Künstliche Künstler

Kann KI der Materie Geist einhauchen?

Inhaltsverzeichnis Anhang

Teil III: Anhang

9 Vertiefungen zur Mathematik

9.1 Grundfunktionen eines Computers

Im Folgenden will ich Sie vertraut machen mit den Grundfunktionen eines Computers. Die Basiselemente eines Rechners können entscheiden, ob auf zwei Inputleitungen gleichzeitig je ein Signal ankommt. Nur dann gibt es ein Signal als Output aus, sonst gibt es nichts aus. Die elektrische Schaltung realisiert ein «UND». Sie verfügt über eine logische Funktion, die man in mathematische Operationen umsetzen kann. Ein solches System kann zwei Zustände darstellen: 1 und 0 oder Ja und Nein. Mit diesen zwei Zuständen kann man bereits ein Zahlsystem konstruieren: 0 (gar nichts), 1 (eine Einheit), 10 (ein Zweier, null Einer. Denken sie nicht «zehn», sondern eins-null), 11 wäre dann drei. Für vier benötigt man schon eine neue Stelle: 100 (ein Vierer, null Zweier, null Einer). Betrachten wir als Beispiel die Addition von 5 und 3 im Zweiersystem: 101 + 11 =? (101 heisst: ein Vierer, null Zweier und ein Einer, 11 heisst: ein Zweier und ein Einer). Wir führen eine schriftliche Addition aus:

$$\begin{array}{r} 101 \\ +\ 11 \\ \hline 1000 \end{array}$$

Aus Erfahrung wissen wir, es gibt acht. Die digitale Grösse 1000 steht tatsächlich für 8: 1000 ist *ein* Achter, *null* Vierer, *null* Zweier, *null* Einer. Wie aber addiert ein Computer diese Zahlen? Betrachten wir die letzte Stelle, die Einer: Eine 1 plus eine 1 muss offensichtlich eine Null ergeben. Wieso? Denken Sie an einen Kilometerzähler: Wenn er bei der höchsten Ziffer ankommt, z.B. bei der 9, dann muss er auf eine neue Stelle umschalten: 9 + 1 = 10. Im Zweiersystem ist diese höchste Stelle 1. Das System muss schalten, wenn es 1 + 1 rechnen will, auf 10. Diese Operation kann man mit einem «UND» und einem «NICHT» bewerkstelligen: Das UND ergäbe zuerst eine 1 und das NICHT eine Null. Die beiden anderen Fälle (0 +1 und 1 + 0) ergeben eine 1. Falls das UND eine Null ergibt, muss das «Behalte» noch realisiert werden; dies kann erneut mit einem UND erreicht werden. Diese logischen Operationen können mit Verstärkern hardwaremässig verwirklicht werden. Früher waren diese Verstärker Röhren, heute sind es Transistoren. Die Erfindung miniaturisierter Verstärker mit Halbleitern, so genannter Transistoren, haben Computer in

der heutigen Leistungsfähigkeit überhaupt erst möglich gemacht. Die Erkenntnis, dass mit Transistoren logische Operationen verwirklicht werden können, gehört zu den grössten Leistungen der frühen Computerwissenschaft (Vgl. Domingos 2015, S. 2).

9.2 Wahrscheinlichkeiten

9.2.1 Wahrscheinlichkeit und Informationsbegriff

Bei dieser Vertiefung geht es um die Begriffe der Wahrscheinlichkeit und der Statistik: Eine Wahrscheinlichkeit ist eine Chance. Wir sprechen von der Wahrscheinlichkeit, beim Würfeln eine Sechs zu erzielen, und charakterisieren diese Chance mit einer Prozentzahl. Allgemein kann man eine Wahrscheinlichkeit definieren als die Anzahl günstiger oder zutreffender Fälle, geteilt durch alle möglichen Fälle:

$$\text{Wahrscheinlichkeit P} = \frac{günstige\ Fälle}{mögli\quad Fälle}$$

Wenn man nun die Wahrscheinlichkeit berechnen will, dass man bei zweimaligem Würfeln eine Doppel-Sechs erzielt, stehen zwei Wege offen: Man überlegt, wie viele zutreffende Fälle es gibt: nur einen. Dann denkt man über alle möglichen Fälle nach: 6 * 6 = 36. Somit ist die Wahrscheinlichkeit einer Doppelsechs

$$P(66) = \frac{1}{36}$$

Man hätte dieses Resultat auch durch eine logische Überlegung gewinnen können: Für eine Doppelsechs braucht es beim ersten Wurf eine Sechs UND beim zweiten Wurf erneut eine Sechs. Bei einer UND Verknüpfung muss man die Wahrscheinlichkeiten multiplizieren:

$$P(66) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$$

Wahrscheinlichkeiten werden dann multipliziert, wenn das Gesamtereignis seltener ist: Zwei Sechser zu würfeln, kommt seltener vor, als einen einzigen Sechser zu erzielen. Wenn eine Vier ODER eine Sechs zu einem Gewinn führen würden, dann vergrössern sich die günstigen Fälle: Die Wahrscheinlichkeiten werden addiert.

$$P\ (4\ \text{oder}\ 6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Nun fehlt nur noch ein kleiner Tick und Sie sind schon eine gewiefte Statistikerin: Wenn wir die Wahrscheinlichkeit aus der Analyse von Daten, z.B. dem Zählen von Wörtern in Texten, gewinnen wollen, dann schwankt die Genauigkeit des Resultates, je nachdem wie viele Wörter wir untersuchen. Wenn «gratis» in einem Textkorpus von einer Million Wörter 50-mal vorkommt, dann ist diese 50 genauer bestimmt, als wenn Sie Texte mit bloss 100'000 Wörter untersuchen und «gratis» 5-mal vorkommt. Der Fehler auf einer experimentell erhobenen zufälligen Grösse ist die Wurzel

der Zahl: der Fehler auf 5 wäre zirka 2.2, der auf 50 zirka 7. Obwohl beide Messexperimente eine Wahrscheinlichkeit von 0.5 Promille ergeben, ist die Messung an der einen Million Wörtern genauer als die Messung an den 100'000 Wörtern. Je grösser die Datenmenge, desto präziser werden die Wahrscheinlichkeiten. Falls Sie in der Schule Statistik gelernt haben und Sie diese Wurzel aus der Zählrate erstaunt: Sie beschreibt die Streuung der Daten, oft auch Sigma genannt. Sie haben gelernt, dieses Sigma sei $\sqrt{n * p * q}$. Dabei ist n die Anzahl Wörter, die Sie untersuchen, p ist die Wahrscheinlichkeit des Auftretens von z.B. «gratis» und q ist 1-p. Da p klein ist, ist q ungefähr 1 und da n * p die Zählrate für «gratis» ergibt, können Sie nachvollziehen, dass diese Faustregel eine schnelle Berechnung der Streuung erlaubt.

Man nennt diese Schwankung auf einer experimentell bestimmten Zahl wie gesagt die Streuung oder auch das Sigma. Im Unterkapitel 7.6.1 haben wir einen Spam-Filter vorgestellt, der in vielen Spam-Texten nach dem Wort «gratis» sucht, und sind davon ausgegangen, dass es in einer Million Wörter im Durchschnitt 50-mal vorkommt. Wenn man sehr viele Datensätze mit je einer Million Wörter untersuchen würde, dann würde «gratis» nicht immer genau 50-mal vorkommen, sondern einmal z.B. 53-mal und dann wieder 46-mal usw. 68 % aller Ergebnisse für «gratis» liegen aber zwischen 43- und 57-mal (50-7 resp. 50+7). Dies ist der so genannte 1-Sigma-Bereich.

Der Wahrscheinlichkeitsbegriff ist dem Informationsbegriff gleichgestellt (Lyre 2002, S.19): Beide messen den Überraschungseffekt beim Auftreten eines Ereignisses. Ein Ereignis mit kleiner Wahrscheinlichkeit, z.B. eine Doppelsechs, hat eine geringere Wahrscheinlichkeit als z.B. das Auftreten zweier gerader Zahlen. Je kleiner die Wahrscheinlichkeit, desto grösser die Überraschung.

9.2.2 Bayes-Theorem

Bei einem positiven Corona-Test erschrecken Menschen im Normalfall. Ein solches Testresultat kann als Illustration einer so genannten bedingten Wahrscheinlichkeit dienen. Nehmen wir an, der Test sei in 98 % der Fälle zuverlässig und in 2 % der Fälle gebe er ein falsches Resultat. Zudem gehen wir davon aus, dass 0.5 % der Menschen tatsächlich mit dem Coronavirus infiziert sind. Unter 1’000 Menschen wären das dann 5 Personen. Der Test würde diese 5 Personen identifizieren können. Er würde aber auch ca. 20 Personen als falsch-positiv bewerten: Sie sind gesund, haben aber ein positives Testresultat. Unter den 25 positiven Testergebnissen (5 tatsächliche, 20 falsch-positive), wären also nur 5 wirklich gefährliche. Die Wahrscheinlichkeit, von Corona infiziert zu sein, beträgt nur 20 %! Dies

nennt man eine bedingte Wahrscheinlichkeit: Die Wahrscheinlichkeit Corona zu haben unter der Bedingung, dass der Test positiv ist.

