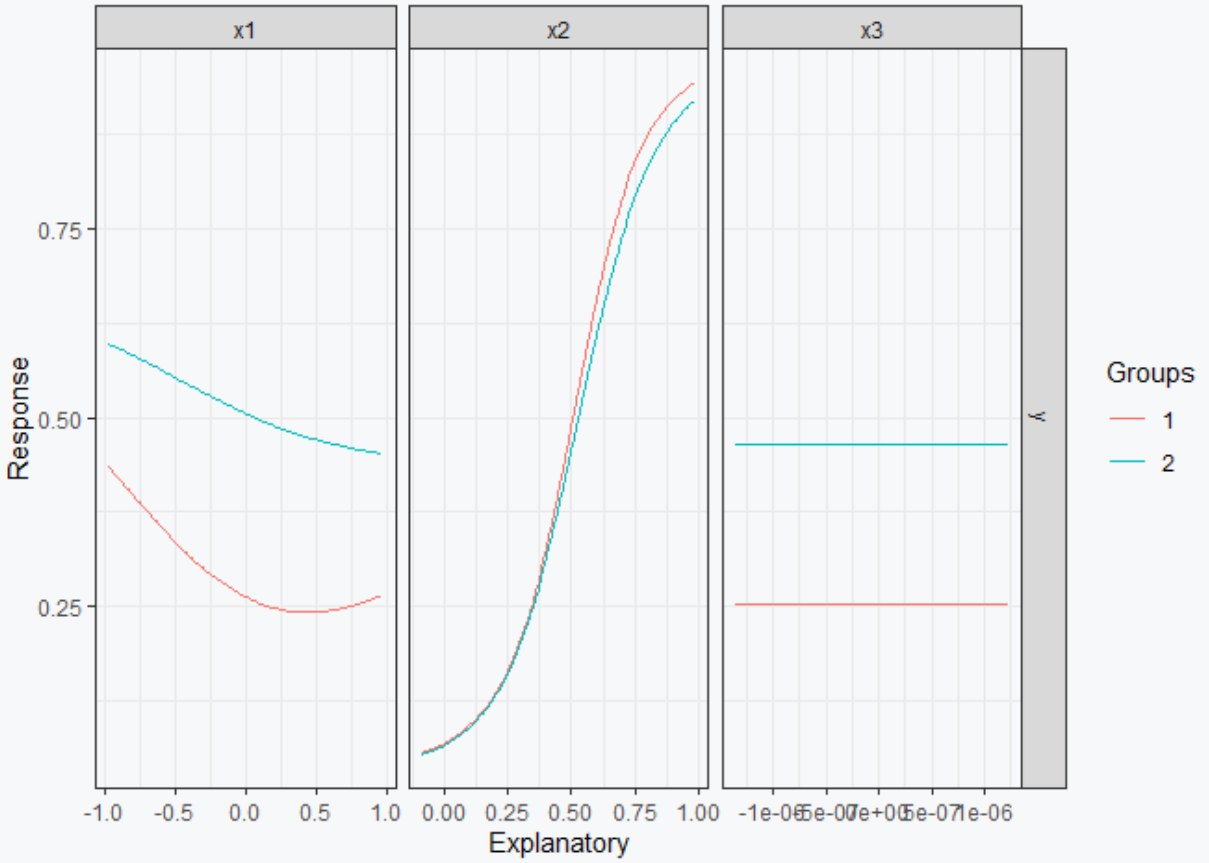
Neuronale Netze

Multi-Layer-Perceptron (MLP)

Diagnostik

Das diagnostizierte Single Layer Two Hidden Neurons Netz ist unser bisheriger Favorit. Jedoch gibt das Lekprofile – anders als der vif - den Hinweis darauf, dass x3 überflüssig ist!

Wir trainieren deshalb erneut nur auf x1 und x2.



