Machine Learning für das KMU

Martin Sterchi

2024-04-01

Contents

Über das Buch

Die Motivation für dieses Buch kam aus der Erkenntnis, dass viele kleine und mittelgrosse Unternehmen (KMU) in der Schweiz zwar über grosse Datenmengen verfügen, aber nicht das nötige Knowhow haben, um die Daten zu analysieren und für die Optimierung von Entscheidungsprozessen zu nutzen. Mit diesem Buch möchte ich einen kleinen Beitrag leisten, den Knowhow Transfer von Fachhochschulen in die Unternehmen zu katalysieren.

Das Buch versucht, sowohl die klassischen Machine Learning Methoden als auch neueste Entwicklungen im Deep Learning mit einem Fokus auf die Anwendung zu vermitteln. Deep Learning kann als eine Teilmenge des Machine Learnings gesehen werden. Das heisst, jede Deep Learning Methode ist automatisch auch eine Machine Learning Methode. Machine Learning entält jedoch weitere Methoden, welche nicht dem Deep Learning zugeordnet werden können. Das Gebiet Machine Learning ist wiederum eine Teilmenge der Methoden der Künstlichen Intelligenz. Letztere enthält weitere Methoden, welche nicht dem Machine Learning zuzuordnen sind. Abbildung 1 versucht diesen Sachverhalt schematisch darzustellen.

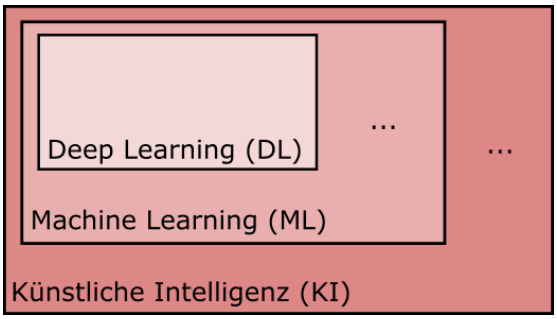


Figure 1: Unterscheidung zwischen KI, ML und DL.

Wir werden im ganzen Buch die folgenden (üblichen) Abkürzungen verwenden:

- Künstliche Intelligenz = KI (oft spricht man auch von AI, was die Abkürzung für den englischen Begriff *Artificial Intelligence* ist).

- Machine Learning = ML
- Deep Learning = DL

Obwohl das Buch einen anwendungsorientierten Ansatz verfolgt, soll die mathematisch-statistische Intuition hinter den beschriebenen Modellen und Methoden nicht zu kurz kommen. Diese Intuition ist aus meiner Sicht zwingend, um beurteilen zu können, ob sich ein Modell überhaupt für ein gegebenes Problem eignet. Am Schluss geht es nämlich darum, dass wir mit dem Einsatz von Machine Learning einen Mehrwert für ein Unternehmen oder für die Gesellschaft schaffen können. Das erfordert, dass wir uns eingehend und kritisch mit den Modellen und deren Eignung für ein gegebenes Problem auseinander setzen.

Zielgruppe

Das Buch richtet sich insbesondere an Fachhochschulstudierende in der deutschsprachigen Schweiz mit einem intrinsischen Interesse an quantitativen Methoden im Allgemeinen und Machine Learning im Besonderen. Vorausgesetzt werden Mathematikkenntnisse auf Stufe Mittelschule (Berufs- oder gymnasiale Matur), d.h. Sie sollten vertraut sein mit den Grundlagen bezüglich mathematischer Funktionen, der Integral- und Differentialrechnung sowie den wichtigsten Resultaten aus der Algebra. Ausserdem gehe ich davon aus, dass Sie bereits eine Einführung in das Thema Statistik besucht haben und Konzepte aus der deskriptiven Statistik (Mittelwert, Median, Varianz, Quantile, etc.) sowie aus der Inferenzstatistik (Verteilungen, statistisches Testen, etc.) bekannt sind.

Bevor Sie sich aber nun Sorgen machen: Kapitel 2 enthält eine Einführung in die wichtigsten Mathematik- und Statistikgrundlagen, die nötig sind für das Verständnis von Machine Learning Modellen.

Da ich mit diesem Buch einen anwendungsorientierten Ansatz verfolge, werden wir auch in das Programmieren einsteigen. Dazu verwenden wir in diesem Buch die Programmiersprache `R`. Es werden keine Vorkenntnisse vorausgesetzt. Kapitel 3 enthält eine kurze Einführung in die Programmiersprache `R` und verweist Sie auf weiterführende Ressourcen zum Thema Programmieren. Jedes Modell, das wir uns anschauen werden, ist mit R-Code dokumentiert, so dass Sie lernen, wie die Modelle in der Praxis angewendet werden können.

Aufbau des Buchs

Das Buch enthält folgende Kapitel:

- Kapitel 1: Einführung in das Thema Machine Learning mit **Definitionen** sowie Anwendungsbeispielen.
- Kapitel 2: Wichtigste **Mathematik- und Statistikgrundlagen**, die für das Verständnis der Modelle in den späteren Kapitel elementar sind.
- Kapitel 3: Einführung in das **Programmieren** mit `R` sowie Überblick über die wichtigsten `R`-Packages, die wir verwenden werden.
- Kapitel 4: Hier erlernen wir die Grundmodelle, um **Regressionsprobleme** zu lösen. Es sind lineare Modelle, was bedeutet, dass die funktionale Form der Modelle linear von den Parametern des Modells abhängen. Grafisch bedeutet dies, dass ein solches Modell im einfachsten Fall durch eine Gerade beschrieben werden kann.
- Kapitel 5: In diesem Kapitel lernen wir die Grundmodelle für das **Klassifikationsproblem** kennen. Diese Modelle führen typischerweise zu einer linearen Entscheidungsgrenze (engl. *Decision Boundary*) zwischen den verschiedenen Klassen, die wir unterscheiden oder klassifizieren wollen.
- Kapitel 6: Damit wir ML in der Praxis anwenden können, lernen wir hier die typische **ML-Pipeline** kennen. Sie werden die Techniken und Methoden kennen lernen, die es braucht, um überhaupt erst an den Punkt zu kommen, um ein ML-Modell rechnen zu können. Oft werden diese Techniken und Methoden unter dem Begriff Preprocessing der Daten zusammengefasst. Doch die Pipeline endet nicht mit dem Rechnen eines ML-Modells. Danach muss ein Modell evaluiert werden und wenn Sie als Analyst*in zufrieden sind, müssen Sie sich Gedanken machen, wie das Deployment des Modells aussehen soll. Das heisst, wie kann Ihr Modell Dritten zur Verfügung gestellt werden? Wir werden uns hier auch kurz mit den wichtigsten Techniken aus dem Unsupervised Learning befassen.
- Kapitel 7: Nach den ersten linearen Modellen für das Regressions- und Klassifikationsproblem lernen wir hier ein flexibleres Modell kennen, nämlich den **Entscheidungsbaum** (engl. *Decision Tree*). Entscheidungsbäume eignen sich sowohl für das Regressions- als auch für das Klassifikationsproblem. Obwohl sie in realen Projekten typischerweise anderen Modellen unterlegen sind, wenn es um die Vorhersagequalität geht, sind sie trotzdem attraktive Modelle, da sie gut visualisierbar sind.
- Kapitel 8: Aufbauend auf den Entscheidungsbäumen aus dem vorherigen Kapitel können sehr mächtige Modelle erstellt werden, die in der Praxis oft die besten Vorhersagen liefern. Weil es sich dabei üblicherweise um eine clevere Aggregierung der Resultate einer grossen Anzahl individueller Entscheidungsbäume handelt, werden diese Modelle **Ensembles** genannt. Wie die individuellen Entscheidungsbäume eignen sich Ensembles sowohl für das Regressions- als auch für das Klassifikationsproblem.
- Kapitel 9: Ein weiteres mächtiges Modell, das sich sowohl für das Regressions- als auch für das Klassifikationsproblem eignet, sind die **Support Vector Machines**. Ihre Popularität ist mit dem Aufstieg von Deep Learning etwas verblasst. Es lohnt sich aber immer noch allemal, diese Familie von Modellen kennen zu lernen, insbesondere auch weil sie nicht als Blackbox-Modelle gelten und theoretisch gut fundiert sind.

- Kapitel 10: Ab diesem Kapitel steigen wir in das Thema Deep Learning ein. Sie werden die Architektur von einfachen **Articial Neural Networks** (ANNs) kennen lernen. Ausserdem schauen wir uns in diesem Kapitel den genialen Backpropagation Algorithmus anhand eines einfachen linearen Regressionsproblems an. Dieser Algorithmus ist der Schlüssel für die viel diskutierten Fortschritte im Bereich der künstlichen Intelligenz, weil er das Trainieren von riesigen Modellen überhaupt erst möglich macht.
- Kapitel 11: Hier lernen wir sogenannte **Convolutional Neural Networks** (CNNs) kennen. Sie sind die Basis für die Fortschritte auf dem Gebiet Computer Vision und erlauben beispielsweise Anwendungen im Bereich automatische Gesichtserkennung in Bildern oder Videos.
- Kapitel 12: Nach ANNs und CNNs lernen wir hier **Recurrent Neural Networks** (RNNs) kennen. Diese Modelle bilden die Basis für Probleme, in denen die Daten als Sequenzen vorliegen. Das können einache Zeitreihen (z.B. Börsenkurse) sein, aber auch komplexere Sequenzdaten wie beispielsweise geschriebene oder gesprochene Sprache oder Tonaufnahmen.
- Kapitel 13: In diesem letzten Kapitel geht es schliesslich um **Generative KI**. Wir beschäftigen uns hier also mit Modellen, die nicht nur einfach ein Vorhersageprobleme lösen können, sondern auch neue Inhalte (z.B. Texte, Musik, Bilder) generieren können. Abbildung 2 enthält als Beispiel den Output einer generativen Software, die basierend auf einem Prompt ein Bild erstellt. Nach dem Lesen dieses Kapitels sollten Sie ein grundlegendes Verständnis für die Funktionsweise von Modellen wie Chat-GPT haben.

Weiterführende Literatur

Ein grosser Teil des vorliegenden Buchs baut auf bestehenden Büchern zum Thema Machine Learning auf. Ich werde im Buch immer wieder auf die Quellen verweisen. Die wichtigsten Referenzen für dieses Buch sind folgende:

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2021). An Introduction to Statistical Learning: with Applications in R. New York: Springer. 2nd Edition.
- Aurélien Géron. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Sebastopol: O'Reilly Media Inc. 3rd Edition.
- Christopher M. Bishop. (2006). Pattern Recognition and Machine Learning. Berlin, Heidelberg: Springer.
- Kevin P. Murphy. (2012). Machine Learning A Probabilistic Perspective. The MIT Press.

Die ersten beiden Referenzen sind einführende Texte und können parallel zum vorliegenden Buch gelesen werden. Die letzten zwei Referenzen sind fortgeschrit-

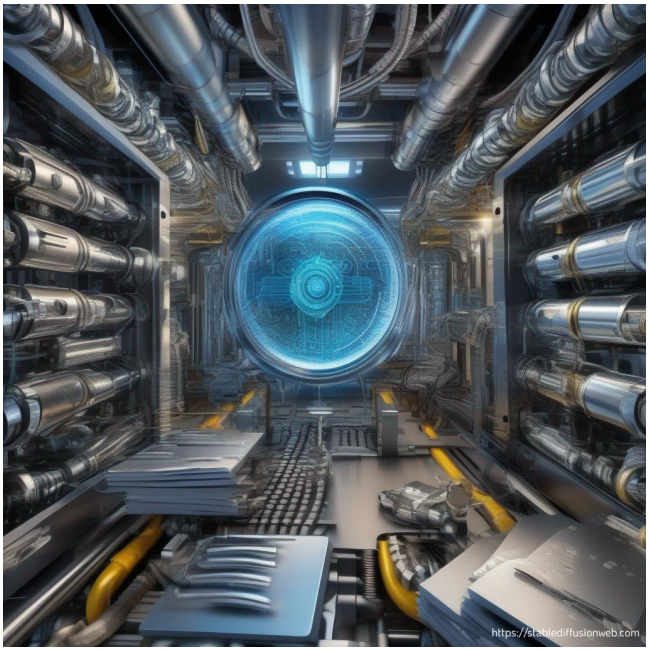


Figure 2: Beispielsoutput einer generativen Bildgenerierungssoftware (https://stablediffusionweb.com/) basierend auf dem Prompt "A title image for a textbook about Machine Learning targeting small and medium companies."

tene Texte und ich empfehle, sie erst nach dem vollständigen Verständnis des vorliegenden Buchs oder der ersten beiden Referenzen zu lesen.

Lizenz

Kontakt

Für Fragen und Anregungen zum Buch stehe ich gerne zur Verfügung:

Martin Sterchi
Riggenbachstrasse 16
4600 Olten
martin.sterchi@fhnw.ch

Chapter 1

Einführung

In diesem Kapitel geht es darum zu verstehen, was ML überhaupt ist, warum es nützlich sein kann und was typische Anwendungsfälle von ML sind. Wir werden ausserdem verschiedene Unterkategorien von ML kennen lernen.

1.1 Was ist Machine Learning?

Im Prinzip geht die Geschichte des MLs weit zurück, nämlich zu den Anfängen der Statistik. Viele Modelle, die heutzutage im ML angewendet werden sind nämlich eigentlich von Statistiker*innen erfundene Modelle. Die Geschichte des MLs und der Statistik sind darum eng verknüpft. Einen eigentlichen Startpunkt des MLs könnte man vielleicht in den 1960er Jahren ausmachen, mit den Arbeiten von Frank Rosenblatt[1], welcher das sogenannte **Perceptron** und einen dazugehörigen Lernalgorithmus prägte (dazu später mehr). Danach blieb es aber rund 20 Jahre relativ ruhig bis die Forschung im Bereich Machine Learning so richtig Fahrt aufnahm. Ein grosser Schub für die Entwicklung von ML ging vom Aufkommen von extrem grossen Datenmengen (**Big Data**) und dem Internet aus. Das führte nämlich dazu, dass sich immer mehr Leute aus den Fachbereichen Informatik und Computer Science mit dem Thema ML befassten und effiziente Hard- und Software sowie algorithmische Kniffs und Tricks beisteuerten. Ausserdem ermöglichte das Internet den Zugang zu gewaltigen Datenmengen an Bildern, Videos, Klicks, etc. - denken Sie beispielsweise nur schon an die Informationen, die jede*r von uns tagtäglich im Internet hinterlässt. Ein weiterer Schub für das Machine Learning war (und ist) zudem die immer besser werdende Rechenleistung von Computern. Diese Entwicklungen haben sich im November 2022 kulminiert in der erstmaligen breiten öffentlichen Wahrnehmung von sogenannten **Large Language Models** wie ChatGPT.