Im Folgenden versuchen wir, das erwähnte schwierige Theorem von Bayes zu verstehen. Wir betrachten Testverfahren, die zur Diagnose seltener Krankheiten eingesetzt werden. Dabei werden wir so genannte bedingte Wahrscheinlichkeiten analysieren und auf das Gesetz von Bayes stossen, auf dem die dritte Schule der KI wesentlich beruht. Wir illustrieren das Verfahren an einem weiteren Beispiel[1]: Ungefähr 1 Prozent der Männer über 60 Jahren leiden an Prostatakrebs. Der so genannte PSA-Test kann diese Krankheit diagnostizieren, allerdings nur mit 80 % Wahrscheinlichkeit. In 20 % bringt er ein falsches Resultat. Wir wollen die Situation zuerst mit einem Diagramm darstellen. Wir stellen uns ein Rechteck vor, dass 100 % der Männer über 60 enthält. Auf der horizontalen Achse tragen wir die Test-Effizienz (Nachweis) ein, auf der Vertikalen das Vorkommen der Krankheit in %. Das ganze Rechteck rechts zeigt die falschen Diagnosen an: Gesunde, die als krank identifiziert werden (schräg schraffiert: falsch-positiv) und Kranke, die als gesund wiedergegeben werden (grau: falsch-negativ). Wie grosse Angst muss

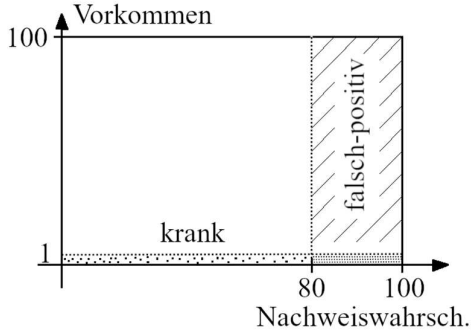


Fig. 9.2 a: Illustration des Bayes-Theorems

man nun haben, effektiv krank zu sein, wenn ein Mann ein positives Testresultat bekommen hat? Günstige oder zutreffende Fälle sind nur die im gepunkteten Bereich. Bei 100'000 Männern hätten 1'000 Prostatakrebs. Beim Test gäbe es 800 zutreffende (günstige) Fälle; 80 % von 1'000. Mögliche Fälle sind aber alle mit einem positiven Testresultat, also die 800 und noch alle falsch positiven: 20% von 99'000. Die Wahrscheinlichkeit tatsächlich krank zu sein ist nur

$$\mathrm{P}(\mathrm{krank} \mid \mathrm{positiv}) = \frac{800}{800 + 19800} = 3.9\ \%$$

[1] Vgl. Hesse Christian: *Achtung Denkfalle!* C.H. Beck, München 2011, S. 65.

Das ist ein erstaunliches und beunruhigendes Resultat. Wie viel Angst wird in einer Gesellschaft erzeugt, wenn man Männer über 60 flächendeckend testet, dabei ca. 20 % ein positives Resultat bekommen, aber nur ca. 4 % von diesen wirklich krank sind? Woran liegt dieses unverhoffte Resultat? Es liegt daran, dass so viele falsch-positiv sind. Das ist immer so, wenn die Auftretenswahrscheinlichkeit der Krankheit gering ist und der Test nicht 100-prozentig gut.

Man kann die Wahrscheinlichkeit als Anzahl zutreffende Fälle geteilt durch Anzahl möglicher Fälle mit Wahrscheinlichkeiten formulieren. Dazu teilt man den Zähler und den Nenner gleichzeitig durch die Gesamtzahl der untersuchten Fälle. Damit wird die Formel eleganter, aber auch undurchsichtiger.

$$\mathrm{P(krank\,|positiv)} = \frac{P(positiv/krank)*P(krank)}{P(positiv)} \qquad \text{(Bayes-Theorem)}$$

Ein Test ist eine Untersuchung an Daten, die uns klüger macht: Zuerst hatten wir nur zwei Fälle: krank (p(krank)= 1 %) oder nicht krank p(nicht-krank)= 99 %. Die beiden Wahrscheinlichkeiten nennt man Voraus-Wahrscheinlichkeiten oder a priori-Wahrscheinlichkeiten oder Basis-Wahrscheinlichkeiten. Nach dem Test sind wir klüger. Wir haben nun vier Fälle:

1. Gesund und negativer Test (so sollte es sein)
2. Gesund und positiver Test (dies dürfte nicht auftreten)
3. Krank und positiver Test (so sollte es sein)
4. Krank und negativer Test (auch das ist schlecht, aber unwahrscheinlich)

Diese Im-Nachhinein-Wahrscheinlichkeiten nennt man a posteriori-Wahrscheinlichkeiten. Bei bedingten Wahrscheinlichkeiten hat man drei Begriffe und zwei Fragestellungen:

- Basiswahrscheinlichkeiten (a priori)
- Bedingung (sie tritt mit 100 % auf)
- Bedingte Ereignisse (sie treten mit einer Wahrscheinlichkeit zwischen 0 und 100 % auf)

Dann gibt es *zwei* Fragestellungen oder zwei verschiedene bedingte Ereignisse. Denken Sie dabei immer zuerst an die Bedingung. Sie ist die Voraussetzung und mit 100 Prozent da:

1. Wie gross ist die Wahrscheinlichkeit, dass ein Mensch krank ist, wenn der Test positiv ist? (Test = positiv ist Bedingung)
2. Wie gross ist die Wahrscheinlichkeit, dass ein Test positiv ist, wenn der Mensch krank ist? (Krank ist die Bedingung).

Wir haben Bedingungen und wir haben bedingte Ereignisse. Viele Menschen verwechseln die Fragestellungen: Prüfen Sie sich selbst. Welche beiden Fragestellungen gibt es bei den folgenden Schlagzeilen?

1. „Fussballer sind die reinsten Bruchpiloten“ (aus dem Stern). Denn „sie verursachen fast die Hälfte der jährlichen rund eine Million Sportunfälle“.
2. „Frauen, hütet euch vor euren Ehemännern!“ Denn „die Hälfte aller ermordeten Frauen werden von ihrem eigenen Mann oder Liebhaber umgebracht“ (aus der Londoner Times).
3. „Schäferhunde sind gefährlich!“ (aus einer lokalen Tageszeitung). „Jeder dritte Biss geht auf das Konto dieser Rasse.“
4. „Zu viel Freizeit erzeugt Kriminalität“ (denn fast alle Vergewaltigungen, Einbrüche und Diebstähle finden ausserhalb der regulären Arbeitszeit der Täter statt). [2]

Der «Skandal» entsteht, weil die Fragestellungen verwechselt werden. Die Antwort passt dann nicht zur Frage: Wie viele Fussballer machen einen Sportunfall, ist die Fragestellung in der Headline. Die Antwort gilt aber für die Frage: Wie viele der Sportunfälle werden von Fussballern gemacht? Es stimmt, dass die Hälfte aller Sportunfälle von Fussballern verursacht werden. Das heisst aber nicht, dass jeder zweite Fussballer einen Sportunfall macht! Bedingte Wahrscheinlichkeiten bereiten uns allen grosse Schwierigkeiten. Ein berühmtes Beispiel ist das so genannte Ziegenproblem, das auch unter Mathematikern eindrückliche Kontroversen auslöste.

Das Bayes-Theorem wird in der KI für das Testen von Theorien an einem Datensatz verwendet. Dabei zeigt sich, dass die Möglichkeit, aus der einen Fragestellung die Antwort für die andere abzuleiten, oft einen grossen Vorteil darstellt. Beim Testen wird aus der Wahrscheinlichkeit, dass der Test positiv ist, darauf geschlossen, ob ein Mensch krank ist. Damit kann etwas über das Auftreten einer Krankheit in einer Bevölkerung gesagt werden, ohne dass man alle Menschen testen muss. Es genügt zu wissen, wie gut der Test ist – das kann man an relativ wenigen Personen überprüfen.

[2] 1.) Wie gross ist die Wahrscheinlichkeit, dass ein Fussballer einen Sportunfall hat? (Fussballer ist Bedingung). Wie gross ist die Wahrscheinlichkeit, dass ein Sportunfall von einem Fussballer verursacht wird? (Sportunfall ist Bedingung).
2) Wie gross ist die Wahrscheinlichkeit, dass bei einem Frauenmord der Ehemann der Täter ist? (Frauenmord ist die Bedingung). Wie gross ist die Wahrscheinlichkeit, dass ein Ehemann seine Frau ermordet? (Ehemann ist die Bedingung).
Die Beispiele stammen aus dem spannenden Buch von Krämer, Walter: *Denkste! Trugschlüsse aus der Welt des Zufalls und der Zahlen.* Campus, NY 1996.

9.2.3 Bayes Theorem bei vielen Merkmalen

Im realen Leben hat man oft nicht nur ein Merkmal, krank oder nicht krank, sondern mehrere. Berühmt ist das Beispiel von Mitchell,[3] ob der Tennisspieler Sam morgen spielen geht. Sein Verhalten auf Grund des Wetters wurde aufgezeichnet und ergab die folgende Liste:

Tag	*Ausblick*	*Temperatur*	*Luftfeuchtigkeit*	*Wind*	*Tennisspielen*
1	Sonnig	Heiß	Hoch	Schwach	Nein
2	Sonnig	Heiß	Hoch	Stark	Nein
3	Bewölkt	heiß	Hoch	Schwach	Ja
4	Regnerisch	Mild	Hoch	Schwach	Ja
5	Regnerisch	Kühl	Normal	Schwach	Ja
6	Regnerisch	Kühl	Normal	Stark	Nein
7	Bewölkt	Kühl	Normal	Strark	Ja
8	Sonnig	Mild	Hoch	Schwach	Nein
9	Sonnig	Kühl	Normal	Schwach	Ja
10	Regnerisch	Mild	Normal	Schwach	Ja
11	Sonnig	Mild	Normal	Stark	Ja
12	Bewölkt	Mild	Hoch	Stark	Ja
13	Bewölkt	Heiß	Normal	schwach	Ja
14	Regnerisch	Mild	Hoch	Stark	Nein

Fig. 9.2 b: Trainingsdaten des Tennisspielers Sam

Wenn es nun morgen sonnig ist und kühl und die Luftfeuchtigkeit hoch und der Wind stark ist, geht er dann spielen? Der Fall kommt in der Liste gar nicht vor und trotzdem können wir seine Wahrscheinlichkeit berechnen! Von den insgesamt 14 erfassten Tagen, geht er an 9 Tagen spielen. Seine a priori-Wahrscheinlichkeit ist also 9/14= 64 %.
Die bedingten Wahrscheinlichkeiten sind:

[3] Auffindbar z.B. in den Vorlesungsunterlagen von Andreas Schätzle: http://u-173-c140.cs.uni-tuebingen.de/lehre/ss06/pro_learning/AusarbSchaetzle.pdf (10.11.2020)

- P(stark/Ja)= 3/9 = 33 %
- P(mild/Ja)= 4/6 = 66 %

Nun könnte man die bedingte Wahrscheinlichkeit ausrechnen, dass er geht, wenn es z.B. sonnig und kühl, die Luftfeuchtigkeit hoch und der Wind stark ist: Das ist eine UND-Verknüpfung und wir müssen die Teilwahrscheinlichkeiten multiplizieren:
P(Ja)*P(sonnig/ja)*P(kühl/Ja)*..=0.0053
Ebenso für nicht gehen:
P(Nein)*P(sonnig/nein)*P(Kühl/nein)* = 0.0206
Wodurch muss man diese Wahrscheinlichkeit für den günstigen Fall teilen? Durch die Wahrscheinlichkeit für alle möglichen Fälle: das ist 0.0053, dass er geht und 0.0206, dass er nicht geht. Man nennt dies die so genannte Normalisierung. Sie ist nötig, da die Bedingung sonnig und kühl und hoch und stark im Datensatz nicht gleich oft vorkommt, wie z.B. irgendeine andere Bedingung.
P(Nein)=$\frac{0.0206}{0.0206+\ .0053} = 79.5\ \%$
Interessant ist, dass man mit diesem Vorgehen Fälle berechnen kann, die so in der Tabelle gar nicht vorkommen. Damit hat man ein Modell oder eine Hypothese, mit der man Sams Verhalten voraussagen kann auch in Fällen, die in den Trainingsdaten nicht enthalten sind.