Resumee über Machine Learning Ansatz

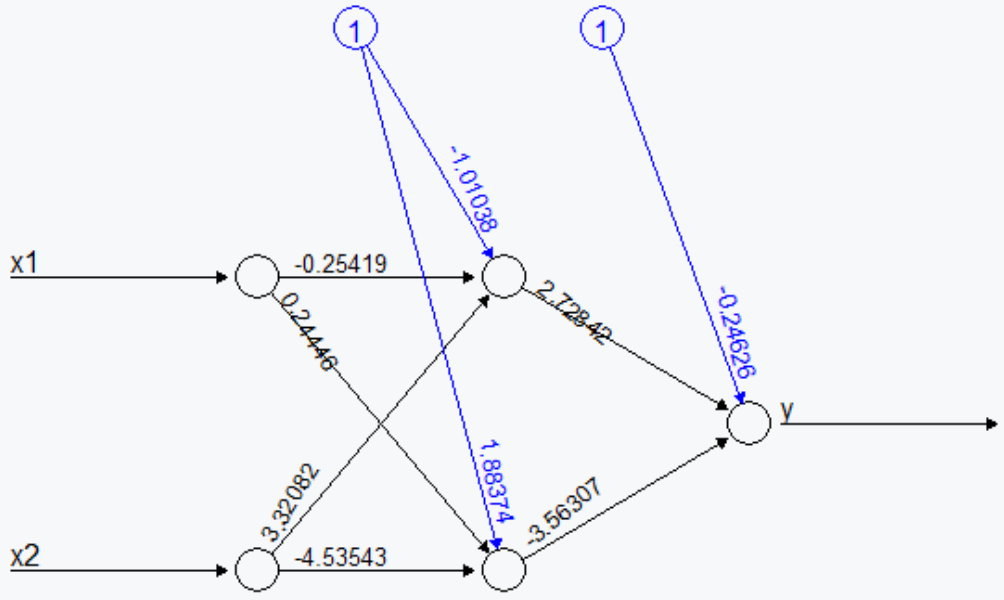
Beispiel Feature Selection

Das MLP sieht auch in der Diagnostik gut aus (vgl. Coding). **Jedoch ist sein R^2 gesunken auf 0,9775** (von 0,9777).

Den in-sample fit sollte man nicht überbewerten.

Hier ging es darum, ob die relevanten Faktoren und die passende funktionale Form entdeckt werden konnten!

Von den Gewichten des MLP kann vermutet werden, dass sie konsistente Schätzer sind, da die Annahmen des Satzes gegeben scheinen.



Resumee über Machine Learning Ansatz

Erkenntnis

Mit dem Training des MLP haben wir **datengetriebene Modellierung** betrieben.

Dabei geschieht Spezifikation und Schätzung/Training in Einem. Zu Beginn wird der "datengenerierende Prozess ... als Black Box angenommen. Das so erweiterte Methodenspektrum ermöglicht das Generieren neuer Erkenntnisse".

Dank des Satzes von White (1989) schätzt man konsistent, wenn bestimmte Annahmen erfüllt sind.

Man nähert sich also der unbekannten Funktion y=f(x) mit zunehmender Stipogröße an, wenn die Störterme bestimmte Annahmen erfüllen.

Deshalb ist auch hier Diagnostik angezeigt.

Und aus (Prognose-)Fehlern sollte man sowieso stets lernen wollen (ATOM).

Quelle: Lübke/Krol (2022) Empirisch-quantitative Abschlußarbeiten, in: Boßow-Thies/Krol (Hrsg.) Quant Forschung im Masterarbeiten, S 502.

Ausblick Projektidee

Motivation

Jemand ohne eigene Daten könnte **weiter auf den Toyota Daten forschen.**
Jedoch muss man sich vorab den Wertebereich des Outputneurons bewusst machen.

Angenommen die Aktivierungsfunktion ist sigmoid. **Welcher Wertebereich wird abgedeckt?**

Angenommen ihr Y ist auch ausserhalb unterwegs. **Was könnte man tun?**

```
MinMaxTrafo <- function(x)
{
(x-min(x))/(max(x)-min(x))
}
```

Bei Nutzung des Pakets neuralnet kann man es auch einfacher haben, indem man das Argument `linear.output = F` auf T setzt.

Ausblick Projektidee

Motivation

Wir wiederholen wir den Schätzoutput von oben unter Verwendung der HAC SE:

```
##                     Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)      -8.6220e+02  5.1066e+03  -0.1688 0.8659582
## Age_08_04        -1.1939e+02  5.8701e+00 -20.3387 < 2.2e-16 ***
## KM               -1.9238e-02  2.1295e-03  -9.0339 < 2.2e-16 ***
## Fuel_TypeDiesel   6.7424e+02  4.6166e+02   1.4605 0.1444828
## Fuel_TypePetrol   1.9525e+03  6.4231e+02   3.0398 0.0024314 **
```

Der Achsenabschnitt von Null ist unschön, weil dadurch der Basispreis Null ist, was keinen Sinn macht. Zudem haben wir sehr viele Regressoren und Daten.