[1] https://en.wikipedia.org/wiki/Frank_Rosenblatt

Wie der Name sagt, geht es im ML darum, dass eine Maschine (oder präziser, ein Computer) aus einem gegebenen Datensatz automatisch Muster lernt, ohne dass ein Mensch dem Computer (explizit) sagen muss, was er lernen soll. Der Mensch gibt jedoch dem Computer die Rahmenbedingungen für das selbständige Lernen vor. Die erlernten Muster sind selbstverständlich nur nützlich, wenn sie **genereller Natur** sind und auch für neue bzw. zukünftige Beobachtungen gelten. Beispiel: ein Spital hat während der Corona Pandemie ein Modell trainiert, um den täglichen Pflegebedarf je nach Wochentag, Saison, und weiteren Indikatoren vorherzusagen. Das Modell funktioniert nun nach der Pandemie aber nicht wunschgemäss und prognostiziert in der Tendenz einen zu hohen Pflegebedarf. Das Problem ist, dass die erlernten Muster nicht gut auf eine Zeit nach der Pandemie generalisierbar sind. Mit anderen Worten: die Trainingsdaten waren nicht repräsentativ genug. ML-Modelle sollen also generell gültige Muster in den Daten erlernen.

Bevor wir etwas konkreter anschauen, wie genau ein Computer selbständig aus Daten lernen kann, schauen wir uns die Definitionen von zwei Experten im Gebiet ML an:

"[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed." Arthur Samuel, 1959

"Machine Learning is the science (and art) of programming computers so they can learn from data." Aurélien Géron[2]

Zusammenfassend lässt sich sagen, dass wir mit ML dem Computer die Möglichkeit geben, automatisch und selbständig aus Daten zu lernen. Nichtsdestotrotz braucht es Sie als ML-Expert*in, und zwar wie folgt:

1. Sie entscheiden sich für ein spezifisches ML Modell. Typischerweise kann ein ML Modell durch eine mathematische Funktion (siehe Kapitel 2) charakterisiert werden. ML Modelle können unterschiedlich flexibel sein und es liegt im Ermessen von Ihnen, wie flexibel das Modell sein soll. Sie müssen bei der Wahl des Modells die Komplexität des Problems berücksichtigen. Grundsätzliche gilt bei der Wahl des Modells, dass flexiblere Modelle komplexere Sachverhalte abbilden können. Ein zu flexibles Modell kann aber zu Overfitting führen, aber dazu später mehr. Dieser Schritt wird im Fachjargon typischerweise **Modellwahl** (engl. *Model Selection*) genannt.
2. Sobald Sie das Modell ausgewählt haben, übergeben Sie dem Computer (etwas vereinfacht gesagt) das Modell, einen Datensatz sowie einen Lernalgorithmus. Nun hat der Computer alle Zutaten, um automatisch zu lernen. Doch was lernt er eigentlich? Der Computer lernt die Parameter Ihres gewählten Modells, so dass das Modell sich optimal an die Daten an-

[2] Aurélien Géron. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Sebastopol: O'Reilly Media Inc. 3rd Edition.

passt. Dieser Schritt wird im Fachjargon **Modelltraining** (engl. *Model Training*) genannt.
3. Falls Sie mit dem erlernten Modell zufrieden sind, können Sie es nun entweder dazu verwenden Vorhersagen zu machen oder um Zusammenhänge in den Daten zu interpretieren und daraus wertvolle Einsichten gewinnen. Dieser Schritt wird im Fachjargon als **Modellinferenz** (engl. *Model Inference*) zusammengefasst. Typischerweise sind Sie in der Realität mit dem ersten erlernten Modell allerdings noch nicht zufrieden und gehen zurück zu Schritt 1 und wählen ein anderes Modell.

Es handelt sich bei dieser Vorgehensweise um eine sehr allgemeine Beschreibung des Machine Learning Prozesses. Wie diese drei Schritte konkret funktionieren, werden Sie in den nachfolgenden Kapiteln dieses Buchs erfahren.

1.2 Wann macht es Sinn ML einzusetzen?

Ein ML Modell zu trainieren kann viel Zeit und Geld kosten. Zum Beispiel müssen Sie unter Umständen überhaupt erst die Daten sammeln (oder von einem Datendienstleister kaufen), um ein Modell zu trainieren. Oder das Projekt ist so komplex, dass Sie als Analyst*in unzählige Stunden benötigen, um die Daten überhaupt erst in eine Form zu bringen, die es erlaubt ein Modell zu trainieren. Für neuartige DL Modelle oder Generative KI kann das Trainieren bzw. Lernen eines Modells durch den reinen Stromverbrauch bzw. die vom Cloud-Betreiber in Rechnung gestellten Kosten so hoch sein, dass sich Ihr ursprüngliches Vorhaben nicht mehr lohnt. Es ist also ungemein wichtig, dass Sie sich vor Projektbeginn gut überlegen, ob ML für Ihr vorliegendes Problem überhaupt Sinn macht und einen Mehrwert generieren kann.

Folgende Daumenregeln[3] können Ihnen dabei helfen, zu entscheiden, ob ML für Ihr Projekt Sinn macht:

- Ihr Problem entspricht einem Standard ML-Problem, das bereits mehrfach gelöst wurde und für das es sogenannte “off-the-shelf” Lösungen gibt. Beispiel: Sie wollen das Sentiment (positive vs. negative Grundhaltung) von Social Media Posts über Ihr Unternehmen automatisch klassifizieren. Dazu gibt es viele vortrainierte Modelle, die teilweise gratis verwendet werden können.
- Der manuelle Arbeitsaufwand ist sehr gross, wenn das Problem durch Menschen gelöst werden soll. Das Problem ist aber ansonsten klar strukturiert und benötigt keinen grossen kognitiven Einsatz eines Menschen. Beispiel: In den Post-Verteilzentren werden die von Hand geschriebenen

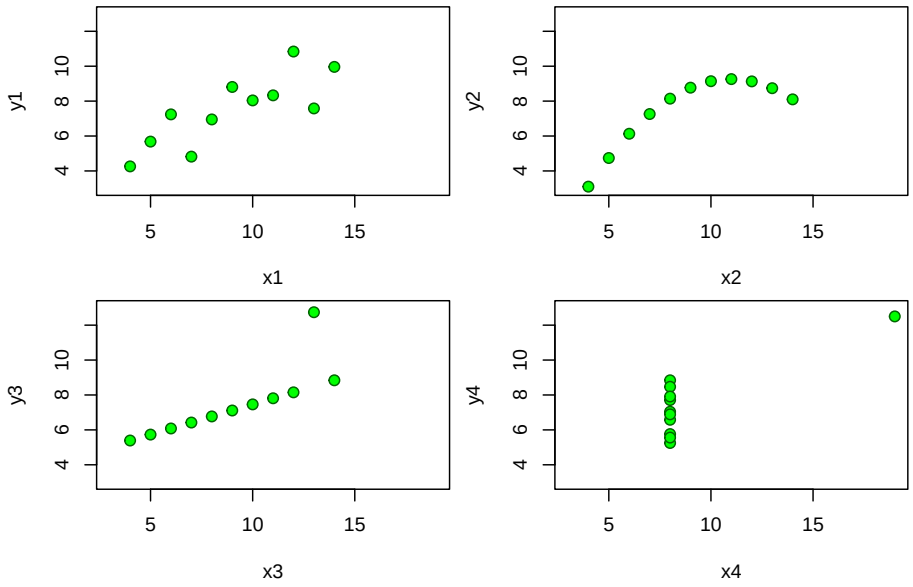
[3] siehe auch Seiten 6 - 7 in Aurélien Géron. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Sebastopol: O’Reilly Media Inc. 3rd Edition.

Postleitzahlen (PLZ) problemlos mittels Computer bzw. ML Modellen erkannt und "gelesen" und die Briefe und Pakete entsprechend sortiert.

- Komplexe Probleme, in denen ein Mensch keinen Überblick hat, weil so grosse und komplexe Datenmengen vorhanden sind. Wir Menschen haben grosse Mühe damit, in Rohdaten (reinen Datentabellen) irgendwelche Muster zu erkennen. In diesem Fall können wir entweder versuchen, die Daten zu visualisieren oder mithilfe von ML Zusammenhänge zu lernen, die wir sonst nicht erkennen könnten. Ein illustratives Beispiel ist das Anscombe Quartett[4], das vier kleine Stichproben mit jeweils elf Datenpunkten enthält. Jeder Datenpunkt wird durch eine x und eine y Variable beschrieben. Die vier x- sowie die vier y-Variablen haben identische Mittelwerte. Erst eine einfache Visualisierung der vier Stichproben mithilfe eines Streudiagramms zeigt die Muster sowie die Unterschiede zwischen den vier Stichproben deutlich auf.

```
#>    x1 x2 x3 x4    y1   y2    y3    y4
#> 1  10 10 10  8  8.04 9.14  7.46  6.58
#> 2   8  8  8  8  6.95 8.14  6.77  5.76
#> 3  13 13 13  8  7.58 8.74 12.74  7.71
#> 4   9  9  9  8  8.81 8.77  7.11  8.84
#> 5  11 11 11  8  8.33 9.26  7.81  8.47
#> 6  14 14 14  8  9.96 8.10  8.84  7.04
#> 7   6  6  6  8  7.24 6.13  6.08  5.25
#> 8   4  4  4 19  4.26 3.10  5.39 12.50
#> 9  12 12 12  8 10.84 9.13  8.15  5.56
#> 10  7  7  7  8  4.82 7.26  6.42  7.91
#> 11  5  5  5  8  5.68 4.74  5.73  6.89
```

[4] https://de.wikipedia.org/wiki/Anscombe-Quartett



1.3 Anwendungsfälle von ML

In diesem Abschnitt stelle ich erfolgreiche Anwendungsfälle von ML vor. Einige davon treffen Sie womöglich tagtäglich in Ihrem Alltag an:

- **Spam Filter** sind ein frühes Beispiel einer erfolgreichen Anwendung von ML. Ein Klassifikationsmodell entscheidet dabei automatisch aufgrund der Inhalte einer Email, des Betreffs sowie des Absenders, ob es sich um eine Spam oder eine sogenannte Ham Email (unproblematische Email) handelt. Falls Sie gängige Email Software verwenden, dann arbeitet im Hintergrund ein Spam Filter daran, Sie vor lästigen Emails zu schützen.
- Ein grosser Teil des wirtschaftlichen Erfolgs von **Google** basiert auf der Idee, dass aufgrund der Suchhistorie hervorgesagt werden kann, welche Nutzerin oder welcher Nutzer mit welcher Wahrscheinlichkeit eine bestimmte Werbung anklickt. Dies erlaubt Google für jede Nutzer*in die Werbung mit den höchsten "Erfolgschancen" zu schalten. Da jeder Klick Einnahmen generiert, ist es für das Geschäftsmodell von Google entscheidend, dass möglichst viele Klicks stattfinden.
- Ein grosser Bereich des MLs und speziell des DLs befasst sich mit **Computer Vision**. Dabei geht es darum, das Hauptmotiv von Bildern zu klassifizieren (z.B. Zeigt ein Bild ein Tier oder einen Menschen?), Objekte in Bildern zu entdecken (z.B. Enthält das Bild eine Person?) und das entdeckte Objekt dann auch zu klassifizieren (z.B. Handelt es sich bei

der Person um XY?). Als konkreteres Beispiel können Sie sich einen Industriebetrieb vorstellen, welcher ein Computer Vision Modell einsetzen möchte, um den Abnützungsgrad der von ihnen produzierten Werkzeuge automatisch zu erkennen und den Kundinnen und Kunden den optimalen Ersatzzeitpunkt für das Werkzeug vorhersagen zu können.

- Ähnlich wie im vorherigen Beispiel gibt es bereits viele Anwendungen im öffentlichen Verkehr, in denen es um **Predictive Maintenance** geht. Z.B. kann der optimale Wartungszeitpunkt für eine Weiche oder einen Gleisabschnitt aufgrund einer Vielzahl an Indikatoren und Messungen vorhergesagt werden.
- Ein grosses Einsatzgebiet für ML ergibt sich im Finanzsektor durch das automatische Erkennen von potentiell **betrügerischen Transaktionen**. Falls Sie auch schon mal eine Kreditkartentransaktion direkt am Telefon einer Kundenberaterin oder einem Kundenberater bestätigen mussten, dann ist es wahrscheinlich, dass Ihre Transaktion von einem ML System zur manuellen Überprüfung geflaggt wurde. In diesem Zusammenhang spricht man manchmal auch vom Erkennen von Anomalien (engl. *Anomaly Detection*).
- Sogenannte **Recommender Systems** sind insbesondere in Online Verkaufspunkten von grossem Nutzen. Betreiben Sie beispielsweise einen grossen Onlinehandel, dann wollen Sie Ihren Kundinnen und Kunden Produkte zum Kauf vorschlagen. Dazu verwenden Sie ein Modell, das basierend auf der Ähnlichkeit zwischen Kundinnen und Kunden potentiell interessante Produkte vorschlägt.
- Die rasanten Entwicklungen im Bereich **Natural Language Processing** (NLP) in den letzten 10 Jahren haben viele neue und interessante Anwendungsgebiete zutage gefördert. Zum Beispiel eignen sich *Large Language Models* (LLMs) als erste Anlaufstelle für Kundinnen und Kunden (automatisierter Kundenservice). LLMs werden vermutlich aber auch immer mehr in internen Prozessen in Unternehmen eingesetzt, z.B. um komplexe Dokumente zusammenzufassen oder Sitzungsprotokolle zu erstellen.

Die obige Liste ist bei weitem nicht komplett und die Entwicklungen im Bereich ML sind aktuell so rasant, dass jeden Tag eine grosse Zahl von neuen ML-basierten Produkten und Dienstleistungen auf den Markt kommen.

1.4 Supervised vs. Unsupervised Learning

Den Unterschied zwischen dem Supervised Learning und dem Unsupervised Learning können wir am besten erklären, indem wir uns mit ein paar mathematischen Grundlagen des Machine Learnings befassen. Keine Sorge, diese Grundlagen sind sehr einfach, aber versuchen Sie, diese bereits gut zu verstehen, denn wir bauen später darauf auf.

Im **Supervised Learning** haben wir einerseits sogenannte Input-Daten und andererseits einen Output, den wir vorhersagen wollen. Für die Input-Daten gibt es ganz viele verschiedene Begriffe, die synonym verwendet werden: z.B. Features, unabhängige Variablen, Attribute, Prädiktoren. Dasselbe gilt für den Output, hier gibt es folgende Synonyme: Zielvariable, abhängige Variable, Label, oder auch einfach y. Unsere Konvention hier ist aber folgende: es gibt Input-Daten (oder Input-Variablen) und einen Output (oder Output-Variable).

Die Input-Daten für eine Beobachtung i schreiben wir mathematisch wie folgt:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix},$$

Diese Notation bedarf ein paar Erklärungen:

- Den Index i brauchen wir, um die verschiedenen Beobachtungen zu kennzeichnen. i kann eine Ganzzahl zwischen 1 und n annehmen, wobei n die Anzahl Beobachtungen im Datensatz bezeichnet. Wenn wir zum Beispiel etwas über die Input-Daten der dritten Beobachtung sagen wollen, dann können wir die Notation $\mathbf{x}_3$ verwenden.
- Für jede Beobachtung i haben wir insgesamt p Variablen, welche die verschiedenen Attribute einer Beobachtung enthalten. x_{i1} bezeichnet also die erste Variable der i-ten Beobachtung, x_{i2} die zweite Variable der i-ten Beobachtung und x_{ip} die p-te (letzte) Variable der i-ten Beobachtung.
- Was Sie oben sehen, ist aus mathematischer Sicht ein Spaltenvektor. Im Moment reicht es, wenn Sie wissen, dass wir mit diesem Spaltenvektor die Input-Daten einer Beobachtung *kompakt* darstellen können.