9.2.4 Naiver Bayes Algorithmus, Information Retrieval

Wie gesagt, der künstliche Umgang mit Wissen, das in sprachlicher Form abgelegt ist, erleichtert den Menschen das Leben enorm. Heute wird wie selbstverständlich gegoogelt: Wissen aus Texten automatisch zu extrahieren, erachten viele Leute als selbstverständlich und denken, das hätte es schon immer gegeben. Dies ist ein grosser Irrtum: Nur Maschinen ermöglichen uns, auf das ganze Wissen der Menschheit innerhalb eines Augenblicks zuzugreifen. Unser künstlicher Künstler hat gewissermassen ein «World Brain», ein die ganze Welt umspannendes Gehirn, das auf alles gespeicherte Wissen zugreifen kann. Um die Leistungsfähigkeit eines solchen künstlichen Künstlers zu würdigen, wäre mindestens ein Grundverständnis seiner Suchstrategien nötig.

Wir haben die Frage «Wann war die Schlacht bei Waterloo?» als Suchanfrage gewählt. Wir sind uns gewohnt, dass ein künstliches System uns die relevanten Texte liefert, die Antworten auf die Frage enthalten. Wir haben ein naives Vorgehen propagiert: Man suche nach der Häufigkeit der Worte von «Schlacht», «Waterloo», «Wann» und «war» und wähle den Text aus, in dem die Häufigkeit dieser Worte am grössten ist. Dieses Vorgehen heisst tatsächlich naiver Bayes-Algorithmus. «Naiv» nennt man ihn, weil nur die Wörter, nicht aber die Wortfolge betrachtet werden. Ein

Satz wie «Schlacht von Waterloo, wann war sie?» würde ebenso einen Treffer ergeben, wie nicht korrekte Sätze: «Waterloo-Schlacht, wann war?» Es geht um eine so genanntes Informations-Suchsystem (Information Retrieval). Wir sollten einen Text den Begriffen «Schlacht», «Waterloo», «Wann» zuordnen können. Damit man die Funktionsweisen und auch die Tücken solcher Suchstrategien versteht, wählen wir wieder ein vereinfachtes Beispiel. Wir lassen uns von einem Exempel inspirieren, das auf der Internetseite monkeylearn.com vorgestellt wird.[4] Genaugenommen geht es dabei um eine Textklassifikation. Also die Zuordnung eines Textes zu einer Textsorte, wie Sport-Nachricht oder Spam.

Die Autoren fragen sich, ob die Aussage «A very close game» als Sport-Nachricht klassifiziert werden kann oder eher nicht. Dazu stellen sie Trainings-Texte zur Verfügung, die schon klassifiziert sind.

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Fig. 9.2 c: Lernsequenz für Textklassifikation

Sie sehen sofort, dass der Satz «A very close game» in keiner Nachricht vorkommt. Einzelne Worte kommen in den Sportnachrichten vor. Aber «close» erscheint nur in den Nicht-Sport-Nachrichten. Also was jetzt? Stellen Sie selber eine Vermutung an, zu welchem Nachrichtentyp der Satz gehört. Mit welcher Zuverlässigkeit können Sie Ihre Aussage treffen: Mit 80 % Wahrscheinlichkeit oder nur mit 60 %? Zuerst schauen wir nur, wie viele Sport-Texte es überhaupt gibt. Es sind drei von den fünf Texten.

$\mathrm{P(Sport)} = \frac{3}{5};\ \mathrm{P(Nicht-Sport)} = \frac{2}{5};$

Das sind die a-priori-Wahrscheinlichkeiten: Sporttexte kommen öfter vor. Nun geht es um die Wortfolge in Sporttexten. Diese Wahrscheinlichkeit berechnet sich als Produkt, da jedes Wort vorkommen muss. Es handelt sich um eine UND Verknüpfung. Für die Wahrscheinlichkeit, dass die Worte in einer Sportnachricht vorkommen gilt:

[4] Vgl. monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/

P(a very close game in Sporttexten) =
P(a)*P(very)*P(close)*P(game)*P(Sport)
Jetzt müssen wir mit Zählen beginnen: Sportnachrichten haben insgesamt 11 Wörter. «a» kommt zwei Mal vor:
$\mathrm{P(a)} = \frac{2}{11}$,
«very» kommt ein Mal vor
$\mathrm{P(very)} = \frac{1}{11}$
«close» kommt gar nicht vor!
Ein Wort, das im Text nicht vorkommt, zerstört unser ganzes Verfahren: Die Wahrscheinlichkeit wird null. Kreative KI-Informatikerinnen entwickelten dazu einen raffinierten Trick. Sie sagen: Jedes Wort, ob es vorkommt oder nicht, hat von Anfang an eine Grundwahrscheinlichkeit. Wir zählen alle Worte in allen Trainingstexten und erhalten insgesamt 14 verschiedene Wörter. Jedes dieser Worte hat eine Grundhäufigkeit von 1. Deshalb müssen wir die Anzahl Möglichkeiten um 14 ergänzen. Mit diesem Trick wird:
$\mathrm{P(a)} = \frac{2+1}{11+1}$; $\mathrm{P(very)} = \frac{1+1}{11+1}$; $\mathrm{P(close)} = \frac{0+1}{11+1}$; $\mathrm{P(game)} = \frac{2+1}{11+1}$
Nach etwas Rechnerei erhalten wir:
P(a very close game in Sporttexten) = 0.000046*P(sport) = 0.000028
Mit der analogen Berechnung für Nicht-Sporttexte erhält man:
P(a very close game in Nicht-Sporttexten) = 0.0000057
Das Auffinden von «a very close game» in Spotttexten ist ca. 5-mal häufiger als in Nicht-Sporttexten.

Verfeinerung des Naiven Bayes-Algorithmus

Es gibt Wörter, die wenig zur Bedeutung eines Satzes beitragen. Z.B: ein, einerseits, immer, selten usw. Sie können aus den Texten entfernt werden. Meist werden in einem Text auch Elemente der Grammatik entfernt wie die Mehrzahl, die Konjugation eines Verbs usw.: Election, elections und elected sind dann das gleiche Wort. Mehr in Richtung eines tatsächlichen Bayes-Modell geht der Einbezug von Wortpaaren. Beim Beispiel «Call the police» wurde das so gemacht: Wenn man «the» vor «police» betrachtet, dann kann man police genauer interpretieren und von anderen Wörtern besser trennen. Man hat dann eine bedingte Wahrscheinlichkeit; sie liefert mehr Information. Solche Techniken verwenden oft auch so genannte n-Grams. Da die Wörter selbst aus Buchstaben bestehen, könnte man alle 3-Buchstaben-Gebilde betrachten. «The» wäre ein 3-Gram. Ein solches n-Gram kann man als Markov-Kette interpretieren (siehe § 9.3). Mit n-Grams

kann man z.B. Sprachen sehr gut als Deutsch, Norwegisch usw. erkennen (Vgl. Russel 2012, S. 997).

Ähnlich wie die nicht bedeutungstragenden Wörter wirken Füllwörter (in, bei, ist). Sie zeichnen sich dadurch aus, dass sie in einem Text sehr oft vorkommen. Im Englischen tritt z.B. das 3-Gram «the» mit einer Wahrscheinlichkeit von ca. 2.7 % auf. Solche häufigen Wörter sind für eine Inhaltsanalyse unergiebig und können mit einer Strafe belegt werden, sodass sie Wahrscheinlichkeiten verringern (Vgl. Russel 2012, S. 1005).

Zusammenhang zum Bayes-Theorem

Wir stellten uns die Frage, wie gross die Wahrscheinlichkeit sei, einen Sporttext vor uns zu haben, wenn wir den Satz «a very close game» als Bedingung setzten. Weil wir diesen Satz in einem Text nur selten so finden, müssen wir uns behelfen, indem wir die Fragestellung umkehren und davon ausgehen, einen Sporttext vor uns zu haben (Bedingung) und nach der Wahrscheinlichkeit zu suchen, darin den erwähnten Satz zu finden (Ereignis). Formal sieht dann das so aus:

$$P(\text{sport} \mid \text{a very close game}) = \frac{P(a\ very\ close\ game/sport) * P(sport)}{P(a\ very\ close\ game;\ in\ allen\ Texten)}$$

Vorerst können wir den Nenner weglassen, weil er bei beiden Textsorten vorkommt und bei einem Vergleich daher keinen Einfluss hat. Dann müssten wir die bedingte Wahrscheinlichkeit des Satzes «a very close game» in allen Sporttexten evaluieren. Das ist oft beinahe unmöglich, weil der Satz sehr selten vorkommt. Es müssten riesige Textmengen analysiert werden. Der Naive Bayes-Algorithmus betrachtet nun nicht die Wortfolge, sondern bloss die Wörter einzeln und unabhängig: Er ist naiv, weil er keine Grammatik kennt und deshalb davon ausgeht, die Worte seien unabhängig voneinander. Das ist in einem normalen Text nicht der Fall. Trotzdem funktioniert der Algorithmus. Zudem kann er auch mit nicht vorhandenen Wörtern umgehen, indem er allen Wörtern eine Grundwahrscheinlichkeit zuschreibt, die dann um das reale Vorkommen der Wörter erhöht wird. Dieses Naive Bayes-Modell hat sich als erstaunlich leistungsfähig erwiesen (Vgl. Russel 2012, S. 589).