Dank unseres Business Understanding entfernen wir **erstens** alle gasbetriebenen Autos (CNG), so dass nur noch die gebräuchlichen Diesel und Benziner übrig bleiben. Wir transformieren diesen Faktor in einen Dummy (1=Petrol).

Ausblick Projektidee

Motivation

Mit Blick auf die 39 Regressoren schätzen wir erneut, nutzen **zweitens** die HAC SE zur Identifizierung hoch signifikanter Faktoren und nutzen **weiteres Business Understanding**, um auf **acht Faktoren** zu reduzieren.

Zu den bereits gezeigten Teilmengen der Regressoren (Tabelle aus Buch von Shmueli) gesellt sich nun diese (Case Toyota Revised.R und csv):

```
coeftest(reg,vcov=NeweyWest(reg))

##
## t test of coefficients:
##
##                    Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)       1.5994e+04  6.1903e+02  25.8374 < 2.2e-16 ***
## Age_08_04        -1.4346e+02  6.6998e+00 -21.4118 < 2.2e-16 ***
## KM               -1.9323e-02  3.4815e-03  -5.5501 3.664e-08 ***
## Petrol_Dummy     -1.9336e+03  4.9542e+02  -3.9029 0.0001015 ***
## HP                5.2630e+01  6.2098e+00   8.4753 < 2.2e-16 ***
## Guarantee_Period  1.6914e+01  1.7002e+01   0.9948 0.3200635
## Automatic_airco   2.7545e+03  3.6592e+02   7.5275 1.160e-13 ***
## Boardcomputer    -3.6729e+02  1.4370e+02  -2.5561 0.0107347 *
## Sport_Model       5.6388e+02  1.2197e+02   4.6231 4.279e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Das R^2 beträgt 86,85% obwohl wir deutlich ausgedünnt haben.

Ausblick Projektidee

Motivation

Auch wenn es im Gegensatz zum **Business Understanding** steht, entfernen wir nun die insignifikante Garantieperiode. Erneute Schätzung und Diagnostik führt zu:

```
coeftest(reg_slim1,vcov=NeweyWest(reg_slim1))

##
## t test of coefficients:
##
##                    Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)      1.6148e+04  6.0784e+02  26.5670 < 2.2e-16 ***
## Age_08_04       -1.4534e+02  6.8365e+00 -21.2600 < 2.2e-16 ***
## KM              -1.9315e-02  3.4924e-03  -5.5305 4.083e-08 ***
## Petrol_Dummy    -1.9147e+03  4.9667e+02  -3.8550 0.0001232 ***
## HP               5.2794e+01  6.1624e+00   8.5671 < 2.2e-16 ***
## Automatic_airco  2.7220e+03  3.6561e+02   7.4451 2.099e-13 ***
## Boardcomputer   -4.1520e+02  1.3931e+02  -2.9803 0.0029495 **
## Sport_Model      5.3703e+02  1.2192e+02   4.4048 1.174e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Predictors	*Estimates*
(Intercept)	16148.37
Age_08_04	-145.34
KM	-0.02
Petrol_Dummy	-1914.68
HP	52.79
Automatic_airco	2722.00
Boardcomputer	-415.20
Sport_Model	537.03

Diese Modell ist besser kommunizierbar. Der Basispreis beträgt 16148 USD. Jedes Jahr senkt den Preis um 145 USD. Usw.

Ein Händler kann anhand weniger Attribute den Preis abschätzen.

Das R^2 beträgt 86,83% obwohl wir nur noch 7 Faktoren haben. Darauf liesse sich nun ein MLP trainieren u.a..

5 Common Problems: Overfitting & Underfitting

Common Problems: Overfitting and Underfitting

Motivation

Heutzutage werden MLP mit sehr vielen Schichten trainiert. Die Anzahl der Gewichte geht bis in die Milliarden. Der universelle Approximator MLP ist stets der Gefahr ausgesetzt, auch Rauschen mitzulernen. Dies nennt man Overfitting.