Neben den Input-Daten haben wir im Supervised Learning aber wie erwähnt auch einen Output und den bezeichnen wir üblicherweise mit y_i. Auch hier hilft uns der Index i dabei, die Beobachtungen eindeutig zu kennzeichnen. Schauen wir uns am besten kurz ein konkretes Beispiel an:

Aufgabe

```
#> Warning: `includeHTML()` was provided a `path` that appears to be a complete HTML document.
#> x Path: exercises/notation.html
#> i Use `tags$iframe()` to include an HTML document. You can either ensure `path` is accessible
```

Wichtig: Beim Supervised Learning geht es um ML Probleme, in denen sowohl Input-Daten als auch ein Output vorhanden ist. Ziel beim Supervised Learning ist es, ein Modell zu trainieren, das basierend auf den Input-Daten möglichst gute Vorhersagen für den Output macht. Es geht also hier um Vorhersageprobleme. In einem gewissen Sinn ist der Output die überwachende Instanz (engl. Supervisor), welche den Lernprozess des Modells kontrolliert.

Im Gegensatz zum Supervised Learning haben wir im **Unsupervised Learning** nur Input-Daten und *keinen Output*. Im Unsupervised Learning geht es darum, aus den Input-Daten interessante Muster zu lernen, welche für bessere unternehmerische Entscheidungen verwendet werden können. Ein einfaches Beispiel ist das Clustering von Kundinnen und Kunden eines Unternehmens in ähnliche Kundengruppen, so dass die verschiedenen Kundengruppen gezielter mit Marketingaktionen angesprochen werden können. Techniken, um komplexe Datensätze zu visualisieren, werden typischerweise auch zum Unsupervised Learning gezählt.

Neben dem Supervised und dem Unsupervised Learning gibt es noch eine dritte Kategorie von Machine Learning, nämlich das **Reinforcement Learning** (RL). Dieser Kategorie gehören Modelle an, die (virtuelle) Agenten so trainieren, dass sie langfristig möglichst optimal handeln. Das bekannteste Beispiel aus dem RL ist Googles AlphaGo Agent, welcher den menschlichen Go Weltmeister im Jahr 2017 schlug.[5]. Reinforcement Learning ist aber auch eine wichtige Komponente in der Optimierung von grossen Sprachmodellen wie ChatGPT. In einer ersten Fassung dieses Buchs werden wir uns nicht (oder nur am Rande) mit RL befassen.

Die Unterscheidung zwischen den drei Arten von Machine Learning ist im oberen Teil der Abbildung 1.1 visualisiert:

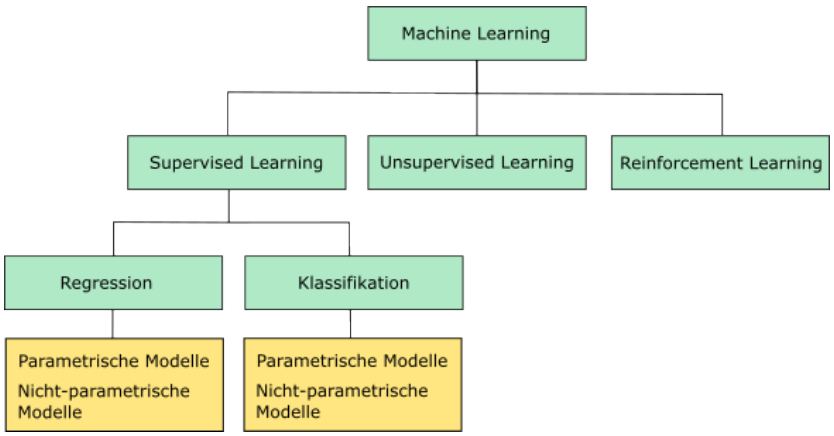


Figure 1.1: Die verschiedenen Kategorien des Machine Learnings und deren Hierarchie.

1.5 Regression vs. Klassifikation

In der Kategorie des Supervised Learnings unterscheiden wir weiter zwischen Regressions- und Klassifikationsproblemen (siehe auch Abbildung 1.1).

[5] https://deepmind.google/technologies/alphago/

Im Regressionsproblem ist der Output eine **stetige** Variable (Intervall- oder Verhältnisskalierung), d.h. die Variable enthält reelle (numerische) Werte. Mathematisch schreibt man dies als $y_i \in \mathbb{R}$, wobei $\mathbb{R}$ die Menge der reellen Zahlen beschreibt.

Im Klassifikationsproblem ist der Output bzw. die Zielvariable eine **kategorische** Variable (Nominal- oder Ordinalskalierung). Mathematisch schreibt man dies als $y_i \in \{1, ..., C\}$, wobei C die Anzahl Kategorien beschreibt. Wenn wir nur $C = 2$ Kategorien haben wie im Beispiel oben mit $y_i \in \{\text{Betrug}, \text{kein Betrug}\}$ sprechen wir von einem binären Klassifikationsproblem. Falls $C > 2$ sprechen wir vom mehrklassigen (engl. *multiclass*) Klassifikationsproblem.

Aufgabe

```
#> Warning: `includeHTML()` was provided a `path` that appears to be a complete HTML document.
#> x Path: exercises/regvsclass.html
#> i Use `tags$iframe()` to include an HTML document. You can either ensure `path` is accessible
```

1.6 Parametrische vs. nicht-parametrische Modelle

Ein ML Modell gehört entweder der Familie **parametrischer** Modelle oder der Familie **nicht-parametrischer** Modelle an. Dabei spielt es keine Rolle, ob wir mit dem Modell ein Regressions- oder ein Klassifikationsproblem lösen wollen.

Womöglich sind Sie in Ihrer Ausbildung bereits **parametrischen Modellen** begegnet, denn das einfache lineare Regressionsmodell ist ein typisches Beispiel für ein parametrisches ML Modell. Das Modell ist vollkommen charakterisiert durch die beiden lernbaren (optimierbaren) Parameter w_0 und w_1[6] und kann wie folgt (mathematisch) aufgeschrieben werden:

$$\widehat{y_i} = f(x_i) = w_0 + w_1 \cdot x_i$$

Wenn Ihnen der obige Ausdruck noch fremd vorkommt, dann ist das nicht schlimm. Wir werden im Kapitel 4 ausführlich auf lineare Regressionsmodelle eingehen. Im Moment müssen Sie nur wissen, dass ein parametrisches Modell wie oben mit einer mathematischen Funktion beschrieben werden kann und dass diese Funktion durch lernbare **Parameter** (hier w_0 und w_1) charakterisiert wird.

Nicht-parametrische Modelle wiederum sind Modelle, welche nicht (oder zumindest nicht explizit) durch Parameter charakterisiert sind. Am besten

[6] In Statistikvorlesungen werden die beiden Parameter oft eher mit b_0 und b_1 oder mit β_0 und β_1 bezeichnet. Im Machine Learning nennt man Parameter oft Gewichte (engl. Weights), weshalb die Parameter typischerweise mit w bezeichnet werden.

schauen wir uns gleich ein einfaches nicht-parametrisches Modell an, nämlich das **K-Nearest-Neighbors** (KNN) Modell. Stellen Sie sich vor, Sie haben einen Datensatz mit 55 Produkten aus Ihrem Sortiment. Sie haben jedes dieser 55 Produkte auf Instagram und auf Tiktok durch Influencer*innen bewerben lassen. Für jedes der 55 Produkte hatten Sie ein Werbebudget für Instagram (x_{i1}) und ein Werbebudget für Tiktok (x_{i2}). Am Ende des Geschäftsjahrs haben Sie für jedes der 55 Produkte bestimmt, ob die Absatzziele erreicht wurden oder nicht (Output y_i). Die erfolgreichen Produkte (= Absatzziel erreicht) sind in untenstehender App als blaue Punkte eingezeichnet. Die roten Dreiecke repräsentieren die nicht-erfolgreichen Produkte. Sie sehen, dass erfolgreiche Produkte tendenziell höhere Instagram und Tiktok Werbebudgets aufwiesen als nicht-erfolgreiche Produkte. Sie möchten nun ein Modell schätzen, dass die Produkte automatisch klassifizieren kann. Dazu verwenden Sie das KNN Modell, das die K nächsten Nachbarn unter den 55 gegebenen Produkten sucht und dann die häufigste Beobachtung unter den K nächsten Nachbarn vorhersagt. In anderen Worten: wir suchen die K **ähnlichsten** Beobachtungen und nutzen diese, um eine Vorhersage zu machen.

Selbstverständlich spielt der konkrete Wert von K hier eine grosse Rolle - sollen wir nur $K = 1$ Nachbarn berücksichtigen? Oder $K = 10$ Nachbarn? Die erste Abbildung in der App zeigt nicht nur die 55 Datenpunkte, sondern auch die **Entscheidungsgrenze** (in schwarz). Untersuchen Sie kurz, wie sich diese Entscheidungsgrenze verändert, wenn Sie K erhöhen oder reduzieren.

Ausserdem können Sie in der ersten Abbildung auch den schwarzen Punkt mit der Maus setzen, wodurch Ihnen die K nächsten Punkte des schwarzen Punkts angezeigt werden.

Die zweite Abbildung zeigt die Entscheidungsregionen mit unterschiedlicher Intensität je nachdem wie sicher sich das Modell ist. In einer Region, in der alle K Nachbarn nicht-erfolgreiche Produkte sind, sind wir uns eher sicher bezüglich der Vorhersage als in einer Region, in der die Anteile zwischen erfolgreichen und nicht-erfolgreichen Produkten ausgeglichen sind.

Um die K nächsten Nachbarn zu finden, müssen wir die Distanzen zwischen Punkten rechnen können. Dazu verwenden wir die Euklidische Distanz, welche wir in Kapitel 2 kennen lernen werden.

Das KNN Modell ist ein sehr einfaches ML Modell, welches in der Praxis allerdings nicht allzu häufig angewendet wird. Warum nicht? Weil es am sogenannten **Fluch der Dimensionalität** (engl. Curse of Dimensionality) leidet. Doch was bedeutet das? Je mehr Input-Variablen wir haben, desto weiter entfernt sind Datenpunkte voneinander (das ist etwas, das man sich nur schwer vorstellen kann, aber Sie können es mir für den Moment einfach mal glauben). Das KNN beruht auf der Grundidee, dass wir K nahe, ähnliche Beobachtungen für die Vorhersage verwenden. Wenn diese K nahen Beobachtungen im hochdimensionalen Raum (= viele Input-Variablen) nicht mehr nahe sind, dann funktioniert auch das Modell nicht mehr gut.

Aufgaben

```
#> Warning: `includeHTML()` was provided a `path` that appears to be a complete HTML document.
#> x Path: exercises/knn.html
#> i Use `tags$iframe()` to include an HTML document. You can either ensure `path` is accessible
```

1.7 Machine Learning Pipeline

Abbildung 1.2 zeigt, wie eine typische ML-Pipeline aussieht.[7]

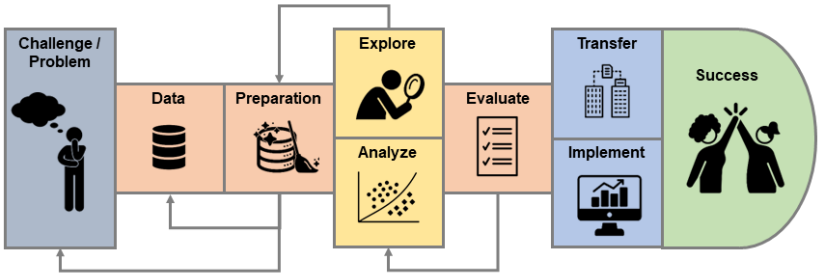


Figure 1.2: Eine typische ML-Pipeline.

Sie starten typischerweise mit einem **Problem** oder einer Herausforderung. Ihr ganzes Projekt sollte darauf ausgelegt sein, dieses Problem zu lösen. Es ist grundsätzlich nicht ratsam, auf Biegen und Brechen eine ML Lösung zu implementieren, wenn kein klar definiertes Problem vorliegt. Nehmen Sie sich also zu Beginn eines Projekts Zeit, das Problem grundlegend zu definieren. Sprechen Sie auch mit den entsprechenden Fachexpert*innen im Unternehmen, um genau zu verstehen, was verbessert oder effizienter gemacht werden soll und was die technischen oder ökonomischen Einschränkungen sind.

Sobald das Problemverständnis vorhanden ist, beginnen Sie, sich mit den **verfügbaren Daten** zu befassen. Auch hier müssen Sie sich wahrscheinlich mit den entsprechenden Expert*innen im Unternehmen (z.B. Datenbankadministrator*innen) austauschen. Es geht hier unter anderem darum abzuklären, welche Daten verfügbar sind, in welchem Format die Daten vorhanden sind wie die Datenqualität ist.

Danach beginnen Sie mit den Datenarbeiten. Häufig wird dieser Schritt **Preprocessing** oder **Data Cleaning** genannt. Oft verschlingt dieser Arbeitsschritt sehr viel Zeit und es ist nicht unüblich, dass 80% der Projektzeit hier aufgewendet werden. Es ist auch völlig normal, wenn Sie von diesem Schritt zurück zur Problemdefinition gehen und sie verfeinern oder anpassen müssen oder zum

[7] Icons stammen von https://thenounproject.com/.

Beispiel nochmals Fragen mit den Datenbankexpert*innen klären müssen, weil Ihr Datenverständnis noch nicht vollständig ist.

Nachdem die Daten vorbereitet wurden, gehen Sie typischerweise zu einer **explorativen Analyse** der Daten über. Das heisst, Sie visualisieren die vorhandenen Variablen univariat (d.h. jede Variable einzeln) oder multivariat (d.h. zwei oder mehr Variablen zusammen). Ein Beispiel einer univariaten Visualisierung ist ein Histogramm einer quantitativen Variable (z.B. Quartalsumsätze). Ein Beispiel einer multivariaten Visualisierung ist ein Streudiagramm zweier quantitativer Variablen (z.B. Quartalsumsätze und Wechselkurse). Auch hier ist es üblich, dass Sie einen Schritt zurück gehen und weitere Datenbereinigungen vornehmen müssen.

Nach der explorativen Analyse der Daten sollten Sie eine erste Idee von den wichtigsten Zusammenhängen in den Daten haben. Basierend darauf können Sie Ihr erstes Modell wählen und trainieren und mit der eigentlichen **Analyse** bzw. der Lösung des Problems beginnen.

Einer der wichtigsten Schritte ist die saubere und gründliche **Evaluation** Ihrer Modelle. Dieser Schritt dient einerseits dazu das beste Modell auszuwählen und andererseits dazu die Qualität Ihrer Lösung bzw. Ihres Modells abzuschätzen. Mit diesem zweiten Schritt wollen Sie nämlich bereits während der Projektphase einschätzen können, wie gut Ihr Modell das gegebene Problem löst oder einen bestehenden Betriebsprozess verbessert oder effizienter macht. Die beiden Schritte Analyse und Evaluation werden typischerweise ein paar Mal iteriert, bis Sie das beste Modell gefunden haben.