9.3 Hidden Markov Modelle (HMM)

Bei «Call the police» haben wir gesagt, hinter einer Folge von Wörtern stehe eine Grammatik. Man würde sie aber einer Software zur

Spracherkennung, einer so genannten NLP,[5] nicht einprogrammieren, sondern sie mit einem verborgenen Regelwerk – einem so genannten Hidden Markov Modell – simulieren. Wir wiesen darauf hin, diese Methode würde dem Problem gleichen, mit einem Würfel die Zahlenfolge 1 3 6 zu werfen. Wir wollen dieses Beispiel-Problem mit einer Markov-Kette lösen – obwohl das nicht nötig wäre, wie wir am Schluss sehen werden. Aber das Beispiel ist einfach und wir können sein Resultat mit logischen Schlussfolgerungen überprüfen. Ein Markov-Modell hat ein Start-Ereignis und einen Rand. Vom Start aus könnte man auf einen inneren Zustand, z.B. eine Eins ① kommen. Von

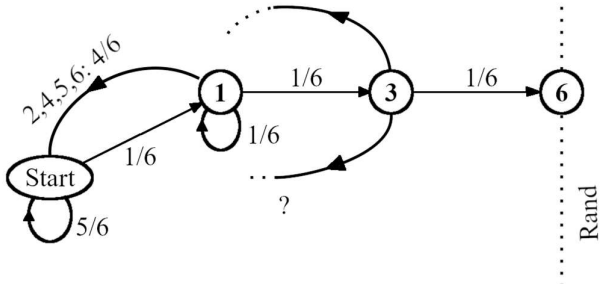


Fig. 9.3 a: Markov-Kette

dort könnte man weiter zu einer Drei ③ kommen und dann zu einer Sechs ⑥: Dann wäre man am Rand und der Markov-Prozess würde stoppen. Aber vom Start aus könnte man auch eine 4 würfeln. Dann sagt man, der Prozess kehre zum Start zurück, weil die gute Zahlfolge ja zwingend mit 1 beginnen muss. Beim inneren Zustand ① könnte man auch eine 4 würfeln, dann würde man wieder zum Start zurückkehren. Alle diese Pfeile kann man nun mit einer Wahrscheinlichkeit versehen: Vom Start mit 1/6 zu ① und mit 5/6 wieder zum Start zurück. Bei ① mit 1/6 zu ③, mit 1/6 zu sich selbst zurück und mit allen übrigen zum Start (4/6). Versuchen Sie selbst die Pfeile von ③ aus zu zeichnen und mit Wahrscheinlichkeiten zu versehen.

Nun kommt der grosse Clou der Markov-Kette: Man gibt die Wahrscheinlichkeit, dass man vom Start aus zum Rand kommt, einfach mit einer *unbekannten* Variablen an: p_s.[6] Den gleichen Trick wendet man auch

[5] NLP ist die Abkürzung für natural language processing, der maschinellen Verarbeitung von natürlicher Sprache.

[6] Als kluger Kopf haben Sie vielleicht schon gemerkt, dass diese Wahrscheinlichkeit 1 sein muss, weil es nur einen Rand gibt und man nach einer gewissen Zeit immer auf ihm landet. Aus didaktischen Gründen habe ich ein Anfangsbeispiel mit nur einem Randzustand gewählt.

für den Zustand ① an: Die Wahrscheinlichkeit, von ihm aus zum Rand zu kommen sei p_1. Das ist typisch für Mathematiker: Wenn sie nicht weiterwissen, bezeichnen sie das, was sie nicht wissen, mit einem Buchstaben, der stellvertretend für eine Zahl steht. So können wir also auch vom Zustand ③ aus mit p_3 zum Rand kommen. Nun stellen wir für z.B. p_s eine Gleichung auf. Vom Start aus gelangen wir mit 1/6 zu ① und von dort mit p_1 zum Rand.
Also scheint vorerst einmal $p_s = 1/6*p_1$ zu sein.
Aber man könnte ja auch wieder zum Start zurückkehren; mit 5/6 Wahrscheinlichkeit. Also ist
$p_s = 1/6*p_1 + 5/6*p_s$ (I)
Wenn Ihnen das etwas unheimlich vorkommt, ist das verständlich. Aber lassen Sie uns zuerst weitermachen, bevor wir zu viel denken. Von ① können wir mit 4/6 zum Start zurückkehren, mit 1/6 zu ③ kommen und mit 1/6 zu uns selbst zurückkehren.
$p_1 = 1/6*p_1 + 1/6*p_3 + 4/6*p_s$ (II)
$p_3 = 1/6*p_1 + 1/6*p_6 + 4/6*p_s$
Von ③ kommt man mit 1/6 zum Rand. Dann bricht der Prozess ab; man ist am Ziel. Die Wahrscheinlichkeit p_6, um von ⑥ zum Rand zu kommen ist 1 oder 100 %; man ist ja schon da. Also kann man für p_3 neu schreiben:
$p_3 = 1/6*p_1 + 1/6*1 + 4/6*p_s$ (III)
Dieses Gleichungssystem könnte man lösen. Aber wir verzichten darauf und überlegen uns, wie viele Schritte, oder Würfe, es im Durchschnitt braucht, bis man zum Rand kommt. Dazu verwenden wird wieder den Variablen-Trick. Vom Start aus brauche es m_s Schritte bis zum Rand. Man könnte vom Start aus einen Schritt tun und bei ① landen. Von dort hätte man dann im Mittel m_1 Schritte bis zum Rand. Also ist m_s vorerst:
$m_s = 1 + 1/6*m_1 + \ldots$
Aber damit sind wir noch nicht fertig: Man könnte ja mit 5/6 wieder zum Start gelangen. Also heisst die vollständige Gleichung:
$m_s = 1 + 1/6*m_1 + 5/6*m_s$ (A)
Dies sieht verdächtig nach Gleichung (I) aus, nur dass am Anfang eine 1 steht; weil man ja immer einen Schritt tun muss, um zu einem nächsten Zustand zu kommen.
$m_1 = 1 + 1/6*m_1 + 1/6*m_3 + 4/6*m_s$ (B)
$m_3 = 1 + 1/6*m_1 + 1/6*0 + 4/6*m_s$ (C)

Warum 1/6 mal 0? Wenn man am Rand ist, braucht man keinen Schritt mehr zu machen. Dieses 1/6*0 kann man natürlich weglassen, ich habe es nur aus didaktischen Gründen aufgeführt. Nun können wir das Gleichungssystem lösen. Meine Schülerinnen und Schüler würden es in ihren Taschenrechner eintippen und dann *solve* drücken. Wir könnten das m_s aus der ersten Gleichung (A) ausrechnen: Es gäbe $m_s = 6 + m_1$. Danach ersetzen wir in

allen folgenden Gleichungen (B, C) m_s. Dann lösen wir die zwei verbleibenden Gleichungen mit Addition. Wir bekämen für $m_1 = 210$ und für $m_s = 216$. Im Durchschnitt benötigen wir 216 Würfe, damit wir eine 1 3 6-Reihe bekommen. Einige von Ihnen werden sagen, diese komplizierte Rechnung hätten wir uns sparen können: Die Wahrscheinlichkeit eine 1 und dann eine 3 und dann eine 6 zu bekommen sei 1/6*1/6*1/6 und ergäbe 1/216. Das heisst, auf 216 mögliche Fälle gibt es einen günstigen. Falls es nicht nur *einen* Randzustand gibt, ist diese einfache Lösung nicht mehr möglich. Dann zeigt der Markov-Prozess seine Kraft, wie wir im Folgenden sehen werden.

9.3.1 Das Unendliche in den Griff bekommen

Was haben wir nun erreicht?

1. Wir sind fähig, bei einem im Prinzip unendlich langen Prozess die mittlere Schrittzahl zu berechnen.
2. Ebenso sind wir fähig, die Wahrscheinlichkeit zu berechnen, dass wir eine Reihe mit 1 3 6 bekommen.[7]
3. Wir lösen ein Problem damit, dass wir nur die unmittelbaren Nachbarn betrachten. Wir verfolgen die Gefechtsstrategie des Gruppenführers. Alle Pfeile, die von einem Zustand wegführen, ergeben zusammen 100%. Es sind nicht sehr viele Pfeile, höchstens 6.
4. Obwohl wir keinen Überblick haben, lösen wir das Problem.

Mit einer Markov-Kette können wir das Unendliche packen! Und: Mit einer Markov-Kette geraten wir nicht in die Komplexitätskatastrophe. Sie ist ein grosser Fallstrick in der KI. Was passiert, wenn wir die Markov-Kette verlängern und komplizierter machen? Wie wächst dann die Anzahl der Pfeile?

Wenn wir unsere Zahlenfolge vergrössern und z.B. sagen, wir wollen eine Serie mit 1 3 6 6 bekommen, dann addieren sich die maximal 6 Pfeile des Zustandes ⑥ zu allen bisherigen. Man sagt, die Pfeilzahl wächst linear mit der Anzahl Zustände (n). Die Pfeilzahl ist dann höchstens 6 * n. Wenn wir ein Netz hätten, bei dem alle Zustände mit *allen* anderen verbunden sein könnten, dann würde die Pfeilzahl mit n * n wachsen. Probieren Sie es mit einfachen Zahlen aus: n * n explodiert.

[7] Diese Wahrscheinlichkeit ist in unserem vereinfachten Falle 100 %. Normalerweise führt eine Markov-Kette aber zu mehreren Randereignissen, dann ist die Wahrscheinlichkeit für jedes Ereignis, das zum Rand führt, interessant.

9.3.2 Zusammenfassung: Verarbeitung natürlicher Sprache

Die Verarbeitung natürlicher Sprache hat sich zu einer eigenen Wissenschaft entwickelt, die erstaunliche Resultate aufweisen kann und in den Bereichen Textklassifikation, Informationsabruf, Informationsextraktion, maschinelle Übersetzung usw. einen Leistungsausweis erbringt, der vor wenigen Jahrzehnten für unmöglich gehalten wurde. Der Fortschritt ist wesentlich darauf zurückzuführen, dass man die ursprüngliche Fixierung auf Grammatik aufgegeben hat und das Vorurteil überwand, dass Wahrscheinlichkeits-Modelle wenig zur Verarbeitung natürlicher Sprache beitragen könnten (Vgl. Russel 2012, S. 1021).