„When you have big data and many inputs, it is easy to overfit the training data so that your model is being driven by noise that will not be replicated in future observations. That adds errors to your predictions, and it is possible that the overfit model becomes worse than no model at all“ (Taddy, 2019, 70).

The next slide from Goodfellow visualizes the issue.

Overfitting and Underfitting

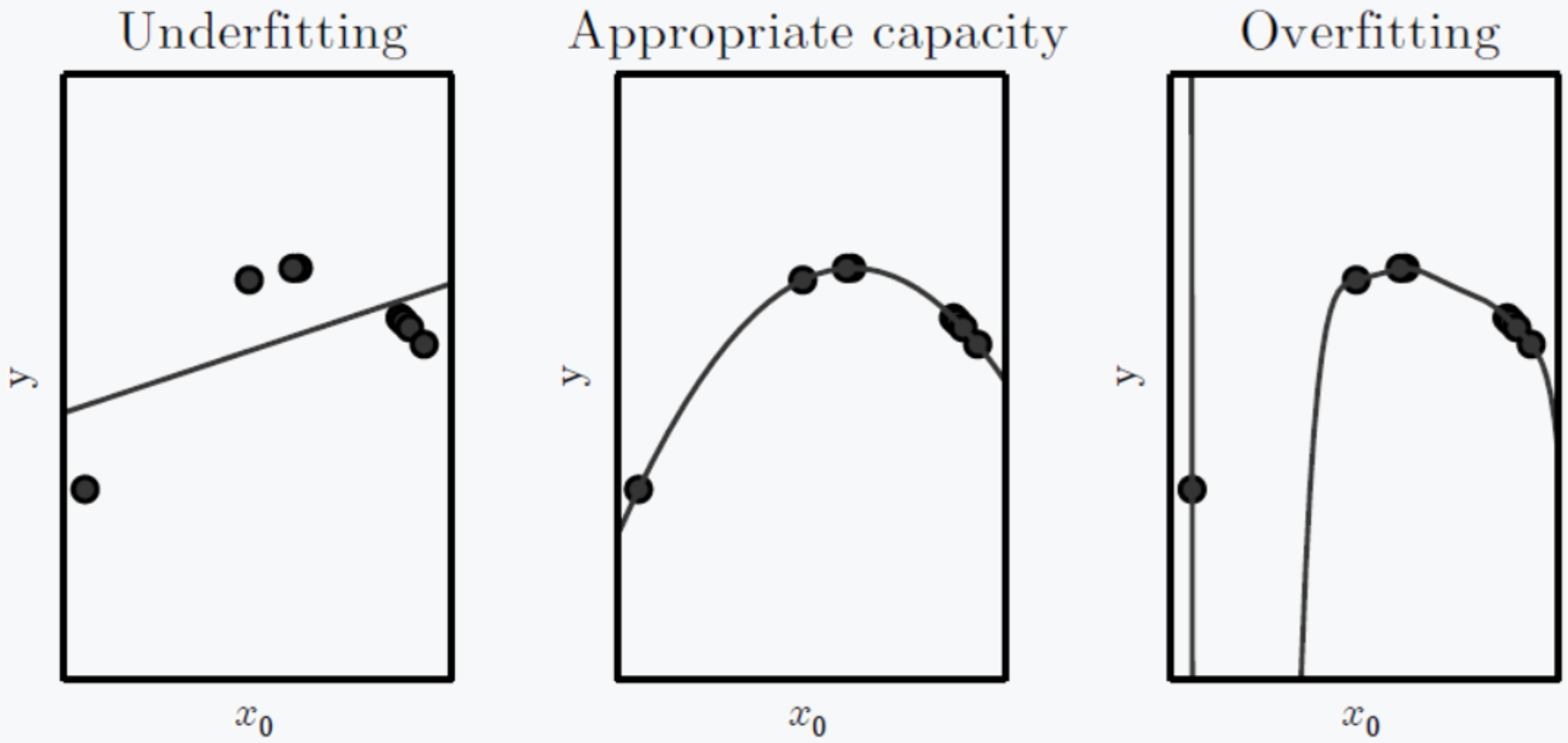


Figure 5.2

(Goodfellow 2016)

Quelle: Goodfellow, 2016, S. 110

Overfitting and Underfitting

Regularization again

“Three situations, where the model family being trained either (1) excluded the true data generating process—corresponding to underfitting and inducing **bias**, or (2) matched the true data generating process, or (3) included the generating process but also many other possible generating processes—the overfitting regime where **variance** rather than bias dominates the estimation error. The goal of **regularization** is to take a model from the third regime into the second regime.” (Goodfellow, 2016, 222)

Assign the numbers (1) to (3) to each of the charts!