Am Schluss geht es darum, dass Sie Ihr Wissen und Ihre Erkenntnisse an die relevanten Fachexpert*innen weitergeben (**Wissenstransfer**) und Ihr finales Modell in einer produktiven Umgebung implementieren (oft **Deployment** genannt). Zum Beispiel können Sie Ihr Modell in einer mobilen App einbetten oder als REST API Service zur Verfügung stellen.

Chapter 2

Mathematik- und Statistik-Grundlagen

In diesem Kapitel repetieren wir die wichtigsten Grundlagen aus der Mathematik und Statistik, die es braucht, um Machine Learning Modelle zu verstehen. Das Thema *Lineare Algebra* wird für die meisten von Ihnen wahrscheinlich Neuland sein.

2.1 Funktionen

Eine Funktion, die wir in der Mathematik typischerweise mit f bezeichnen, ordnet jedem **Argument** x aus dem Definitionsbereich D (engl. *Domain*) **genau einen Wert** y aus dem Wertebereich W (engl. *Codomain*) zu. Oft sind D und W die Menge der reellen Zahlen, also $\mathbb{R}$. Die Menge der reellen Zahlen enthält alle möglichen Zahlen, die Sie sich vorstellen können.[1] Zum Beispiel die Zahlen 3, -4.247, $\sqrt{14}$, $5/8$, etc.

Wie eine Funktion grafisch aussieht, ist aus Panel (a) der Abbildung 2.1 ersichtlich. Hier zeigen wir die Form einer Funktion in einem kartesischen Koordinatensystem. Die Funktionskurve weist jedem Wert x auf der x-Achse genau einen Wert y auf der y-Achse zu. Der wichtigste Teil der oben aufgeführten Definition ist der Teil "genau einen Wert", denn eine Funktion kann einem Element x nicht zwei oder mehr Werte zuweisen, sondern nur genau einen. Genau aus diesem Grund handelt es sich bei Panel (b) in Abbildung 2.1 *nicht* um eine Funktion, da gewissen x-Werten mehrere Werte y zugeordnet werden. *Wichtig*: das heisst aber nicht, dass zwei verschiedenen x-Werten, nennen wir sie x' und x'', derselbe y-Wert zugeordnet werden kann (vgl. Panel (a)).

[1] Einzige Ausnahme sind die komplexen Zahlen.

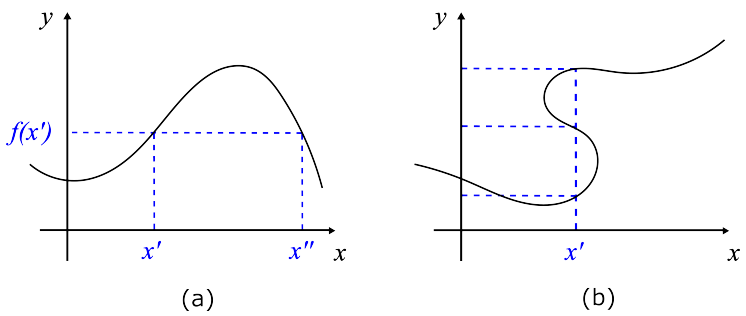


Figure 2.1: (a) Eine Funktion, die jedem x-Wert genau einen y-Wert zuweist. (b) Keine Funktion.

Mathematisch wird diese allgemeine Definition einer Funktion häufig wie folgt beschrieben:

$$f : x \mapsto y$$

Wir haben also eine Funktion f, die jedem Element x genau einen Wert y zuweist. Der Pfeil in obiger mathematischer Schreibweise beschreibt genau dieses Mapping. Wie genau dieses Mapping einem Argument x den entsprechenden y-Wert zuordnet, wird durch die Funktion $f(x)$ beschrieben. In den folgenden Abschnitten schauen wir uns typische Beispiele von Funktionen an, angefangen mit linearen Funktionen. Doch vorher wollen wir uns kurz überlegen, warum Funktionen für das Machine Learning überhaupt wichtig sind. Ein grosser Teil des Machine Learnings, der **Supervised Learning** genannt wird, befasst sich mit dem Problem, wie eine Zielvariable y mithilfe von einem oder mehreren Prädiktoren x vorhergesagt werden kann. Ein Machine Learning Modell ist darum nichts anderes als eine Funktion $y = f(x)$, die basierend auf den Prädiktoren x die Zielvariable y möglichst gut beschreiben kann.[2]

2.1.1 Lineare Funktionen

Nun schauen wir uns an, wie eine **lineare** Funktion aussieht. Eine lineare Funktion kann allgemein wie folgt geschrieben werden:

$$y = f(x) = a \cdot x + b$$

Obige Funktionsgleichung besagt, dass wir den entsprechenden y-Wert kriegen, indem wir den Wert des Arguments x mit a multiplizieren und danach eine Konstante b addieren. a und b sind die **Parameter** dieser Funktion. Die konkreten Zahlenwerte dieser beiden Parameter definieren, wie die Funktion am Schluss genau aussieht.

[2] Zumindest aus einer nicht-probabilistischen Perspektive.

Eine lineare Funktion hat auch eine geometrische Interpretation und zwar entspricht eine lineare Funktion einer Gerade. Das ist auch der Grund, warum wir diese Funktionen **linear** nennen, sie können graphisch durch eine "Linie" dargestellt werden. Der Parameter a ist die Steigung dieser Geraden und der Parameter b entspricht dem Ort, wo die Gerade die y-Achse schneidet (sogenannter y-Achsenabschnitt).

Am besten schauen wir uns ein paar konkrete Beispiele an (Abb. 2.2).

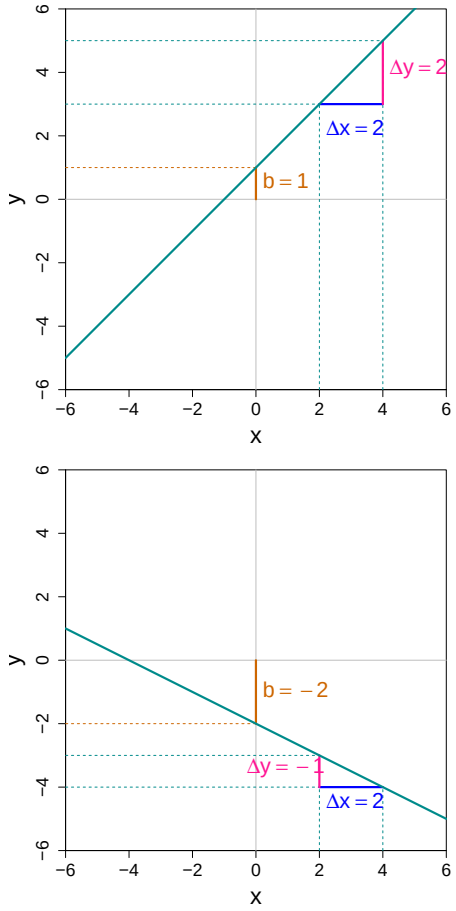


Figure 2.2: Beispiele linearer Funktionen.

Aus der linken Abbildung können wir ablesen, dass die Steigung dieser Geraden $\frac{\Delta y}{\Delta x} = \frac{2}{2} = 1$ ist und dass die Gerade die y-Achse am Ort 1 schneidet. Die entsprechende lineare Funktion kann dementsprechend als $y = x + 1$ geschrieben werden.[3]

[3]Wir müssen hier die Steigung 1 nicht explizit schreiben, aber selbstverständlich ist es nicht

Aus der rechten Abbildung können wir ablesen, dass die Steigung $\frac{\Delta y}{\Delta x} = \frac{-1}{2} = -0.5$ ist und dass die Gerade die y-Achse am Ort -2 schneidet. Die entsprechende lineare Funktion kann dementsprechend als $y = -0.5 \cdot x - 2$ geschrieben werden.

Es ist wichtig zu sehen, dass der Effekt einer Veränderung von x (also Δx) auf y überall derselbe ist. Es spielt also keine Rolle, ob wir von $x = -2$ zu $x = -1$ gehen oder von $x = 100$ zu $x = 101$, die entsprechende Veränderung in y (also Δy) wird dieselbe sein. Das muss so sein, denn die Gerade steigt (oder sinkt) mit konstanter Steigung.

Aufgaben

1. Zeichnen Sie die Funktion $y = 2 \cdot x$ in ein Koordinatensystem ein. Warum fehlt der Parameter b?
2. Zeichnen Sie die Funktion $y = -3$ in ein Koordinatensystem ein. Ist das überhaupt eine Funktion nach obiger Definition?

2.1.2 Quadratische Funktionen

Nun wollen wir uns eine etwas interessantere (und flexiblere) Familie von Funktionen anschauen, nämlich **quadratische** Funktionen. Auch hier wollen wir die Funktion erstmal allgemein aufschreiben:

$$y = f(x) = a \cdot x^2 + b \cdot x + c$$

Eine quadratische Funktion hat drei **Parameter**, nämlich a, b und c. Grafisch entspricht die quadratische Funktion einer **Parabel** (vgl. Abb. 2.3). Die Parameter sind hier nicht mehr so einfach grafisch zu interpretieren, aber die vier Beispiele in unten stehender Abbildung geben Anhaltspunkte, was passiert, wenn die Parameterwerte sich ändern.

Aufgaben

1. Sie haben folgende quadratische Gleichung: $y = 2 \cdot x^2 + x - 2$. Berechnen Sie mit der bekannten Lösungsformel $x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ die Orte auf der x-Achse, wo die Parabel die Achse schneidet (oder einfacher gesagt die Nullstellen).
2. Verwenden Sie folgenden R-Code, um beliebige quadratische Funktionen grafisch darzustellen, indem Sie die Parameterwerte auf der ersten Code-Zeile verändern.

falsch die lineare Funktion als $y = 1 \cdot x + 1$ zu schreiben.

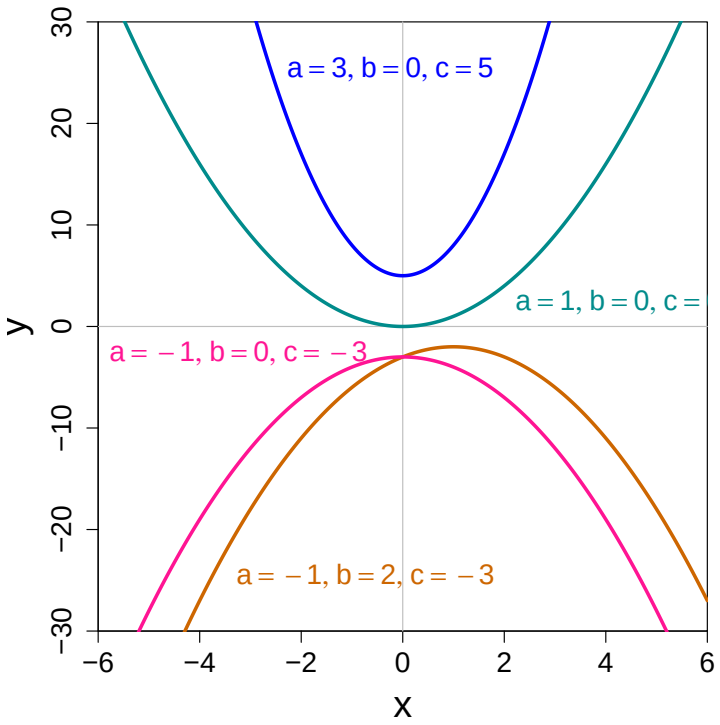


Figure 2.3: Beispiele quadratischer Funktionen.

```r
# Parameter setzen
a <- 2; b <- 0; c <- 1
# Quadratische Funktion
quad <- function(x, a, b, c) {a * x^2 + b * x + c}
# x-Werte
x <- seq(-6, 6, 0.01)
# y-Werte
y <- quad(x, a, b, c)
# Plot
plot(x, y, type = "l", lwd = 2, col = "darkcyan")
```

Sie wundern sich nun vielleicht, könnte man nicht auch eine Funktion antreffen, in der x^3, x^4, etc. vorkommen? Das ist selbstverständlich möglich. In diesem Fall spricht man dann von einem sogenannten **Polynom**. Die höchste Potenz des Arguments x definiert den Grad des Polynoms.

Schauen wir uns doch am besten gleich wieder ein Beispiel an:

$$y = f(x) = 1 \cdot x^4 - 2 \cdot x^3 - 5 \cdot x^2 + 8 \cdot x - 2$$

Die Visualisierung dieser Funktion ist in Abb. 2.4 gegeben. Diese Funktion ist nun bereits enorm flexibel und kann je nach Parameterwerten ganz unterschiedliche Zusammenhänge abbilden.

Aufgaben

1. Eine quadratische Funktion ist ein Polynom welchen Grades?
2. Handelt es sich bei der Funktion $y = 2x^5 + x + 1$ immer noch um ein Polynom? Falls ja, ein Polynom welchen Grades?
3. Handelt es sich bei der Funktion $y = x^{0.5} + 2$ um ein Polynom?

2.1.3 Funktionen mehrerer Argumente

Bisher haben wir nur Funktionen mit **einem Argument** x angeschaut, doch die meisten für das Machine Learning interessanten Funktionen sind Funktionen **mehrerer Argumente**.

Der Einfachheit halber schauen wir uns hier nur mal eine **lineare** Funktion zweier Argumente, nennen wir sie x_1 und x_2, an, denn diese können wir in 3D immer noch visualisieren. Wir betrachten folgende Funktion: $y = f(x_1, x_2) = 1 \cdot x_1 + 0.5 \cdot x_2 + 5$.

Aha! Während eine lineare Funktion eines Arguments grafisch einer Gerade entspricht, sehen wir nun, dass eine lineare Funktion zweier Argumente nichts anderes als eine Ebene darstellt. Wir sehen, dass die Ebene die y-Achse am Punkt 5 schneidet. Etwas schwieriger zu sehen ist die Steigung der Ebene in

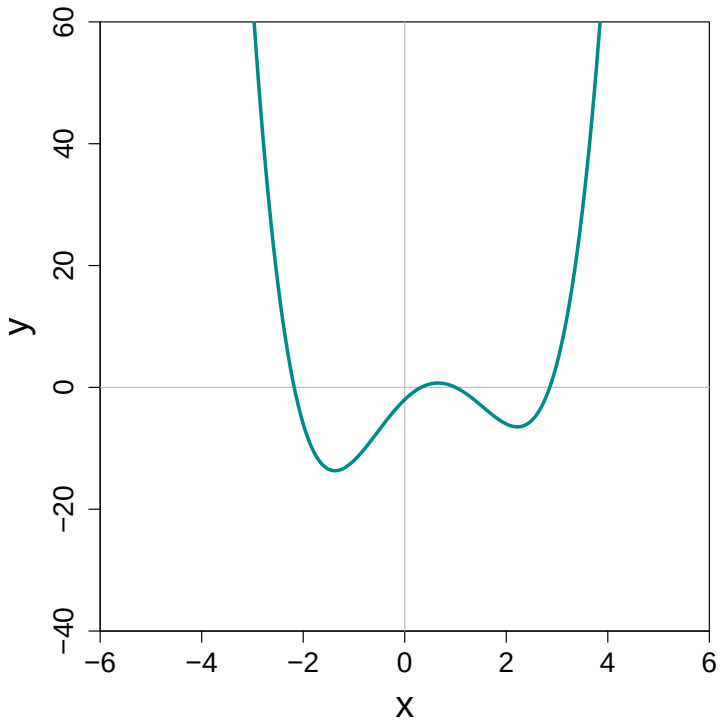


Figure 2.4: Beispiel einer polynomischen Funktion vierten Grades.