9.4 Stellenwertsystem

Was ist die Bedeutung der Null? Betrachten Sie dazu römische Zahlen. Man kann sie nur im Kopf addieren: C-X = XC. Mit der Null bringen Sie eine zusätzliche Ordnung ins Zahlensystem: 100 – 10 = 90. Die Ziffer Eins hat nicht nur einen Zahlwert (1), sondern auch einen Stellenwert: Hunderter oder Zehner oder Einer. Die Bezeichnung dessen, was abwesend ist (0), erhöhte den Informationsgehalt der Ziffer. Die 9 bei 90, bedeutet nicht nur 9 Einheiten, sondern auch, es sind 9 Pakete, die selbst wieder aus 10 Dingen bestehen.

Etwas Abwesendes mit einem Symbol zu bezeichnen erachten wir bei der Null als selbstverständlich. Stellen Sie sich aber die indische Mathematikerin oder den arabischen Mathematiker vor, die das erste Mal sagten, man solle für das Nichts ein Symbol schreiben. Die Leute werden sie für verrückt gehalten und gesagt haben: «Nichts ist doch nichts, wozu dafür etwas schreiben?» Vielleicht hätten wir etwas mehr Verständnis für die Schwierigkeiten von Erstklässlerinnen und Erstklässlern beim Zehnerübergang, wenn wir die geistige Revolution würdigen könnten, die die Null darstellt.

9.5 Komplexe Systeme

9.5.1 Digitalisierungsproblem, Grenzwert

Das Problem des Zeitpunkts, zu dem Achilles die Schildkröte überholt, wird heute mit dem so genannten Grenzwert gelöst. Es hat mehr als 2000 Jahre gedauert, bis man diese Problematik mathematisch exakt fassen konnte (Bolzano um 1800). Die Lösung besteht darin, um die Zahl 11.1.. einen kleinen Gartenzaun mit Radius von sagen wir 0.05 zu spannen (von 11.10 bis 11.20). Dann definiert man, eine solche Zahl sei endlich, wenn *einige wenige* Zahlen ausserhalb liegen und alle

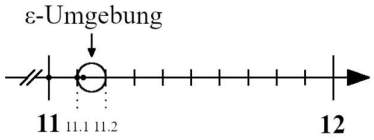


Fig. 9.5 a: ε-Umgebung

anderen (unendlich viele) innerhalb. Endlich heisst, Achilles kann die Schildkröte überholen. Der Gartenzaun heisst ε-Umgebung und die Zahl heisst Grenzwert. Die Zahlen laufen auf den Grenzwert zu. Die Folge der Zahlen ist so genannt konvergent. Die Zahl selbst kann man nicht genau angeben, man kann nur den *Prozess* beschreiben, der zu ihr führt.

9.5.2 Schmetterlingseffekt

Der Begriff Schmetterlingseffekt wurde in Zusammenhang mit der Wetterbildung geprägt. Sie ist ein Paradebeispiel für einen komplexen Prozess. Ein Gewitter über Zürich kann man im Prinzip gut erklären. Dass es aber exakt über Zürich um 15.03 abregnet, ist von sehr vielen Faktoren abhängig. Manchmal sogar auch von einem Schmetterling, der in Berlin aufgeflogen ist und dessen Flügelschlag eine kleine Welle auslöste, die ähnlich wirkte, wie die zerschlagene Flasche des Betrunkenen (aus § 2.1.3): deshalb der Begriff «Schmetterlingseffekt».

9.5.3 Erzeugen künstlicher Gesichter

Was haben die im Unterkapitel 2.2.1 abgebildeten Personen gemeinsam? Es gibt sie nicht – ein neuronales Netz hat die Gesichter künstlich erzeugt. Obwohl diese Personen uns vollständig natürlich erscheinen, existieren sie in der Realität nicht. Wenn Sie sich auf den Hintergrund konzentrieren, kann man bei einigen Darstellungen seltsame Formen erkennen, sie sind oft ein Hinweis auf die Künstlichkeit. Ein Computer hat auch Schwierigkeiten,

die Bilder als künstlich oder natürlich einzuordnen. Allerdings verblüffen seine Kriterien oft (Vgl. Unterkapitel 8.4.2, der Vergleich der zwei Schulbus-Bilder).

9.5.4 Anzahl Verbindungen

Die Berechnung der Anzahl Verbindungen kann man sich folgendermassen überlegen: Das erste Neuron kann im Prinzip mit jedem Neuron der zweiten Ebene eine Verbindung haben. Ein Neuron der zweiten Ebene ebenso. Es gäbe 300‘000 * 300‘000 * 300‘000….je nach Anzahl Ebenen: und dies 300‘000-mal – für jedes Neuron der Eingangs-Ebene. Damit hätte ein KNN für Bildverarbeitung mit hoher Auflösung sehr schnell viel mehr Verbindungen als das Alter des Universums, gemessen in Sekunden. Es versteht sich von selbst, dass die Neuronenzahl reduziert werden muss.

9.6 Komplexe Zahlen

Zu ihrem Verständnis wäre es günstig, sich die normalen Zahlen auf einem Strahl als Pfeile vorzustellen: Die positiven finden sich auf der rechten Seite von Null, die negativen auf der linken. Denken wir uns die Zahl Drei: Sie sei ein Pfeil, der von Null bis + 3 reicht. Nun multiplizieren wir diesen Pfeil mit -1. Das ergibt vorerst -3 oder als Pfeil gedacht, einen Zeiger, der von Null auf -3 zeigt. *Diese Multiplikation mit -1 entspricht also der Drehung des Pfeiles um 180°.*

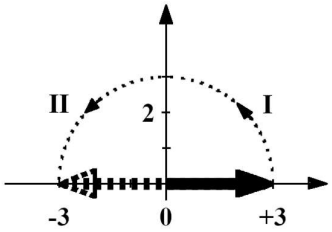


Fig. 9.6 a: Eine Multiplikation mit -1 ist eine Drehung

Jetzt kommt der Clou: Wir teilen diese Drehung in *zwei* Schritte auf: zuerst um 90 ° (I) und dann nochmals um 90° (II). Denken Sie nicht darüber nach, warum man das macht. Wir nennen diese Drehung z.B. q. Wenn wir q auf den + 3-Pfeil anwenden, dann zeigt der Pfeil senkrecht nach oben und ist q*3. Jetzt drehen wir nochmals um 90°, also nochmals um q und erhalten -3: oder q*q*3. Der Vergleich ergibt, dass q*q = -1 sein muss. Damit haben

wir aber etwas «Unerlaubtes» produziert: *Wir fabrizieren eine Zahl q, deren Quadrat -1 ergibt.* Das ist innerhalb der normalen Zahlen nicht erlaubt, und deshalb gibt man dieser Zahl einen neuen Namen. Sie heisst in der Mathematik normalerweise «imaginäre Einheit i» und bedeutet: Drehung um 90 °.

Denken Sie nochmals an die erste Drehung. Das Resultat ist ein Pfeil, der

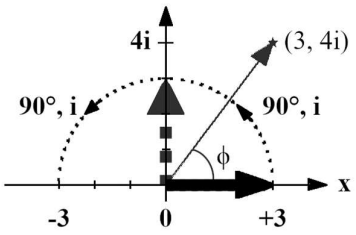


Fig. 9.6 b: Drehung um 90°

nach oben zeigt und die Länge q*3 oder 3i hat. Es gibt also neu eine senkrechte Zahlengerade mit den Einheiten 1i, 2i, 3i, 4i … Sie spannt zusammen mit der horizontalen Achse (x) der normalen Zahlen die Ebene der so genannten komplexen Zahlen auf. Wir könnten nun einen beliebigen Punkt (*) in dieser neuen Zahlenebene anschauen: Er hat einen horizontalen-Wert, der eine normale Zahl darstellt (3) und einen senkrechten Wert, der eine i-Zahl darstellt (4i). Man kann eine solche Zahl als Paar (3, 4i) schreiben. Den Pfeil vom Ursprung zu diesem Punkt könnte man aber auch mit seiner Länge r und seinem Winkel (ϕ) zur x-Achse beschreiben. Das ist der Grund, wieso man Pfeile mit so genannten komplexen Zahlen darstellen kann. Die Quantenmechanik fusst nicht auf diesen komplexen Zahlen, wie viele Laien meinen. Komplexe Zahlen sind nur ein praktisches Verfahren, um den Wellencharakter darzustellen. Die Addition von Funktionen mit komplexen Zahlen stellt deshalb automatisch unsere Pfeiladdition und damit die Überlagerung von Wellen sicher.

Leserinnen und Leser, die schon einmal etwas von Sinus und Cosinus gehört haben, können nun noch einen Schritt weitergehen. Die Projektion des Pfeiles auf die horizontale Achse ergibt eine Cosinusfunktion. Wenn wir einen Papierstreifen vorbeiziehen würden, ergäbe die Projektion eine Cosinus-Welle. In der Mathematik drehen Pfeile im Gegenuhrzeigersinn und sie beginnen auf der horizontalen Achse mit dem Winkel $\phi = 0$.

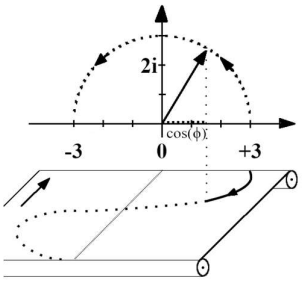


Fig 9.6 c: Projektion ergibt Cosinus-Welle

Ein Pfeil, z.B. |R> für Reflexion an der Rückseite, kann nun geschrieben werden als:

$|R> = \cos(\phi) + i*\sin(\phi)$

Wenn seine Länge nicht 1 ist, würde man noch einen Faktor r vor diese Formel setzen.