Overfitting and Underfitting

The Dilemma

Underfitting	Overfitting
But similar forecasting quality when applied to new data (e.g. validation, test set)	Fails when applied to new data because (past) random error realizations have been memorized
Less costly to set up / maintain model	High costs, violation of Ockhams´s rule
Incomplete Fitting Quality within training data	Fifference in Fitting Quality between data sets

"… the overfitting regime where **variance** rather than **bias** dominates the estimation error" (Goodfellow, 2016, 222)

What is your choice? Which errors do you prefer? What is particularly devastating in practice, e. g. during CRISP's last step of deployment?

Nice video on topic: https://www.youtube.com/watch?v=EuBBz3bI-aA&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=5

Overfitting and Underfitting

The Dilemma

As a formula:

"It is possible to show that the expected test MSE, for a given value x_0, can always be decomposed into the sum of three fundamental quantities: the **variance** of $\hat{f}(x_0)$, the squared **bias** of $\hat{f}(x_0)$ and the variance of the error term

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

Expected **Test** Set MSE = Wie ändert sich $\hat{f}$ bei **anderen Trainingsdaten?** + Fehler durch Modellvereinfachung + Unvermeidliche Varianz aus Störterm

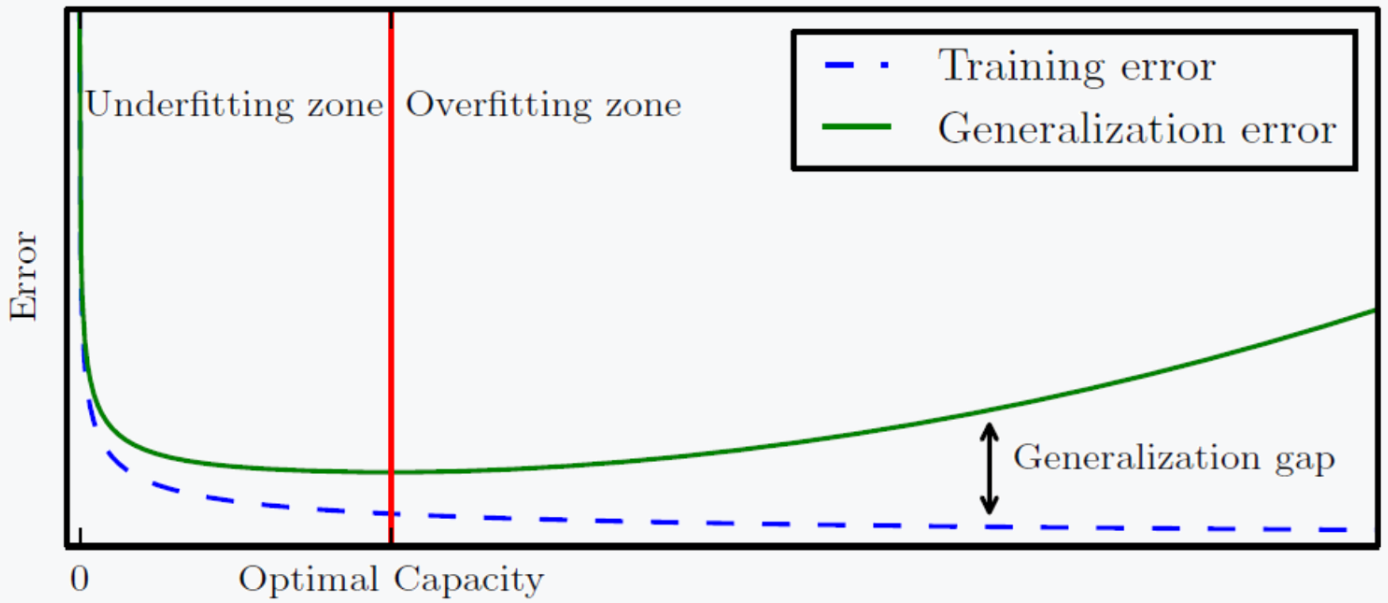
Wichtig: Es ist keine Betrachtung auf dem vorhandenen Datensatz, sondern eine Überlegung dazu was passiert, wenn neue Daten kommen, d. h. welchen Fehler man dann im Durchschnitt macht.