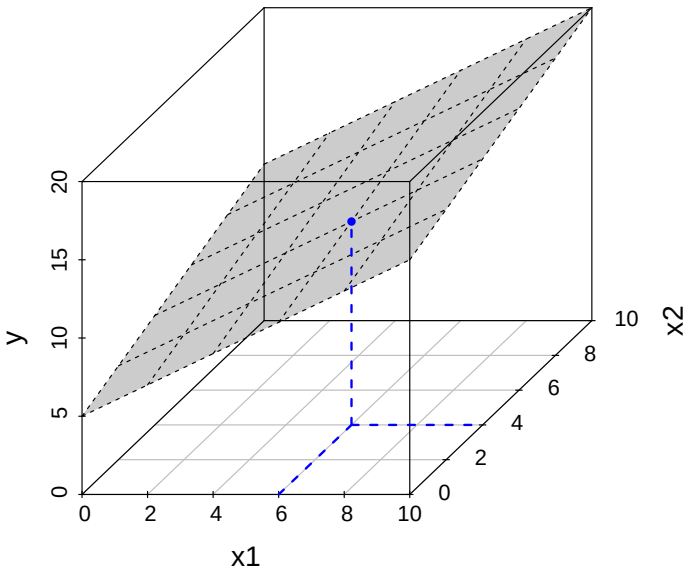


Figure 2.5: Lineare Funktion zweier Argumente (Ebene).

die Richtung der x_1-Achse und in die Richtung der x_2-Achse. Sie können aber vielleicht bereits erraten, dass die (partiellen) Steigungen 1 und 0.5 betragen.

Die Funktion ordnet jeden möglichen Punkt (x_1, x_2) einem Punkt auf der Ebene zu. Wir können zum Beispiel für den in Abb. 2.5 eingezeichneten Punkt $(6, 4)$ den entsprechenden Punkt auf der Ebene ausrechnen:

$$\begin{aligned} y &= 1 \cdot x_1 + 0.5 \cdot x_2 + 5 \\ &= 1 \cdot 6 + 0.5 \cdot 4 + 5 \\ &= 13 \end{aligned}$$

Selbstverständlich könnten wir uns nun auch quadratische Funktionen oder Polynome mehrerer Argumente anschauen, aber darauf verzichten wir vorerst.

2.1.4 Potenzen und Logarithmen

Blabla…

2.2 Integral- und Differentialrechnung

Olteanu materials: Local vs. global minima From a maximization to a minimization problem Basic definition of derivative Differentiation rules local min., max. and saddle point Second derivative test Partial derivatives What is a gradient? What is Hessian? What is Jacobian? Chain rules Lagrange optimization

2.3 Lineare Algebra

Olteanu materials: What is a scalar? What is a vector? What is a matrix? Vector norms Inner products Symmetric, diagonal, square and identity matrix Associative, commutative laws for matrices Matrix addition and multiplication Matrix inversion Eigenvectors and eigenvalues Quadratic form and positive (semi-) definiteness Differentiation rules for matrices

2.4 Wahrscheinlichkeitsrechnung

Olteanu materials: Sample space and axioms of probability Conditional probability definition Discrete vs. continuous random variables Joint probability distributions Expectation and variance, covariance (always for discrete and continuous) Bernoulli, Binomial, Normal, Multivariate Normal, Laplace

2.4.1 Diskrete Zufallsvariablen

Wir werden später sehen, dass im Machine Learning oftmals Dinge als **Zufallsvariablen** modelliert werden. Eine Zufallsvariable X ist eine Variable, für die der konkrete Wert nicht von vornherein klar ist. Wir können mit X zum Beispiel das Resultat eines Münzwurfs modellieren. Die zwei möglichen Resultate sind Kopf und Zahl. Vor dem Münzwurf ist nicht klar, ob Kopf oder Zahl erscheinen wird. Genau darum modellieren wir das Resultat des Münzwurfs als Zufallsvariable.

Es gibt in diesem einfachen Beispiel nur zwei mögliche Resultate (Kopf und Zahl), d.h. die Anzahl möglicher Resultate ist endlich (= nicht unendlich). Darum handelt es sich in diesem Fall um eine **diskrete** Zufallsvariable.

2.5 Verteilungen

Chapter 3

Einführung in das Programmieren mit R

leaRn Materialen

tidymodels

Referenzen auf andere Ressourcen (Hadley et al.)

Chapter 4

Lineare Regression

In diesem Kapitel werden wir uns eingehend mit dem einfachsten Modell für das Regressionsproblem auseinander setzen, nämlich dem linearen Regressionsmodell. Liegt ein Regressionsproblem vor, dann macht es in der Praxis fast immer Sinn mit diesem Modell zu starten und dann die Komplexität nach Bedarf zu erhöhen.

4.1 ML-Modelle im Allgemeinen

Wie bereits in Kapitel 1 gesehen, geht es beim Regressionsproblem darum, eine stetige Variable $y_i \in \mathbb{R}$ möglichst optimal vorherzusagen. Dazu verwenden wir eine oder mehrere Input-Variablen, welche wir kompakt als Vektor $\mathbf{x}_i$ schreiben.

Das Problem ist nur lösbar, falls es tatsächlich einen Zusammenhang zwischen den Input-Variablen $\mathbf{x}_i$ und dem Output y_i gibt. Wir nehmen ganz allgemein an, dass der Zusammenhang zwischen dem Output y_i und den Input-Variablen $\mathbf{x}_i$ mathematisch wie folgt ausgedrückt werden kann:

$$y_i = f(\mathbf{x}_i) + \epsilon$$

- Die Funktion $f(\mathbf{x}_i)$ bezeichnet die **systematische Information**, die wir aus $\mathbf{x}_i$ im Hinblick auf y_i lernen können.
- ϵ ist ein Fehlerterm, der die Differenz zwischen y_i und $f(\mathbf{x}_i)$ abbildet,[1] also den **nicht-lernbaren** (unsystematischen) **Teil**. Der Fehlerterm beinhaltet einerseits den Effekt von Variablen, die uns nicht zur Verfügung stehen, aber einen Einfluss auf den Output y_i haben und andererseits nicht-messbare Variation, oft auch einfach *Noise* genannt. Grob gesagt:

[1] $\epsilon = y_i - f(\mathbf{x}_i)$

alles nicht-messbare. Auch wichtig zu sehen: der Fehler ist **additiv**, wir addieren ihn zum lernbaren Teil hinzu.

Der Output y_i ergibt sich also aus der Addition eines systematischen Teils $f(\mathbf{x}_i)$ sowie eines Fehlerterms ϵ.

Wichtig: Ziel des Machine Learnings ist es, eine Funktion $\hat{f}(\mathbf{x}_i)$ zu trainieren (schätzen), die der wahren aber unbekannten Funktion $f(\mathbf{x}_i)$ so nahe wie möglich kommt. Im (unrealistischen) Idealfall ist unser trainiertes Modell gleich der wahren Funktion, also $\hat{f}(\mathbf{x}_i) = f(\mathbf{x}_i)$ und wir haben die systematische Information perfekt gelernt. Jedes ML-Modell, das wir uns in diesem Buch anschauen werden, kann als eine mathematische Funktion $\hat{f}(\mathbf{x}_i)$ der Input-Variablen $\mathbf{x}_i$ aufgeschrieben werden. Sobald wir $\hat{f}(\mathbf{x}_i)$ trainiert haben, können wir damit Vorhersagen machen, denn die Vorhersage für einen gegebenen Input-Vektor $\mathbf{x}_0$ ist nichts anderes als der Wert der trainierten Funktion an diesem Punkt, also $\hat{y}_0 = \hat{f}(\mathbf{x}_0)$.

4.2 Das Modell (ausgeschrieben)

Nun wollen wir uns konkret mit dem **linearen Regressionsmodell** befassen. Das bedeutet nun nichts anderes, als dass wir die allgemein geschriebene Funktion $f(\mathbf{x}_i)$ durch eine konkrete mathematische Funktion ersetzen. Im Machine Learning ist das der erste wichtige Schritt, nämlich die Modellwahl (engl. *Model Selection*). Das Modell kann wie folgt geschrieben werden:

$$f(\mathbf{x}_i) = w_0 + w_1 \cdot x_{i1} + w_2 \cdot x_{i2} + ... + w_p \cdot x_{ip}$$

Wir verzichten hier bewusst darauf, den Hut für f zu schreiben, da es sich lediglich um eine allgemein gültige Funktion handelt und noch nichts geschätzt bzw. trainiert wurde. Dieses Modell bzw. diese Funktion hat sogenannte **Parameter**, die es zu schätzen gilt. Hier sind dies die Parameter w_0, w_1, ..., w_p. Wegen der Konstante w_0 haben wir immer einen Parameter mehr als es Input-Variablen hat, also $p + 1$ Parameter.

Diese Parameter sind die Schlüsselzutat in einem ML-Modell. Wir wollen sie **optimieren**, so dass die trainierte Funktion $\hat{f}(\mathbf{x}_i)$ der wahren Funktion $f(\mathbf{x}_i)$ möglichst nahe kommt.

Wir schauen uns in diesem Kapitel ein ganz einfaches Beispiel an mit nur einer Input-Variable x_i, so dass der Zusammenhang zwischen dem Output y_i und dem Input x_i in 2D dargestellt werden kann. In diesem Zusammenhang spricht man vom **einfachen linearen Regressionsmodell**. Ausserdem haben wir nur vier Beobachtungen, welche in Abbildung 4.1 in einem Streudiagramm dargestellt werden:

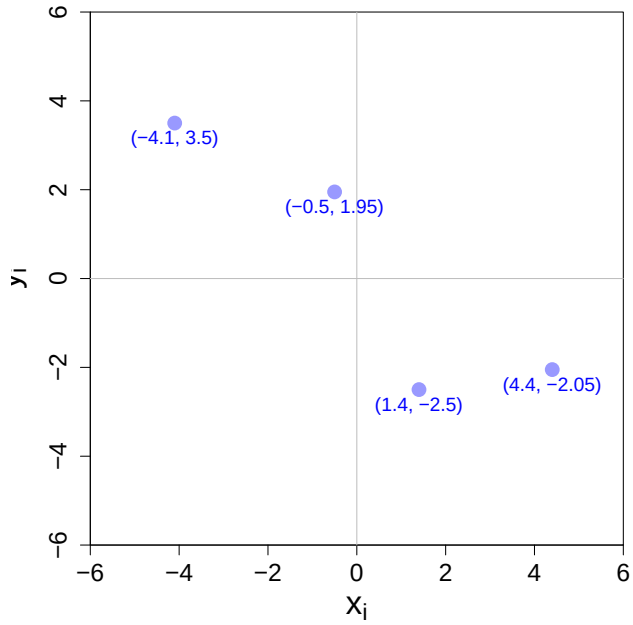


Figure 4.1: Einfaches Regressionsbeispiel. Die vier Beobachtungen werden in einem Streudiagramm dargestellt. Auf der x-Achse ist der Wert der Input-Variable und auf der y-Achse der Wert der Output-Variable ablesbar.

4.3 Das Modell (kompakt)

Sie sehen oben, dass es ziemlich umständlich sein kann, das lineare Regressionsmodell aufzuschreiben, insbesondere wenn wir viele Input-Variablen haben. Mithilfe von **Vektoren und Matrizen** können wir das Modell viel kompakter aufschreiben.

Wir haben in Kapitel 1 bereits gesehen, dass die Input-Variablen für eine Beobachtung i als Spaltenvektor geschrieben werden können. Wir modifizieren diesen Spaltenvektor in einem ersten Schritt, indem wir an erster Stelle eine 1 einfügen,[2] also:

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Nun stecken wir die Parameter des Modells ebenfalls in einen Spaltenvektor:

$$\mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix}$$

Wir können nun das lineare Regressionsmodell (für die Beobachtung i) als **Skalarprodukt** dieser beiden Vektoren aufschreiben:

$$f(\mathbf{x}_i) = \mathbf{w}'\mathbf{x_i} \tag{4.1}$$

$$= \begin{pmatrix} w_0 & w_1 & w_2 & ... & w_p \end{pmatrix} \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \tag{4.2}$$

$$= w_0 \cdot 1 + w_1 \cdot x_{i1} + w_2 \cdot x_{i2} + \cdots + w_p \cdot x_{ip} \tag{4.3}$$

Die Form $\mathbf{w}'\mathbf{x_i}$ ist schon ziemlich kompakt, aber es geht noch besser. Wir können nämlich das Modell gleich für alle n Beobachtungen (und nicht nur für die i-te Beobachtung) aufschreiben. Dazu müssen wir die Input-Variablen für jede Beobachtung i in einer Matrix anordnen:

[2] So müssen wir die Konstante w_0 nicht separat aufschreiben.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Die Matrix $\mathbf{X}$ wird typischerweise **Design Matrix** genannt. Die erste Zeile enthält die Input-Variablen für die erste Beobachtung, die zweite Zeile die Input-Variablen für die zweite Beobachtung, usw. Nun können wir das Modell mithilfe einer Multiplikation zwischen der Design Matrix $\mathbf{X}$ und dem Spaltenvektor $\mathbf{w}$ in einem Schritt für alle Beobachtungen aufschreiben:

$$f(\mathbf{X}) = \mathbf{X}\mathbf{w} \tag{4.4}$$

$$= \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \cdots \\ w_p \end{pmatrix} \tag{4.5}$$

$$= \begin{pmatrix} w_0 \cdot 1 + w_1 \cdot x_{11} + w_2 \cdot x_{12} + \cdots + w_p \cdot x_{1p} \\ w_0 \cdot 1 + w_1 \cdot x_{21} + w_2 \cdot x_{22} + \cdots + w_p \cdot x_{2p} \\ \cdots \\ w_0 \cdot 1 + w_1 \cdot x_{n1} + w_2 \cdot x_{n2} + \cdots + w_p \cdot x_{np} \end{pmatrix} \tag{4.6}$$

Überprüfen wir doch noch kurz die Dimensionen von obigem Matrix-Vektor Produkt. Die Matrix $\mathbf{X}$ hat n Zeilen und $p+1$ Spalten und darum eine Dimensionalität von $n \times (p+1)$. Der Spaltenvektor $\mathbf{w}$ hat Dimensionalität $(p+1) \times 1$. Das Matrix-Vektor Produkt hat dementsprechend eine Dimensionalität von $n \times 1$, genau was wir erwarten würden, nämlich einen Vektor mit den Vorhersagen für alle n Beobachtungen.