Haben Sie Lust zu einem weiteren Challenge? Der folgende Abschnitt ist im ersten Anlauf schwierig zu verstehen. Er beruht darauf, dass es in der Mathematik nur eine *einzige* Funktion gibt, die abgeleitet sich selbst reproduziert. Diese Summe von Cosinus- und Sinusfunktion ist eigentlich eine Exponentialfunktion. Wir schreiben vorerst:

$e^{i\phi} = \cos(\phi) + i*\sin(\phi)$

Sie können diese Aussage mit der Methode des Ableitens überprüfen: Leiten Sie die linke und die recht Seite je separat ab und vergewissern Sie sich, dass wieder eine Sinus- und eine Cosinusfunktion auf der rechten Seite und bloss eine neue e-Funktion auf der linken Seite entstehen. Stellen Sie sich unter dieser Funktion einen rotierenden Pfeil vor – denken Sie nicht an die normale Exponentialfunktion, die steil ansteigt. Dieses $e^{i\phi}$ ist die mathematische Darstellung des Pfeiles in der so genannten Polarform der komplexen Zahlen. Er hat die Länge 1. Wenn er eine andere Länge haben soll, dann kann man ihn wie gesagt mit r multiplizieren. Nun bleibt uns nur noch die Aufgabe, den Winkel in Abhängigkeit von der Zeit (t) zu bestimmen. Normalerweise geben Physikerinnen und Physiker Winkel nicht im Gradmass, sondern im Bogenmass an. Sie sagen, der volle Winkel (360°) sei der Umfang eines Kreises mit Radius 1. Also entspricht ihm der

Wert 2π. Neunzig Grad entsprächen dann 2π/4 usw. Mit dem Begriff Periodendauer (T) bezeichnet man die Zeit, die der Pfeil braucht, um sich ein ganzes Mal zu drehen. Deshalb ist der Winkel in Abhängigkeit von der Zeit t:

$\varphi(\mathrm{t}) = \frac{2*\pi}{T} * t = \omega * t$

Die Grösse ω nennt man Kreisfrequenz. Sie gibt an, wie viel Strecke die Pfeilspitze auf der Peripherie des Einheitskreises in einer Sekunde überstreicht. Nun haben wir eine imposante Form der Wellenfunktion gefunden:

$|R> = r*e^{i\omega t}$

Das Vorzeichen des Exponenten ist willkürlich: Bei Zeiten wählt man üblicherweise ein negatives Zeichen.

$|R> = r*e^{-i\omega t}$

Halten Sie sich vor Augen: Diese schwerverständliche Form ist nichts anderes als ein Pfeil der Länge r, der sich mit der Zeit dreht. Wenn ω gross ist, dreht er schnell, sonst dreht er langsam.

10 Vertiefung Physikalischer Konzepte

10.1 Einführung in die Quantenmechanik

10.1.1 Quantenmechanik und Realität

Wie hängt das Modell der Quantenmechanik mit der Realität zusammen? Man sollte ein Berechnungs-Modell nicht mit der Realität verwechseln. Das ist für die moderne Wissenschaft selbstverständlich. Zudem sind viele Physikerinnen und Physiker der Meinung, der Quantenmechanik-Formalismus sei ein unglückliches Modell: Er konstruiere zuerst ein kompliziertes System von Funktionen, die vom Minus-Unendlichen bis zum Plus-Unendlichen reichen und eine Unmenge von Information enthalten. Wenn man dann etwas berechnen wolle, schlage man alle schönen Funktionen zusammen und reduziere sie auf eine einzige Zahl. Ein solches Modell sei unnötig kompliziert. Es ist aber in der Form der so genannten Quanten-Elektro-Dynamik das Modell mit der grössten Genauigkeit in der Physik. (Vgl. Feynman 1985, S.17).

10.1.2 Quantenmechanik: Einführung in den Formalismus

Gemäss meiner Erfahrung bleibt Quantenmechanik (QM) für viele Studierende ein Mysterium. Das ist schade, weil dadurch keine adäquate und breite Diskussion der wesentlichen Erkenntnisse zu Stande kommt. Zudem ist diese so genannte QM und die darauf aufbauende Quantenelektrodynamik (QED) die beste Theorie, die wir in der Physik überhaupt haben: Mit einer ihr adäquaten Genauigkeit könnte man die Strecke Los Angeles – New York bis auf eine Haaresbreite präzise vorhersagen. Wir wollen deshalb eine knappe Einführung in die grundlegenden Ideen vorstellen. Meiner Meinung nach ist die Erörterung von Richard Feynman in seinem Büchlein QED (Feynman 1985) noch immer vorbildlich – auch als didaktisches Lehrstück. Wir orientieren uns wieder an unserem Photon, das auf eine Scheibe zufliegt und durch sie hindurchgeht oder an ihr reflektiert wird. Wir haben gesagt, das geschehe im Maximum mit 16 % und im Minimum mit 0 %. Die Scheibe kann so dick sein, wie sie will, die Wahrscheinlichkeit schwankt immer zwischen diesen beiden Werten (Feynman 1985, S.33). Feynman sagt, wir müssten ein Modell bauen, um diese beiden Wahrscheinlichkeiten prognostizieren

zu können. In seiner anschaulichen Sprache redet er davon, wir müssten einen Mechanismus haben, um «die Bohnen zu zählen».

Feynman schlägt vor, jedem Photon einen Pfeil in der Form des Zeigers einer Uhr mitzugeben.[8] Er hat ursprünglich die Länge 1, startet bei 12-Uhr, wenn das Photon losfliegt, und dreht sich. Das Lichtteilchen kann an der Vorderseite der Scheibe zurückgeworfen werden oder an der Hinterseite.[9] Bei einer Reflexion wird der Zeiger gestaucht und bei der Reflexion an der Vorderseite springt er zusätzlich um 180°.[10] Wenn nun ein Ereignis wie die Reflexion auf zwei verschiedenen Wegen zu Stand kommen kann – durch Reflexion an der Vorder- (Pfeil 1) *oder* Rückseite (Pfeil 2) – werden die

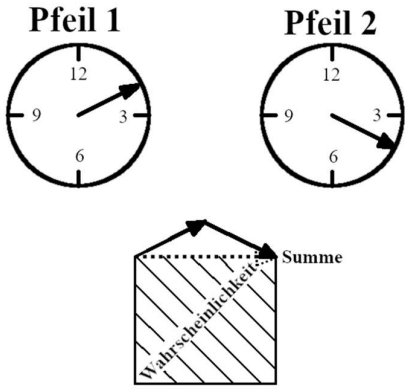


Fig.10.1 a: Pfeiladdition nach Feynman

[8] Für Menschen, die den Formalismus schon kennen: Er stützt sich bekanntlich stark auf das Konzept der komplexen Zahlen. Feynman umgeht diese Schwierigkeit mit dem Bild des Pfeiles. Wir werden den Zusammenhang zu den komplexen Zahlen später herstellen.

[9] Für eine kritische Leserin mag diese Einschränkung als unzulässig erscheinen. Wenn das Modell einmal steht, kann man zeigen, dass alle möglichen Reflexionen *innerhalb* des Glases sich gegenseitig aufheben (Feynman 1985, S. 118 ff). Anschaulich kann man dies an einem See nachvollziehen, der keine «Rückseite» aufweist. Bei ihm spiegelt nur die Oberfläche.

[10] Wie wir sehen werden, dient der Zeiger zur Darstellung einer Welle. Eine solche Welle könnte z.B. mit einem Seil erzeugt werden, das man an der einen Seite auf und ab bewegt und das an der anderen Seite an einer Wand befestigt ist. Dabei kann das Seil fix mit der Wand verbunden sein, oder nur lose, weil zwischen Wand und Seil z.B. ein Gummiband gespannt ist. Bei der fixen Befestigung kommt ein Berg als Tal zurück, während bei der losen Befestigung der Berg als Berg reflektiert wird. Eine Welle, die von Luft auf Glas trifft, erfährt die Reflexion an einer fixen Befestigung: Der Zeiger spring um 180°, z.B. von einem Berg zu einem Tal.

beiden Pfeile addiert. Zur Ermittlung der Wahrscheinlichkeit für eine Reflexion wird die Länge des resultierenden Pfeiles quadriert.

Nun sind Sie bereits in der Lage, wie eine Quantenmechanikerin zu rechnen: Z.B. Was passiert bei einer sehr dünnen Scheibe? Nach dem Entstehen fliegen beide Pfeile gleich lang, bis sie auf die Scheibe treffen. Sie schauen also unmittelbar vor der Scheibe in die gleiche Richtung. Nun wird der Pfeil für die Reflexion an der Vorderseite um 180 ° gedreht und schaut deshalb in die Gegenrichtung des Pfeiles, den erst die Hinterseite reflektiert. Dieser ist wegen der dünnen Scheibe aber unwesentlich weiter geflogen und die Addition der beiden Pfeile ergibt einen Pfeil, der praktisch null ist.

Was passiert, wenn die beiden Pfeile nach der Reflexion in die gleiche Richtung schauen? Der resultierende Pfeil wird doppelt so gross. Es entsteht die maximale Wahrscheinlichkeit. Wie dick muss die Scheibe sein, damit dies eintritt? Der Hinterseiten-Pfeil muss die Scheibe zwei Mal durchqueren, er muss sich dabei aber nur um 180 ° drehen. Wenn die Pfeile für eine ganze Drehung eine Strecke der Länge λ im Glas zurücklegen, dann muss die Scheibe also nur eine Dicke von ¼ λ aufweisen.

Wie stark muss der Pfeil gestaucht werden? Wir haben gesagt, die maximale Reflexion sei 16 %. Das Quadrat des zusammengesetzten Pfeiles ergibt dann 0.16. Die Länge des resultierenden Pfeiles ist daher die Wurzel aus 0.16: also 0.4. Die Länge entsteht durch die Addition von zwei Pfeilen, die in die gleiche Richtung schauen. Ein einzelner Pfeil ist somit 0.2 lang. Er muss daher beim Auftreffen auf 0.2 gestaucht werden.

Wenn die Scheibe dicker wird, kommt das Konzept des Pfeiles voll zum Tragen. Nehmen wir an, ein Pfeil stehe auf 3-Uhr. Dann wissen Sie nicht, wie oft er sich schon rundherum gedreht hatte. Das kann *keinmal* sein, *einmal*, *zweimal* usw. Deshalb wird die Platte immer nach ¼ λ an Dickezunahme wieder das gleiche Verhalten zeigen: Der Vorgang ist periodisch.