(Gareth, 2021, 33)

Overfitting and Underfitting

Konsequenzen für das Training von MLP

Für das Trainieren neuronaler Netze beobachtet man deswegen den MSE auf den Trainings- und sogenannten Validierungs- oder Generalisierungsdaten.

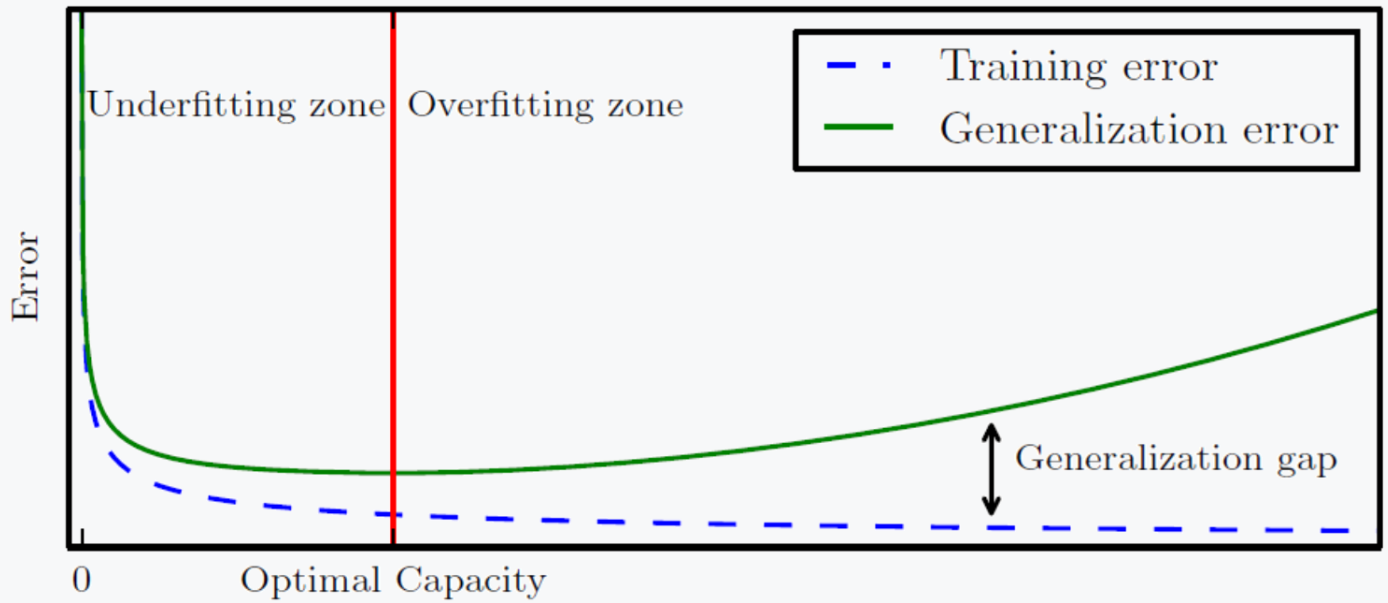


Sobald die grüne Kurve ansteigt vermutet man ein Auswendiglernen von alten Störtermrealisationen. Da insbesondere zukünftige Ziehungen des Störterms unbekannt sind – wie sie sich in anderen Daten niederschlagen – erhalten wir also eine Idee der Antwort zu: **Wie ändert sich $\hat{f}$ bei anderen Trainingsdaten?**

(Goodfellow, 2016, 127, Fig. 5.6)

Overfitting and Underfitting

Konsequenzen für das Training von MLP



Die blau gestrichelte Kurve zeigt, wie der **Fehler durch Modellvereinfachung** durch die schrittweise verbesserte Bestimmung der Gewichte gemäß Back Propagation reduziert wird. Da sich in den Trainingsdaten nur ein durch den Störterm verrauschter Zusammenhang y(x) befindet, ist eine zu gute Anpassung an die Trainingsdaten gefährlich. Es gilt auch hier: **History doesn’t repeat itself, but it rhymes.**

(Goodfellow, 2016, 127, Fig. 5.6)