Für unser einfaches Beispiel kann das Modell wie folgt in Matrixform geschrieben werden:

$$f(\mathbf{X}) = \mathbf{X}\mathbf{w} \tag{4.7}$$

$$= \begin{pmatrix} 1 & -4.1 \\ 1 & -0.5 \\ 1 & 1.4 \\ 1 & 4.4 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \tag{4.8}$$

$$= \begin{pmatrix} w_0 \cdot 1 - w_1 \cdot 4.1 \\ w_0 \cdot 1 - w_1 \cdot 0.5 \\ w_0 \cdot 1 + w_1 \cdot 1.4 \\ w_0 \cdot 1 + w_1 \cdot 4.4 \end{pmatrix} \tag{4.9}$$

Warum wir all das tun, werden wir weiter unten sehen. Es wird unser Leben viel einfacher machen! Versuchen Sie diesen Abschnitt hier gut zu verstehen, so dass Sie sobald wie möglich mit der Matrixschreibweise von Modellen vertraut sind.

4.4 Modelltraining

Wir werden uns hier anschauen, dass für das Training (oft auch *Fitting* genannt) des linearen Regressionsmodells **zwei verschiedene Perspektiven** eingenommen werden können, welche am Schluss beide zum selben Schluss kommen.

4.4.1 Perspektive 1: Funktionsoptimierung

In der ersten Perspektive behandeln wir das Modelltraining als Optimierungsproblem. Wir wollen nämlich eine sogenannte **Kostenfunktion** (engl. *Loss Function*) aufstellen, die es danach zu minimieren gilt. Sie werden gleich sehen, dass die Kostenfunktion für das lineare Regressionsmodell von den Modellparameter $w_0, w_1, ..., w_p$ abhängt. Das Ziel wird also sein, die optimalen Werte für die Modellparameter zu finden, so dass die Kostenfunktion so klein wie möglich ist.

Doch wie sieht denn nun diese Kostenfunktion für das lineare Regressionsmodell konkret aus? Wir werden uns hier der Einfachheit halber nur ein **einfaches lineares Regressionsmodell** mit nur einer Input-Variable x_i anschauen (wie in unserem einfachen Beispiel). Die Kostenfunktion sieht in diesem Fall so aus:

$$J(\hat{w}_0, \hat{w}_1) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2$$

Sie sehen, dass die Kostenfunktion $J(\hat{w}_0, \hat{w}_1)$ eine Funktion der beiden (trainierten) Modellparameter ist. Vielleicht wundern Sie sich nun, wie diese Kostenfunktion von den Modellparameter abhängt, da diese in obiger Formel ja gar nicht direkt ersichtlich sind. Schreiben wir die Kostenfunktion doch mal etwas um:

$$J(\hat{w}_0, \hat{w}_1) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2 \tag{4.10}$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - (\hat{w}_0 + \hat{w}_1 \cdot x_i) \right)^2 \tag{4.11}$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \hat{w}_0 - \hat{w}_1 \cdot x_i \right)^2 \tag{4.12}$$

$$\tag{4.13}$$

Nun ist offensichtlich, wie die Kostenfunktion J von den Modellparameter $\hat{w}_0$ und $\hat{w}_1$ abhängt. Im ML gibt es nun viele verschiedene Arten, wie man für die beiden Modellparameter die optimalen Werte findet. Hier ist die Lösung zum Glück einfach, denn es gibt eine sogenannte **analytische Lösung**, d.h. es ist möglich für $\hat{w}_0$ und $\hat{w}_1$ je eine Formel zu finden, die uns erlaubt die optimalen Parameterwerte direkt auszurechnen. Die Herleitung dieser Formeln ist nicht besonders schwierig, denn wir wenden nämlich ein altbekanntes Prinzip aus der Differenzialrechnung an: wir berechnen die erste Ableitung der Funktion nach den Modellparameter, setzen sie gleich Null und lösen nach dem Parameter auf.

Machen wir dies in einem ersten Schritt für $\hat{w}_0$:

$$\frac{\partial J(\hat{w}_0, \hat{w}_1)}{\partial \hat{w}_0} = \frac{1}{2n} \sum_{i=1}^{n} 2 \cdot (y_i - \hat{w}_0 - \hat{w}_1 \cdot x_i) \cdot (-1) \tag{4.14}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{w}_0 - \hat{w}_1 \cdot x_i) \tag{4.15}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} y_i + \frac{1}{n} \sum_{i=1}^{n} \hat{w}_0 + \frac{1}{n} \sum_{i=1}^{n} \hat{w}_1 \cdot x_i \tag{4.16}$$

$$= -\bar{y} + \frac{1}{n} \cdot n \cdot \hat{w}_0 + \hat{w}_1 \cdot \bar{x} \tag{4.17}$$

$$= -\bar{y} + \hat{w}_0 + \hat{w}_1 \cdot \bar{x} \tag{4.18}$$

Nun setzten wir die Ableitung gleich Null und lösen nach $\hat{w}_0$ auf:

$$-\bar{y} + \hat{w}_0 + \hat{w}_1 \cdot \bar{x} = 0 \tag{4.19}$$

$$\hat{w}_0 = \bar{y} - \hat{w}_1 \cdot \bar{x} \tag{4.20}$$

Wir sehen, dass die Lösung für $\hat{w}_0$ von den beiden Mittelwerten $\bar{y}$ und $\bar{x}$ sowie von $\hat{w}_1$ abhängt. Suchen wir nun also in einem zweiten Schritt die Lösung für $\hat{w}_1$:

$$\frac{\partial J(\hat{w}_0, \hat{w}_1)}{\partial \hat{w}_1} = \frac{1}{2n} \sum_{i=1}^{n} 2 \cdot (y_i - \hat{w}_0 - \hat{w}_1 \cdot x_i) \cdot (-x_i) \tag{4.21}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} (y_i \cdot x_i - \hat{w}_0 \cdot x_i - \hat{w}_1 \cdot x_i^2) \tag{4.22}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} y_i \cdot x_i + \hat{w}_0 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i + \hat{w}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 \tag{4.23}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} y_i \cdot x_i + \hat{w}_0 \cdot \bar{x} + \hat{w}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 \tag{4.24}$$

$$\tag{4.25}$$

Nun können wir wiederum die Ableitung gleich Null setzen und für $\hat{w}_0$ setzen wir unsere Lösung von oben ein. Danach lösen wir nach $\hat{w}_1$ auf:

$$-\frac{1}{n} \sum_{i=1}^{n} y_i \cdot x_i + \hat{w}_0 \cdot \bar{x} + \hat{w}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 = 0 \tag{4.26}$$

$$(\bar{y} - \hat{w}_1 \cdot \bar{x}) \cdot \bar{x} + \hat{w}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1}{n} \sum_{i=1}^{n} y_i \cdot x_i \tag{4.27}$$

$$\bar{y} \cdot \bar{x} - \hat{w}_1 \cdot \bar{x}^2 + \hat{w}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1}{n} \sum_{i=1}^{n} y_i \cdot x_i \tag{4.28}$$

$$\hat{w}_1 \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right) = \frac{1}{n} \sum_{i=1}^{n} y_i \cdot x_i - \bar{y} \cdot \bar{x} \tag{4.29}$$

$$\hat{w}_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} y_i \cdot x_i - \bar{y} \cdot \bar{x}}{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2} \tag{4.30}$$

Vielleicht erkennen Sie die Ausdrücke im Zähler und Nenner der Lösung für $\hat{w}_1$: es sind dies die **Kovarianz** zwischen y_i und x_i im Zähler und die **Varianz** von x_i im Nenner.

Yay! Nun haben wir die Formeln für die Berechnung der optimalen Parameterwerte des einfachen linearen Regressionsmodells gefunden. Diese Methode wird **Kleinstquadratemethode** (engl. *Least Squares*) genannt, weil die optimalen Parameter die Summe über die **quadrierten** Differenzen zwischen y_i und den Vorhersagen $\hat{f}(x_i)$ minimieren.

Aufgabe

```
#> Warning: `includeHTML()` was provided a `path` that appears to be a complete HTML do
#> x Path: exercises/simplelinreg.html
#> i Use `tags$iframe()` to include an HTML document. You can either ensure `path` is a
```

Das in der obigen Aufgabe berechnete Modell ist in Abbildung 4.2 (links) grafisch als blaue Gerade dargestellt. Der Parameter $\hat{w}_0$ ist der Ort, an dem die Gerade die y-Achse durchkreuzt, während der Parameter $\hat{w}_1$ der Steigung der Geraden entspricht. Unser optimales Modell minimiert die Summe über die quadrierten Differenzen zwischen den tatsächlichen y_i Werten und den Vorhersagen gemäss unserem Modell $\hat{f}(x_i)$ (als rot gestrichelte Linien dargestellt).

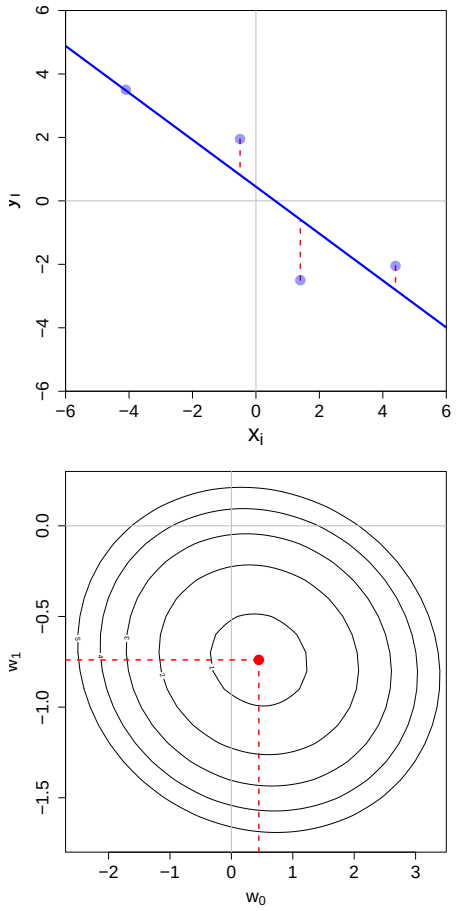


Figure 4.2: Einfaches Regressionsbeispiel. Das geschätzte Modell ist als blaue Gerade eingezeichnet. Die vertikalen roten Linien stellen die Abweichungen der wahren Outputs von den Vorhersagen dar. Rechts ist ein Konturplot der Kostenfunktion mit der optimalen Parameterwert-Kombination dargestellt.

Die Abbildung 4.2 (rechts) zeigt sogenannte **Konturlinien** unserer Kostenfunktion. Die optimale Parameterwert-Kombination ist als roter Punkt eingezeichnet. Jede Konturlinie zeigt alle Parameterwert-Kombination, welche jeweils zum gleichen Kostenwert führen. Die fünf eingezeichneten Linien zeigen

beispielsweise die Parameterwert-Kombination für die Kostenwerte 1 bis 5 (von innen nach aussen). Man kann sich unsere Kostenfunktion also wie eine Schüssel vorstellen mit dem roten Punkt als Boden der Schüssel. Es handelt sich bei unserer Kostenfunktion um eine Funktion, die **quadratisch** in den Parameterwerten $\hat{w}_0$ und $\hat{w}_1$ ist. In diesem Fall finden wir immer **genau eine Parameterwert-Kombination**, welche dem absoluten Minimum der Kostenfunktion entspricht. Manchmal spricht man auch von einer **konvexen** Kostenfunktion.

Optional: Kleinstquadratemethode in Matrixform

Die obige Herleitung funktioniert nur für das einfache lineare Regressionsmodell mit einer Input-Variable x_i. Wir schauen uns hier nun kurz die allgemeine Lösung in Matrixform an. Wir nehmen an, dass die Werte unseres Outputs alle in einem Spaltenvektor $\mathbf{y}$ organisiert sind und unsere Modellvorhersagen als $\mathbf{X\hat{w}}$ geschrieben werden können.

Dann können wir unsere Kostenfunktion von oben wie folgt in Matrixform schreiben:

$$J(\hat{\mathbf{w}}) = \frac{1}{2n}(\mathbf{y} - \mathbf{X\hat{w}})'(\mathbf{y} - \mathbf{X\hat{w}}) \tag{4.31}$$

Das sieht schlimmer aus als es ist, denn $(\mathbf{y} - \mathbf{X\hat{w}})$ ist lediglich ein Spaltenvektor mit den Differenzen zwischen den wahren y_i und den Vorhersagen unseres Modells. Wenn wir diesen Spaltenvektor $\mathbf{e}$ nennen, dann kann obiger Ausdruck als $\frac{1}{2n}\mathbf{e}'\mathbf{e}$ geschrieben werden, wobei $\mathbf{e}'\mathbf{e}$ ein Skalarprodukt ist und dementsprechend einen Skalar bzw. eine einzige Zahl zurück gibt. Diese Zahl multipliziert mit $\frac{1}{2n}$ ist dann nichts anderes als der Wert unserer Kostenfunktion. Sie sehen also, dass wir mit dem Skalarprodukt $\mathbf{e}'\mathbf{e}$ die Summe ersetzen können.

Nun wenden wir die bekannten Matrix-Rechenregeln an, um die Kostenfunktion umzuschreiben:

$$\begin{aligned} J(\hat{\mathbf{w}}) &= \frac{1}{2n}(\mathbf{y} - \mathbf{X\hat{w}})'(\mathbf{y} - \mathbf{X\hat{w}}) && (4.32)\\ &= \frac{1}{2n}(\mathbf{y}' - \hat{\mathbf{w}}'\mathbf{X}')(\mathbf{y} - \mathbf{X\hat{w}}) && (4.33)\\ &= \frac{1}{2n}(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X\hat{w}} - \hat{\mathbf{w}}'\mathbf{X}'\mathbf{y} + \hat{\mathbf{w}}'\mathbf{X}'\mathbf{X\hat{w}}) && (4.34) \end{aligned}$$

Wenn Sie sich kurz anhand der Dimensionalität der einzelnen Komponenten überlegen, was das Endprodukt des Ausdrucks $\mathbf{y}'\mathbf{X\hat{w}}$ ist, dann werden Sie sehen, dass ein Skalar (Dimensionalität 1×1) resultiert. Darum muss zwingend auch die transponierte Form davon, $(\mathbf{y}'\mathbf{X\hat{w}})' = \hat{\mathbf{w}}'\mathbf{X}'\mathbf{y}$ ein Skalar sein, was dazu

führt, dass die beiden mittleren Terme in der letzten Zeile von obiger Kostenfunktion identisch sein müssen. Deshalb können wir die Kostenfunktion wie folgt umschreiben:

$$J(\hat{\mathbf{w}}) = \frac{1}{2n}(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\mathbf{w}} + \hat{\mathbf{w}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{w}}) \tag{4.35}$$

So, nun können wir die Kostenfunktion nach dem Spaltenvektor mit den Modellparameter $\hat{\mathbf{w}}$ ableiten. Man spricht in diesem Fall nun nicht von einer Ableitung, sondern von einem **Gradienten**. Auch die mathematische Schreibweise ist etwas anders:

$$\nabla_{\hat{\mathbf{w}}} J(\hat{\mathbf{w}}) = \frac{1}{2n}(-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\mathbf{w}}) \tag{4.36}$$

$$= \frac{1}{n}(-\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\hat{\mathbf{w}}) \tag{4.37}$$

Diesen Ausdruck können wir nun wie gewohnt gleich Null setzen (wobei wir hier rechts einen Nullvektor $\mathbf{0}$ setzen) und mit den Matrix-Rechenregeln nach $\hat{\mathbf{w}}$ auflösen:

$$\frac{1}{n}(-\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\hat{\mathbf{w}}) = \mathbf{0} \tag{4.38}$$

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}'\mathbf{y} \tag{4.39}$$

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{4.40}$$

Wichtig: Die Matrix $\mathbf{X}'\mathbf{X}$ hat eine Dimensionalität von $(p+1) \times (p+1)$, ist also quadratisch. Sie ist nur invertierbar, wenn die Design Matrix mehr Zeilen als Spalten hat, also wenn $n > (p+1)$.