10.1.3 Entanglement: Interferenz der Einzeleffekte

Durch die Pfeiladdition sind die Reflexion an der Vorderseite und die an der Rückseite untrennbar miteinander verheiratet. Man kann die Reflexionen nicht einzeln bestimmen und dann die Resultate addieren: Mit den Einzeleffekten alleine würde man nie eine Null-Wahrscheinlichkeit bekommen und die Reflexion an der Vorder- oder Rückseite wäre genau 4%. Man sagt deshalb, die beiden Ereignisse würden miteinander interferieren, sie seien entangelt: Sie verstärken sich, aber sie löschen sich auch aus! Einen solchen Effekt kann man nur durch eine Überlagerung der

beiden Pfeile für die unterschiedlichen Wege erreichen. Üblicherweise schreibt man den Pfeil für die Reflexion an der Vorder- oder Rückseite mit einem eigenartigen Symbol: |V> oder |R>. Zudem zieht man die *Länge* des Pfeiles vor das Symbol und verwendet *zwei* Buchstaben, weil die Reflexionen nicht notwendigerweise gleichwahrscheinlich sein müssen:
$|\psi> = a*|V> + b*|R>$
Wobei wir die allgemein übliche Konvention verwenden, dass ein quantenmechanischer Zustand mit dem griechischen Buchstaben ψ gekennzeichnet wird. Wenn man entscheiden will, ob das Photon an der Vorderseite reflektiert wurde oder auf der Rückseite, dann muss man eine Messung machen. Diese Messung zerstört aber das Entanglement; sie hebt die Fähigkeit zum Verstärken oder Auslöschen aus.

Für unsere Überlegungen zur Frage, ob man die Prozesse der realen Welt vollständig auf einem Computer simulieren könnte, ist diese Aussage von entscheidender Bedeutung. Die Messung spielt in unserem Bild die Rolle des Tunnels für das Gewinnen von Information aus der Natur. Der Einfluss dieses Tunnels ist im Falle der Quantenmechanik so radikal, dass er das System selbst zerstört. Aus diesem Grund sind heute viele versierte Theoretikerinnen und Theoretiker der Meinung, man könne die Quantenwelt – und damit einen wesentlichen Teil unserer Welt – nicht vollständig auf einem gewöhnlichen Computer simulieren. Dazu bräuchte es einen Quantencomputer, der das Entanglement handhaben kann, ohne es zu zerstören. In der Theorie funktioniert ein solcher Quantencomputer – in der Praxis sind die Hürden bisher noch enorm hoch.[11]

Wie hängt das Pfeilmodell mit dem Wellencharakter von Licht zusammen? Stellen Sie sich vor, sie würden den Zeiger einer Uhr von oben her auf den Boden projizieren. Sein Schatten hätte die Länge 1, wenn er auf 3 Uhr steht. Er würde sich auf 0 verkürzen, wenn er gegen 6 Uhr fortschreitet und würde bei -1 umkehren, wenn der Zeiger 9 Uhr passiert. Wenn Sie auf dem Boden ein Papier vorbeiziehen würden, dann wäre auf ihm eine schöne Wellenlinie aufgezeichnet (vergleiche Fig. 9.6 c). Der Pfeil beschreibt also eigentlich eine Welle. Zwei Pfeile entsprechen zwei Wellen. Wenn diese ungünstig gegeneinander versetzt sind, dann löschen Berge und Täler sich aus. Wenn sie gut passen, addieren sich Berg und Berg wie auch Tal und Tal. Diese so genannte Interferenz ist ein typisches Wellenphänomen. Der Pfeil, der dem Photon mitgegeben wird, stellt dessen

[11] Eine gut lesbare Einführung in Quantencomputer geben Michael Nielsen und Andy Matuschak: Andy Matuschak and Michael A. Nielsen, "Quantum Computing for the Very Curious", auffindbar unter: https://quantum.country/qcvc, San Francisco (2019).

Wellennatur sicher. Allerdings ist weder der Pfeil noch die hier beschriebene Welle in der Quantenmechanik ein reales Objekt: Sie sind bloss ein Modell, um die Bohnen zu zählen. Wenn Sie einmal diese Grundideen verstanden haben, dann sollte die Formalisierung kein allzu grosses Problem mehr sein. Eine Hürde stellt noch die mathematische Darstellung eines Pfeiles dar. Auch hier wollen wir Ihnen einen Zugang zur Welt der komplexen Zahlen öffnen – eine Welt, die leider den meisten Menschen Zeit ihres Lebens verschlossen bleibt.

10.1.4 Darstellung des Pfeils mit komplexen Zahlen

Im Unterkapitel 9.6 bei den mathematischen Ergänzungen führe ich in die so genannten komplexen Zahlen ein. Sie stellen Drehungen dar. Wenn Sie diese Objekte nicht kennen, oder sie in Ihrem Studium nie verstanden haben, lesen Sie bitte zuerst dort nach, bevor Sie weiterfahren. Mit komplexen Zahlen kann man den Pfeil der Wellenfunktion in einer eindrücklichen Form darstellen:

$|R\rangle = r * e^{-i\omega t}$

Sie ist nichts anderes als ein rotierender Pfeil mit der Länge r – den es in der Realität so nicht gibt. Er dient uns bloss als Modell, um Wahrscheinlichkeiten auszurechnen: weil die Natur geizig ist und uns bei kleinen Dimensionen nicht genügend Information liefert. Ein Photon nach der Reflexion könnte man dann beschreiben mit:

$\psi(t) = a * e^{-i\omega t_V} + b * e^{-i\omega t_R}$

Wobei t_V die Zeitdauer bezeichnet für eine Reflexion an der Vorderseite und t_R diejenige für eine Reflexion an der Rückseite. Was heisst nun, die Natur sei geizig? Bei einem einzelnen Photon liefert sie uns keine Information über die Grösse der beiden Summanden. Wir wissen also bei einem einzelnen Lichtteilchen nicht, wie die Reflexion an der Vorderseite mit der an der Rückseite gemischt ist, oder mit anderen Worten, ob das Photon vorne oder hinten reflektiert wurde. Die beiden Grössen $a * e^{-i\omega t_V}$ und $b * e^{-i\omega t_R}$ werden deshalb oft als verborgene Variablen bezeichnet. Die Natur liefert uns keine Angaben über deren Werte für ein *einzelnes* Teilchen – für viele aber schon. Da wir die Grösse der beiden Summanden nicht bestimmen können, wissen wir auch nicht, wie sich dieses $\psi(t)$ mit der Zeit entwickelt. Das ist höchst ungewöhnlich (Vgl. Friebe 2018, S. 34). Damit haben wir auch bei der Information eine Art Quantisierung: Wir bekommen nicht beliebig viel Information. Die Menge der Information ist endlich und

beschränkt. Der Mechanismus dieser Beschränkung ist hochinteressant und noch lange nicht abschliessend verstanden.[12]

Mit der Darstellung dieser Überlagerung, auch Superposition oder Entanglement genannt, können Sie nun der Diskussion in der so genannten Quantenphilosophie folgen: Sie wissen jetzt, was es bedeutet, wenn jemand sagt, die Wellenfunktion könne man gar nicht messen. Sie verstehen, dass der Geiz der Natur uns nichts sagt über die beiden Teile der Wellenfunktion. Sie sagt uns nur etwas über das Quadrat von deren Summe. Das ist das Resultat einer Messung. Viele Messungen geben dann einen Wert für a und einen für b. Sie verstehen auch, wieso man vom «Kollaps der Wellenfunktion» spricht, wenn man das Quadrat bildet. Dann kollabiert das ganze schöne $b * e^{-i\omega t_R}$ auf die Zahl b^2: Das ω und das t verschwinden.

Viele Studierende lernen die Quantenmechanik über die so genannte Schrödingergleichung kennen. Mit unserer Herleitung stehen wir unmittelbar vor ihr: Zuerst denken wir daran, dass die Teilchennatur des Lichtes 1905 von Mileva Marić und Albert Einstein begründet wurde mit ihrem wegweisenden Postulat, die Energie eines Teilchens sei eine Konstante ($\hbar$) mal die Kreisfrequenz (ω).

$E = \hbar * \omega$

Dies Gleichung begründet die Doppelnatur von kleinen Teilchen: Sie haben Teilchen- (E) und Wellencharakter (ω). Damit können wir nun ω durch die Gesamtenergie E ersetzen und erhalten für unseren rotierenden Pfeil:

$$\psi(t) = r * e^{-i\frac{E}{\hbar}*t}$$

Für diese Funktion ergibt sich aus der Ableitung der folgende Zusammenhang, der als Schrödingergleichung bezeichnet wird:

$$i\hbar\frac{d}{dt}\psi(t) = E * \psi(t)$$

Die Schrödingergleichung beschrieb ursprünglich real existierende Wellen. In der sogenannten Kopenhagener Deutung wurde das ψ später als Wahrscheinlichkeit, genau genommen als Pfeil oder Wahrscheinlichkeits-Amplitude, interpretiert. Ich finde den Zugang von Feynman überzeugender: Er sagt unmissverständlich, wir hätten bei der Quantenmechanik völlig neue Phänomene vor uns, und dazu bräuchten wir ein neues Modell – um die Bohnen zu berechnen. Dieses Modell ist vorerst

[12] Wir leben in einer spannenden Welt: 1993, ca. 60 Jahre nach der Entdeckung der Quantenmechanik, wies man experimentell nach, dass man den Zustand |ψ> übertragen kann – obwohl man ihn wie gesagt nicht vollständig kennt. Der Empfänger hat ihn dann und kann sogar mit Sicherheit sagen, ob ihn bei der Übertragung jemand lesen wollte oder nicht. Der Sender zerstört ihn beim Übertragen: Er muss einen Teil von ihm messen. Diese Entdeckung hat unter anderem zum Konzept von Quantencomputern geführt und ist ein Gebiet, das für uns noch einige Überraschungen bereithält.

einmal in unserem Kopf; es muss nicht in der Wirklichkeit realisiert sein. Natürlich hantieren wir Physiker mit Begriffen, die abschreckend wirken und viele Menschen vom Mitdenken abhalten. Wir wollen einige davon benennen und sie mit dem bisher Erklärten verbinden.