4.4.2 Perspektive 2: Wahrscheinlichkeitstheorie

Nun werden wir sehen, dass wir die Lösung oben (aus Perspektive 1) auch mit einer probabilistischen Sicht auf die Dinge erhalten. Dazu schreiben wir nochmals kurz den allgemein angenommenen Zusammenhang zwischen dem wahren Output y_i und den Input-Variablen auf und konkretisieren ihn dann gleich für das lineare Regressionsmodell:

$$y_i = f(\mathbf{x}_i) + \epsilon \tag{4.41}$$

$$= \mathbf{w}'\mathbf{x_i} + \epsilon \tag{4.42}$$

$$\tag{4.43}$$

Nun nehmen wir an, dass der Fehlerterm ϵ normalverteilt ist mit Mittelwert 0 und Varianz σ^2, also $\epsilon \sim N(0, \sigma^2)$. Weil wir annehmen, dass $\mathbf{w'x_i}$ fix ist (also keine Zufallsvariable), ist unser Output y_i normalverteilt mit Mittelwert $\mathbf{w'x_i}$ und Varianz σ^2:

$$y_i \sim \mathcal{N}\left(\mathbf{w'x_i}, \sigma^2\right)$$

Grafisch zeigen!

Nun möchten wir wissen, was die **gemeinsame Verteilung** aller Output-Werte in unserem Datensatz ist. D.h. wie sieht die Wahrscheinlichkeit $p(y_1, y_2, ..., y_n|\mathbf{w}, \mathbf{X}, \sigma^2)$ aus? Weil wir annehmen, dass alle Beobachtungen i in unserem Datensatz unabhängig sind, sieht die Antwort auf die Frage folgendermassen aus:

$$p(y_1, y_2, ..., y_n|\mathbf{w}, \mathbf{X}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{w'x_i}, \sigma^2\right)$$

Maximum Likelihood

Die gemeinsame Wahrscheinlichkeit $p(y_1, y_2, ..., y_n)$ wird in der Fachsprache **Likelihood** genannt. Die zentrale Idee hier ist, dass wir die Modellparameter $\mathbf{w}$ so wählen, dass die *Likelihood* maximal wird. Der daraus folgende Ausdruck für $\mathbf{w}$ wird **Maximum Likelihood** Schätzer genannt und oft als ML abgekürzt, was sehr verwirrlich sein kann, da wir ja auch Machine Learning so abkürzen.

Wir können nun in der Likelihood oben anstelle von $\mathcal{N}\left(\mathbf{w'x_i}, \sigma^2\right)$ jeweils die Dichtefunktion der Normalverteilung einsetzen:

$$p(y_1, y_2, ..., y_n|\mathbf{w}, \mathbf{X}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{w'x_i}, \sigma^2\right) \tag{4.44}$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mathbf{w'x_i}}{\sigma}\right)^2\right) \tag{4.45}$$

Nun vollziehen wir einen kleinen mathematischen Trick, der vielfach angewendet wird: anstelle der *Likelihood* verwenden wir nun den natürlichen Logarithmus der *Likelihood* (*Log-Likelihood*). Das ist möglich, weil sich so das Optimierungsproblem nicht verändert. Das Logarithmieren vereinfacht das Problem ungemein, denn der Logarithmus eines Produkts wird zu einer Summe der logarithmierten Elemente:

$$\ln p(y_1, y_2, \dots, y_n|\mathbf{w}, \mathbf{X}, \sigma^2) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mathbf{w}'\mathbf{x_i}}{\sigma}\right)^2\right)\right) \tag{4.46}$$

$$= \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mathbf{w}'\mathbf{x_i}}{\sigma}\right)^2\right)\right) \tag{4.47}$$

$$= \sum_{i=1}^{n} \ln(1) - \ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2}\left(\frac{y_i - \mathbf{w}'\mathbf{x_i}}{\sigma}\right)^2 \tag{4.48}$$

$$= \sum_{i=1}^{n} \ln(1) - \sum_{i=1}^{n} \ln\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n} \frac{1}{2}\left(\frac{y_i - \mathbf{w}'\mathbf{x_i}}{\sigma}\right)^2 \tag{4.49}$$

$$= n \cdot \ln(1) - n \cdot \ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{w}'\mathbf{x_i})^2 \tag{4.50}$$

Wow, nun haben wir ein tolles Resultat gefunden: je kleiner der Term $\sum_{i=1}^{n} (y_i - \mathbf{w}'\mathbf{x_i})^2$ in obiger Gleichung, desto grösser ist der natürliche Logarithmus der *Likelihood*. Das heisst nichts anderes, als dass die Kleinstquadratemethode auch der *Maximum Likelihood* Schätzer ist.

Wir haben diesen Abschnitt damit begonnen anzunehmen, dass unser Output y_i normalverteilt ist, d.h. $y_i \sim \mathcal{N}(\mathbf{w}'\mathbf{x_i}, \sigma^2)$. Wir haben nun herausgefunden, dass wir den Spaltenvektor mit den Parameter mit der Kleinstquadratemethode berechnen können. Um die Normalverteilung vollkommen zu spezifizieren, benötigen wir nun noch eine Formel, um die Varianz σ^2 zu rechnen. Dazu leiten wir den obigen Ausdruck der *Log-Likelihood* nach σ ab:

$$\frac{\partial \ln p(y_1, y_2, \dots, y_n|\mathbf{w}, \mathbf{X}, \sigma^2)}{\partial \sigma} = -n \cdot \frac{\sqrt{2\pi}}{\sigma\sqrt{2\pi}} - (-\frac{2}{\sigma^3}) \cdot \frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{w}'\mathbf{x_i})^2 \tag{4.51}$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \mathbf{w}'\mathbf{x_i})^2 \tag{4.52}$$

Nun können wir wie gewohnt die Ableitung gleich Null setzen und nach σ auflösen:

$$-\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}\left(y_i - \hat{\mathbf{w}}'\mathbf{x_i}\right)^2 = 0 \tag{4.53}$$

$$\frac{n}{\hat{\sigma}} = \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}\left(y_i - \hat{\mathbf{w}}'\mathbf{x_i}\right)^2 \tag{4.54}$$

$$\frac{\hat{\sigma}^3}{\hat{\sigma}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\mathbf{w}}'\mathbf{x_i}\right)^2 \tag{4.55}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\mathbf{w}}'\mathbf{x_i}\right)^2 \tag{4.56}$$

$$\tag{4.57}$$

Sehr schön, dieses Resultat macht ebenfalls viel Sinn. Die geschätzte Varianz $\hat{\sigma}^2$ ist nichts anderes als der durchschnittliche quadrierte Fehler (engl. *Mean Squared Error*).

Aufgabe

Berechnen Sie …

4.5 Regularisierte Regression

Das zentrale Problem der oben kennen gelernten Kleinstquadratemethode ist, dass sie extrem anfällig auf **Overfitting** ist. Beim linearen Regressionsmodell ist Overfitting vor allem dann ein Problem, wenn die Anzahl Input-Variablen p relativ gross ist im Vergleich zur Anzahl Beobachtungen n. Im Extremfall haben wir mehr Input-Variablen als Beobachtungen ($(p+1) > n$), was dazu führt, dass der Kleinstquadrateschätzer mathematisch nicht rechenbar ist, weil $\mathbf{X}'\mathbf{X}$ nicht invertierbar ist. Das sollte auch intuitiv Sinn machen, denn wie soll eine Schätzung funktionieren, wenn wir im Schnitt weniger als eine Beobachtung pro zu schätzenden Parameter haben.

Wir können das Problem des Overfittings weitgehend lösen, indem wir ein **regularisiertes** Regressionsmodell rechnen. Regularisierung bedeutet eigentlich nichts anderes, als dass wir die ursprüngliche Kostenfunktion für das lineare Regressionsmodell modifizieren. Dabei gibt es zwei bekannte Regularisierungsarten, nämlich **Ridge** oder **LASSO**. Wir fokussieren in einem ersten Schritt auf die Ridge Regularisierung, weil wir in diesem Fall nach wie vor eine analytische Lösung finden.

4.5.1 Ridge Regressionsmodell

Die Kostenfunktion für das Ridge Regressionsmodell sieht wie folgt aus:

$$J(\mathbf{w}) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \hat{f}(\mathbf{x}_i) \right)^2 + \frac{\lambda}{2} \cdot \sum_{j=1}^{p} w_j^2$$

Diese modifizierte Kostenfunktion hat etwas Erklärungsbedarf:

- Wir versuchen hier Modellparameter **w** zu finden, welche **gleichzeitig** den durchschnittlichen quadrierten Fehler sowie eine Summe über die quadrierten Modellparameter so klein wie möglich machen. Das sind zwei **konkurrenzierende Ziele** und während des Modelltrainings muss der beste Tradeoff gefunden werden.
- Der Regularisierungsterm ist eine Summe über die quadrierten Modellparameter. Das Quadrieren stellt sicher, dass sich positive und negative Parameterwerte nicht gegenseitig kompensieren.
- Der **Hyperparameter** λ legt fest, wie viel (relatives) Gewicht der Regularisierungsterm im Verhältnis zum durchschnittlichen quadrierten Fehler bekommt. Je grösser λ, desto stärker "bestrafen" wir komplexe Modelle. Wir werden später sehen, wir wir den optimalen Wert für λ via **Cross-Validation** finden können.
- Der Regularisierungsterm enthält die Konstante w_0 **nicht** (Summe startet bei $j = 1$ und nicht bei $j = 0$).

Aufgabe

- Was passiert wenn $\lambda = 0$?
- Was passiert wenn $\lambda \to \infty$?

Aufgabe

```
#> Warning: `includeHTML()` was provided a `path` that appears to be a complete HTML document.
#> x Path: exercises/ridgederiv.html
#> i Use `tags$iframe()` to include an HTML document. You can either ensure `path` is accessible
```

Optional: Ridge Regression in Matrixform

Der Einfachheit halber nehmen wir hier an, dass die Outputwerte y_i hier standardisiert[3] wurden, so dass der Mittelwert über die standardisierten Outputwerte Null ist. So entfällt die Konstante w_0 aus dem Modell, was uns die Matrixform für das Ridge Modell erleichtert, denn der Regularisierungsterm soll ja die Konstante nicht enthalten und wenn es diese nicht gibt, dann gibt es keine Probleme.

Wie weiter oben gesehen, können wir die Kostenfunktion für das nicht-regularisierte Modell wie folgt schreiben:

[3] Formel für die Standardisierung: $\frac{y_i - \bar{y}}{s_y}$

$$J(\hat{\mathbf{w}}) = \frac{1}{2n}(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\mathbf{w}} + \hat{\mathbf{w}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{w}}) \tag{4.58}$$

Der Regularisierungsterm kann sehr einfach in Matrixform geschrieben werden, nämlich als Skalarprodukt $\frac{\lambda}{2}\hat{\mathbf{w}}'\hat{\mathbf{w}}$. Damit kriegen wir folgende Kostenfunktion:

$$J(\hat{\mathbf{w}}) = \frac{1}{2n}(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\mathbf{w}} + \hat{\mathbf{w}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{w}}) + \frac{\lambda}{2}\hat{\mathbf{w}}'\hat{\mathbf{w}} \tag{4.59}$$

Um den Gradienten dieser Kostenfunktion zu finden, gehen wir nun sehr ähnlich wie oben vor:

$$\nabla_{\hat{\mathbf{w}}} J(\hat{\mathbf{w}}) = \frac{1}{2n}(-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\mathbf{w}}) + \frac{2\lambda}{2}\hat{\mathbf{w}} \tag{4.60}$$

$$= \frac{1}{n}(-\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\hat{\mathbf{w}}) + \lambda\hat{\mathbf{w}} \tag{4.61}$$

Diesen Ausdruck können wir nun wie gewohnt gleich Null setzen und mit den Matrix-Rechenregeln nach $\hat{\mathbf{w}}$ auflösen:

$$\frac{1}{n}(-\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\hat{\mathbf{w}}) + \lambda\hat{\mathbf{w}} = \mathbf{0} \tag{4.62}$$

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{w}} + \lambda\hat{\mathbf{w}} = \mathbf{X}'\mathbf{y} \tag{4.63}$$

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})\hat{\mathbf{w}} = \mathbf{X}'\mathbf{y} \tag{4.64}$$

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \tag{4.65}$$

$$\tag{4.66}$$

Wichtig: $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})$ ist immer invertierbar, auch wenn $p > n$. Wir haben nun also ein analytisch lösbares Regressionsmodell gefunden, dass gut gegen Overfitting schützt.

Standardisierung der Input-Variablen

Es ist eminent wichtig, dass Sie alle numerischen Input-Variablen vor der Anwendung eines regularisierten Modells standardisieren, so dass alle Variablen auf der selben Skala "leben". Warum ist das so wichtig? Sie haben gesehen, dass wir beim Ridge Modell die Grösse der Parameter mit dem Regularisierungsterm beschränken. Wenn jedoch die Input-Variablen alle auf unterschiedlichen Skalen "leben", dann sind die Parameter nur schon deshalb unterschiedlich. Durch die Standardisierung der Input-Variablen erreichen wir, dass die Parametergrössen vergleichbar werden und die Regularisierung so auch korrekt funktioniert.

4.5.2 LASSO Regressionsmodell

Blabla...

4.6 Bias-Variance Tradeoff

Wir haben oben bereits gesehen, dass der Hyperparameter λ die Komplexität des regularisierten Regressionsmodells bestimmt. Um noch besser zu verstehen, warum diese Komplexität überhaupt wichtig ist, wollen wir uns nun mit einem ganz wichtigen Konzept beschäftigen, nämlich dem **Bias-Variance Tradeoff**. Dieses Konzept kann *intuitiv* für alle Bereiche des Supervised Learnings angewendet werden. Für das Regressionsproblem können wir diesen Tradeoff jedoch auch *mathematisch* herleiten und genau das tun wir jetzt hier.

Stellen Sie sich vor, dass wir eine grosse Anzahl Datensätze zur Verfügung haben und mit jedem dieser Datensätze versuchen wir den wahren funktionalen Zusammenhang $f(\mathbf{x}_i)$ möglichst gut mit $\hat{f}(\mathbf{x}_i)$ zu schätzen. Für jeden Datensatz sieht das geschätzte Modell $\hat{f}(\mathbf{x}_i)$ etwas anders aus. Das geschätzte Modell $\hat{f}$ variiert also je nach Datensatz und ist dementsprechend eine **Zufallsvariable**.

Ausserdem treffen wir folgende Annahmen:

- Von oben wissen wir, dass $y_i = f(\mathbf{x}_i) + \epsilon$ gilt.
- Wir nehmen an, dass der Erwartungswert des nicht-lernbaren Teils ϵ Null ist, also $\mathbb{E}[\epsilon] = 0$.
- Allgemeine Regel zur Varianz einer Zufallsvariable: $\text{Var}(\epsilon) = \mathbb{E}[\epsilon^2] - \mathbb{E}[\epsilon]^2 = \mathbb{E}[\epsilon^2] - 0^2 = \mathbb{E}[\epsilon^2]$.

Um den Bias-Variance Tradeoff zu zeigen, leiten wir nun den **Erwartungswert des quadrierten Fehlers** für eine gegebene Testbeobachtung her, die wir als $(y_0, \mathbf{x}_0)$ bezeichnen. Dies wäre der durchschnittliche quadrierte Fehler, den wir für diese Beobachtung kriegen würden, wenn wir mit jedem geschätzten Modell (jedes auf einem unterschiedlichen Datensatz trainiert) die Vorhersage für diese Testbeobachtung rechnen würden.

In einem ersten Schritt erweitern wir den quadrierten Fehler, indem wir einmal den wahren Funktionswert an der Stelle $\mathbf{x}_0$ einmal abziehen und einmal hinzuzählen. Zusammen gibt das Null und verändert darum die rechte Seite der Gleichung nicht:

$$\mathbb{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathbb{E}\left[\left(y_0 - f(\mathbf{x}_0) + f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] \tag{4.67}$$

Nun verwenden wir die bekannte polynomische Expansion $(a+b)^2 = a^2+2ab+b^2$, aber hier behandeln wir $y_0 - f(\mathbf{x}_0)$ als a und $f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)$ als b. Dadurch kriegen wir folgende Gleichung:

$$\mathbb{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathbb{E}\Big[\left(y_0 - f(\mathbf{x}_0)\right)^2 \tag{4.68}$$

$$+ 2\left(y_0 - f(\mathbf{x}_0)\right)\left(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right) \tag{4.69}$$

$$+ \left(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\Big] \tag{4.70}$$

Nun wissen wir aus obigen Annahmen, dass der Erwartungswert von y_0 folgender ist: $\mathbb{E}[y_0] = \mathbb{E}[f(\mathbf{x}_0) + \epsilon] = f(\mathbf{x}_0)$. Dadurch entfällt der erste Teil des zweiten Terms, weil $\mathbb{E}[(y_0 - f(\mathbf{x}_0))] = f(\mathbf{x}_0) - f(\mathbf{x}_0) = 0$. Dadurch lässt sich das Ganze massiv vereinfachen zu:

$$\mathbb{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathbb{E}\left[\left(y_0 - f(\mathbf{x}_0)\right)^2\right] + \mathbb{E}\left[\left(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] \tag{4.71}$$

Nun setzen wir im ersten Erwartungswert auf der rechten Seite anstelle von y_0 den Term $f(\mathbf{x}_0) + \epsilon$ ein und kriegen folgendes:

$$\mathbb{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathbb{E}\left[\left(f(\mathbf{x}_0) + \epsilon - f(\mathbf{x}_0)\right)^2\right] + \mathbb{E}\left[\left(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] \tag{4.72}$$

$$= \mathbb{E}\left[\epsilon^2\right] + \mathbb{E}\left[\left(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] \tag{4.73}$$

$$= \mathrm{Var}(\epsilon) + \mathbb{E}\left[\left(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] \tag{4.74}$$

Das ist schon mal ein erstes wichtiges Zwischenresultat. Der Erwartungswert des quadrierten Fehlers wird eine untere Grenze haben, die genau der Varianz des Noises $\mathrm{Var}(\epsilon)$ entspricht. Diese untere Grenze des erwarteten Fehlers wird dann erreicht, wenn unser geschätztes Modell genau dem wahren entspricht und darum der zweite Term oben entfällt.

Nun wollen wir diesen zweiten Term oben noch etwas weiter aufspalten. Dazu brauchen wir wiederum den Trick, den wir oben bereits angewendet haben. Wir ziehen den Erwartungswert des geschätzten Modells $\mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]$ einmal ab und fügen ihn einmal hinzu:

$$\mathbb{E}\left[\left(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathbb{E}\left[\left(f(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right] + \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right] - \hat{f}(\mathbf{x}_0)\right)^2\right] \tag{4.75}$$

$$= \mathbb{E}\left[\left(f(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right] - \left(\hat{f}(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)\right)^2\right] \tag{4.76}$$

Ähnlich wie weiter oben können wir diese Gleichung mit einer polynomischen Expansion wie folgt umschreiben:

$$\mathbb{E}\left[(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0))^2\right] = \mathbb{E}\Bigg[\left(f(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)^2 \tag{4.77}$$

$$- 2\left(f(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)\left(\hat{f}(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right) \tag{4.78}$$

$$+ \left(\hat{f}(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)^2\Bigg] \tag{4.79}$$

Auch hier entfällt der mittlere Term, wenn wir den Erwartungswert in die Klammern reinnehmen, weil der zweite Teil $\left(\mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right] - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right) = 0$ ist. Was übrig bleibt ist folgendes:

$$\mathbb{E}\left[(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0))^2\right] = \mathbb{E}\left[\left(f(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)^2\right] + \mathbb{E}\left[\left(\hat{f}(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)^2\right] \tag{4.80}$$

$$= \left(f(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)^2 + \mathbb{E}\left[\left(\hat{f}(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)^2\right] \tag{4.81}$$

Schauen wir uns kurz die beiden Komponenten auf der rechten Seite etwas genauer an:

- $\left(f(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)^2$ ist der **quadrierte Bias** und misst die systematische Abweichung unseres geschätzten Modells $\hat{f}$ vom wahren unbekannten Modell f. Je kleiner der Bias, desto tiefer der erwartete quadrierte Fehler. Wir können diesen Term der Einfachheit halber mit $\left[\text{Bias}\left(\hat{f}(\mathbf{x}_0)\right)\right]^2$ bezeichnen.
- $\mathbb{E}\left[\left(\hat{f}(\mathbf{x}_0) - \mathbb{E}\left[\hat{f}(\mathbf{x}_0)\right]\right)^2\right]$ ist nichts anderes als die Varianz unseres geschätzten Modells $\hat{f}$. Sie misst, wie stark sich $\hat{f}$ im Schnitt verändert, wenn wir einen anderen Datensatz für das Training verwenden. Ein Modell mit hoher Varianz passt sich jeweils sehr stark an die Daten an. Je kleiner diese Varianz, desto tiefer der erwartete quadrierte Fehler. Wir bezeichnen diesen Term der Einfachheit halber als $\text{Var}\left(\hat{f}(\mathbf{x}_0)\right)$.

Nun sind wir endlich am Ziel angelangt und können den erwarteten quadrierten Fehler für die Beobachtung $(y_0, \mathbf{x}_0)$ wie folgt aufschreiben:

$$\mathbb{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \text{Var}(\epsilon) + \left[\text{Bias}\left(\hat{f}(\mathbf{x}_0)\right)\right]^2 + \text{Var}\left(\hat{f}(\mathbf{x}_0)\right)$$

Ein Modell mit **viel Bias** führt zu einer schlechten Vorhersagequalität (auf Trainings- und Testdaten), weil das Modell zu rigide ist, um den wahren Zusammenhang zwischen der Output-Variable und den Input-Variablen zu modellieren. Beispiel: wir verwenden ein einfaches lineares Regressionsmodell, um einen stark nicht-linearen Zusammenhang zwischen y_i und $\mathbf{x}_i$ zu modellieren. Im Fall von Modellen mit viel Bias spricht man auch von **Underfitting**.

Ein Modell mit **viel Varianz** führt zu einer hervorragenden Vorhersagequalität auf den Trainingsdaten, aber zu einer sehr schlechten Vorhersagequalität auf den Testdaten. Das Problem hier ist, dass das Model zu flexibel ist gemessen an der Grösse des Trainingsdatensatzes. Das Modell passt sich so zu stark an die Trainingsdaten an und modelliert auch sogenanntes **Noise** (und nicht nur das **Signal** in den Daten). Beispiel: wir modellieren ein neuronales Netzwerk, haben aber nur einen Trainingsdatensatz von einigen hundert Beobachtungen. Im Fall von Modellen mit viel Varianz spricht man auch von **Overfitting**.

Warum spricht man von einem **Tradeoff**? Flexiblere Modelle haben oft einen kleinen Bias, aber hohe Varianz, während unflexible Modelle oft eine kleine Varianz, aber einen hohen Bias haben. Es existiert also ein Tradeoff zwischen Bias und Varianz und wir wollen beim Modellieren und vor allem beim Hyperparameter Tuning den optimalen Tradeoff finden.

In unserem Beispiel wenden wir ein regularisiertes Regressionsmodell an. Hier spielt der Hyperparameter λ eine zentrale Rolle für den Tradeoff zwischen Bias und Variance. Ein zu tiefer Wert für λ kann zu einem zu flexiblen Modell mit viel Varianz führen. Ein zu hoher Wert für λ führt zu einem zu rigiden Modell mit viel Bias.

4.7 Polynomische Regression

Wir machen hier nun einen kurzen Abstecher in die **polynomische Regression**, denn diese eignet sich sehr gut, um den Bias-Variance Tradeoff zu illustrieren.

Ein **ganz wichtiger Punkt**: das polynomische Regressionsmodell ist immer noch **linear in den Parametern**, es handelt sich also immer noch um ein lineares Modell. Sie sehen aber an obigen Modellkurven, dass dieses "lineare" Modell sehr wohl in der Lage ist, nicht-lineare Zusammenhänge zwischen x und y zu fitten!

4.8 Lineare Regression in R

Base R vs. `tidymodels`

4.9 Weiterführende Themen

Bayesianische Regression

Grob gesagt rechnen wir ein ML-Modell in zwei Schritten. In einem **ersten Schritt** entscheiden wir uns für die funktionale Form unseres Modells $\hat{f}(\mathbf{x}_i)$. Man nennt dies in der Fachsprache **Model Selection**. Wir betrachten hier nur mal den vereinfachten Fall, in dem wir nur eine x_i-Variable pro Beobachtung als Input haben. Folgende Funktionen bzw. Modelle sind mögliche Kandidaten:

- $f(x_i) = b_0 + b_1 \cdot x_i$ (einfache lineare Regression)
- $f(x_i) = b_0 + b_1 \cdot x_i + b_2 \cdot x_i^2$ (polynomische Regression)
- $f(x_i) = \begin{cases} \bar{y}_1, & \text{falls } x_i > x^* \\ \bar{y}_2, & \text{sonst} \end{cases}$

Wir werden mit unserer Wahl der Funktion nie genau die wahre aber unbekannte Funktion $f(\mathbf{x}_i)$ treffen, aber wir versuchen möglichst nahe daran zu kommen.

"No Free Lunch" Theorem

Das *No Free Lunch* Theorem besagt, dass es kein universal bestes Modell gibt. Das heisst, dass es je nach Problem und Datensatz andere Modelle bzw. Funktionen braucht, um gute Vorhersagen zu machen. Das ist der Hauptgrund, warum wir Ihnen möglichst viele verschiedene Tools mit auf den Weg geben wollen.

Im Vergleich zur Summe der quadrierten Residuen haben wir hier noch den Faktor $\frac{1}{2n}$ drin. Dieser Faktor macht daraus eine Art Mittelwert und darum wird diese Kostenfunktion typischerweise **Mean Squared Error** (MSE) genannt.

Optional: Zerlegung des Vorhersagefehlers

Wir wollen hier kurz anschauen, wie der **Erwartungswert** des quadrierten Fehlers, $\left(y_i - \hat{f}(\mathbf{x}_i)\right)^2$, in zwei Komponenten zerlegt werden kann.

Dazu gilt folgendes:

- Von oben wissen wir, dass $y_i = f(\mathbf{x}_i) + \epsilon$ gilt.
- Wir nehmen an, dass der Erwartungswert des unsystematischen Teils ϵ Null ist, also $\mathrm{E}(\epsilon) = 0$.
- Allgemeine Regel zur Varianz einer Zufallsvariable: $\mathrm{Var}(\epsilon) = \mathrm{E}(\epsilon^2) - \mathrm{E}(\epsilon)^2 = \mathrm{E}(\epsilon^2) - 0^2 = \mathrm{E}(\epsilon^2)$.
- $\hat{f}$ und $\mathbf{x}_i$ sind fix und gegeben (keine Zufallsvariablen) und darum gilt $\mathrm{E}\left(\hat{f}(\mathbf{x}_i)\right) = \hat{f}(\mathbf{x}_i)$.

Nun können wir den **Erwartungswert** des quadrierten Fehlers rechnen:

$$\mathrm{E}\left[\left(y_i - \hat{f}(\mathbf{x}_i)\right)^2\right] = \mathrm{E}\left[\left(f(\mathbf{x}_i) + \epsilon - \hat{f}(\mathbf{x}_i)\right)^2\right] \tag{4.82}$$

$$= \mathrm{E}\left[f(\mathbf{x}_i)^2 - 2 \cdot f(\mathbf{x}_i) \cdot \hat{f}(\mathbf{x}_i) + \hat{f}(\mathbf{x}_i)^2 + 2 \cdot \epsilon \cdot f(\mathbf{x}_i) - 2 \cdot \epsilon \cdot \hat{f}(\mathbf{x}_i) + \epsilon^2\right] \tag{4.83}$$

$$= f(\mathbf{x}_i)^2 - 2 \cdot f(\mathbf{x}_i) \cdot \hat{f}(\mathbf{x}_i) + \hat{f}(\mathbf{x}_i)^2 + 2 \cdot \mathrm{E}(\epsilon) \cdot f(\mathbf{x}_i) - 2 \cdot \mathrm{E}(\epsilon) \cdot \hat{f}(\mathbf{x}_i) + \mathrm{E}(\epsilon^2) \tag{4.84}$$

$$= f(\mathbf{x}_i)^2 - 2 \cdot f(\mathbf{x}_i) \cdot \hat{f}(\mathbf{x}_i) + \hat{f}(\mathbf{x}_i)^2 + 2 \cdot 0 \cdot f(\mathbf{x}_i) - 2 \cdot 0 \cdot \hat{f}(\mathbf{x}_i) + \mathrm{Var}(\epsilon) \tag{4.85}$$

$$= f(\mathbf{x}_i)^2 - 2 \cdot f(\mathbf{x}_i) \cdot \hat{f}(\mathbf{x}_i) + \hat{f}(\mathbf{x}_i)^2 + \mathrm{Var}(\epsilon) \tag{4.86}$$

$$= \left(f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\right)^2 + \mathrm{Var}(\epsilon) \tag{4.87}$$

Der erste Teil auf der rechten Seite der Formel beschreibt den **reduzierbaren Fehler** und der zweite Teil den **nicht-reduzierbaren Fehler**. Wir sehen also auch hier: es ist sehr wichtig, dass wir eine Funktion $\hat{f}(\mathbf{x}_i)$ schätzen, welche dem wahren funktionalen Zusammenhang $f(\mathbf{x}_i)$ möglichst nahe kommt.

Chapter 5

Lineare Klassifikation

Chapter 6

Machine Learning Pipeline

Chapter 7

Decision Trees

Chapter 8

Ensembles

Chapter 9

Support Vector Machines

Chapter 10

Artificial Neural Networks

Chapter 11

Convolutional Neural Networks

Chapter 12

Recurrent Neural Networks

Chapter 13

Generative AI