Hilbertraum: Es ist der Raum der Zustände, die wir z.B. mit |R> oder |V> bezeichnet haben und für die wir die Beschreibung mit Pfeilen einführten. In diesem abstrakten Raum wirken *Operatoren.* Sie sind eine Art von Maschinen, die diese Zustände umformen. Oben bei der Schrödingergleichung haben Sie einen solchen Operator am Werk gesehen: Die Ableitung des ψ nach der Zeit ist eine Maschine: Sie werfen ψ rein und heraus kommt wieder ψ, multipliziert mit einer Zahl. Einfache Multiplikationen sind linear. Daraus kann man drei interessante Sachverhalte ableiten, die in etwas technischeren Diskussionen der Quantenmechanik immer wieder vorkommen:

1. Der Hilbertraum ist linear, er kann mit so genannten Eigenvektoren als Basis dargestellt werden.
2. Solche Funktionen, z.B. ψ, gehen bei einer Operation in sich selber über, multipliziert mit einer Zahl.
3. Die Zeit und die Energie hängen miteinander zusammen: Sie sind kanonisch konjugiert, wie man sagt. Da die Ableitung nach der Zeit eine kleine zeitliche Änderung beschreibt, sieht man, dass diese Änderung durch die Energie hervorgerufen wird. Der Operator (Ableitung nach der Zeit) entspricht einer Grösse, die das System charakterisiert, der Energie E.

Der Zugang zur Quantenmechanik über den Hilbertraum ist abstrakter als der Zugang mit Pfeilen. Eine auch für Laien sehr lesbare Darstellung findet sich im ersten Kapitel des Buches *Philosophie der Quantenphysik* von Friebe und anderen Autoren (Friebe 2018).

10.1.5 Wellenpakete, Heisenbergsche Unschärfe

Die Darstellung von Teilchen als Wellen in der beeindruckenden Formel als komplexer Exponentialfunktion ist nicht ganz ohne Tücken. Um dies zu verstehen, betrachten wir der Einfachheit halber nur einen einzelnen Pfeil (ebene Welle).

$$\psi(t) = r * e^{-i\frac{E}{\hbar}*t}$$

Wenn wir deren Projektion auf die x-Achse betrachten, wie dies in Fig. 9.6 c dargestellt ist, dann würde sich auf dem Papierstreifen eine Cosinus-Funktion ergeben. In der Figur 10.1 b ist cos(3t) abgebildet. Wir lassen alle Einheiten

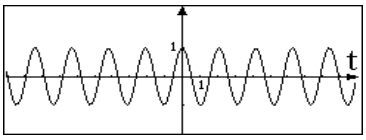


Fig. 10.1 b: Ebene Welle

weg und denken, die Energie sei 3. Die Physikerinnen nennen dies eine *ebene Welle*. Zu welcher Zeit wurde dieses Teilchen dargestellt, wann wurde es fotografiert? Man kann es nicht sagen: Das Teilchen ist über die ganze Zeit von minus unendlich bis plus unendlich verschmiert! Mit einer solchen Wellenfunktion kann man zwar die Energie des Teilchens ermitteln, indem man den Abstand zwischen zwei Wellenbuckeln misst, man kann aber nicht sagen, ob es heute, gestern oder zum Beginn unserer Zeitrechnung existierte.

Um diese zeitliche Verschmierung zu umgehen, stellt man ein Teilchen als Wellenpaket dar: Man addiert Energien oder Frequenzen auf, die nahe bei E liegen, und bekommt dann eine Überlagerung von Wellen, deren grösste Höhen sich um eine feste Zeit gruppieren. Ein Teilchen besteht dann nicht aus einer einzigen Energie, sondern aus einem Energiebündel wie Weinberg sagt. In der Figur 10.1 c ist die Summe der Cosinusfunktionen von 2.6 t bis 3.4 t dargestellt. Man sieht, dass sich die Wellenfunktion nun um 0 herum konzentriert. Damit ist die Wahrscheinlichkeit gross, dass das Teilchen zur Zeit t=0 gefunden werden kann.

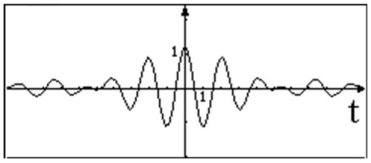


Fig. 10.1 c: Wellenpaket

An einem Wellenpaket kann man nun auch sehr schön das Wirken der so genannten Heisenbergschen Unschärfe (siehe auch § 10.1.8) nachvollziehen. Die Energie des Wellenpakets erstreckt sich über einen Bereich von E - 0.4 bis E + 0.4, sie hat zwar noch immer einen Mittelwert von 3 aber bei einer Messung würde die Energie in einem Unschärfeband von $\Delta E = 0.8$ liegen. Dafür ist nun aber auch die Zeit bestimmbar. Die Figur 10.1 c zeigt: Sie liegt in der Gegend von null, innerhalb eines Bandes von Δt in der Grössenordnung von 8. Würde man alle Einheiten berücksichtigen

und zudem mit der Wahrscheinlichkeit, dem Quadrat der Wellenfunktion, rechnen, dann ergäbe sich:

$$\Delta E * \Delta t \geq \frac{h}{2\pi} = \hbar$$

Weil man bei der ebenen Welle sehr viele Buckel in gleichmässigem Abstand hat, kann man die Energie ganz genau bestimmen ($\Delta E = 0$), allerdings zum Preis, dass man von der Zeit gar nichts weiss ($\Delta t = \infty$).

10.1.6 CBH-Theorem

Es gibt spannende Ansätze in der Physik, etwa um die Nullerjahre entwickelt, die von der beschränkten Information ausgehen und daraus die Wellenfunktion ableiten. Carlo Rovelli fasst Information als Verbindung zweier Systeme miteinander auf: Die Systeme sind miteinander korreliert, sie sind entangelt. Er sagt (Rovelli 1997, S. 9): «Correlation is «information» in the sense of information theory.» Damit hat das eine System Information vom anderen. Wenn zwei Teilchen (z.B. Photonen) mit einem Eigendrehimpuls 1, einem so genannten Spin 1, produziert werden, dann weiss das eine Teilchen vom anderen. Dieses Wissen ist in seinem eigenen Spin-Zustand gespeichert: Hat es +1, so weiss es vom anderen, dass dieses -1 hat.

Wheeler hat wie gesagt schon 1988/1989 den Begriff „It from Bit" geprägt (Wheeler 1989, S.309).[13] Er sagt, dass Physik wesentlich aus Information aufgebaut sei und zwar aus binärer Information. Dafür gibt er drei Beispiele: Das Photon, das in einem Detektor – oder im Auge – Klick oder Nicht-Klick macht, die Messung des magnetischen Flusses und die Beckenstein-Formel, dass die Information im schwarzen Loch begrenzt ist. Wheeler formuliert vier Verneinungen (Wheeler 1989, S. 313):

1. Es gibt keinen schichtweisen Aufbau des Universums aus beliebig vielen Ebenen: „Now tower of turtles". Es gibt eine Grenze.
2. Es gibt keine präexistierenden Gesetze. Das Universum ist ein sich selbst entwickelndes System.
3. Es gibt kein Kontinuum: Das ist der grosse Unterschied zwischen Mathematik und Physik.
4. Ohne Raum gibt es keine Zeit.

In der Theorie der QM hat man bis zum Ende des letzten Jahrhunderts darum gerungen, eine vollständige Theorie zu generieren. Meiner Meinung nach ist dieser Ansatz wenig zielführend. Das Studium der Information in

quantenmechanischen Systemen und deren Beschränkung ist ein fruchtbarer Ansatz.

10.1.7 Interpretation der QM

Vielleicht kann man das Vorgehen der Quantenmechanik besser verstehen, wenn man einen Würfel im Flug beschreiben will: Er könnte ja eine Eins oder eine Zwei oder eine Drei … bis hin zu einer Sechs zeigen, wenn er schliesslich auf einer Unterlage zur Ruhe kommt. Bei einem idealen Würfel wären die Wahrscheinlichkeiten je 1/6. Man nennt eine solche Beschreibung im Fluge eine Zustandsfunktion. Sie ist eine Summe von 6 Teil-Zustandsfunktionen. Jede Teilzustandsfunktion hat eine Wahrscheinlichkeit von 1/6. Die Summe nennt man wie gesagt auch Überlagerung.[14] Beim Lichtteilchen hat man z.B. eine Überlagerung von zwei (Teil-)Zuständen: an der Vorderseite reflektiert werden oder an der Rückseite. Man könnte aber auch weitere Zustände überlagern, z.B. hindurchgehen oder reflektiert werden. Diese Überlagerung von Zuständen ist nun der Knackpunkt. Wie soll man eine solche Überlagerung «verstehen»? Meiner Meinung nach kann man drei grundsätzliche Interpretationsweisen identifizieren: die Modell-Theorie, die Viele-Welten-Theorie und die Naiver-Realismus-Theorie.

1. Den Standpunkt der Modell-Theoretiker habe ich in dieser Darstellung eingenommen; sie gehen davon aus, die Überlagerung von Zuständen

[14] Falls Sie eine höhere Schulbildung genossen haben, könnte Sie ein bisschen mehr Mathematik interessieren. Man fasst die Zustandsfunktion als Vektor auf. Seine Basiszustände könnten mit $\vec{e}_1 = \begin{pmatrix}1\\ \vdots \\ 0\end{pmatrix}$, $\vec{e}_2 = \begin{pmatrix}0\\ 1 \\ \vdots\end{pmatrix}$.. beschrieben werden, wobei e1 bedeutet, der Würfel zeigt eine 1, e2 der Würfel zeigt eine 2 usw. Die Zustandsfunktion wird traditionell mit einem ψ bezeichnet und wäre dann eine Summe:
$\vec{\Psi} = a * \begin{pmatrix}1\\ 0 \\ \vdots\end{pmatrix} + b * \begin{pmatrix}0\\ 1 \\ \vdots\end{pmatrix} + \ldots$. Wenn man nun eine Wahrscheinlichkeit ausrechnen will, verwendet man das Skalarprodukt von ψ mit z.B. a*$\begin{pmatrix}1\\ \vdots \\ 0\end{pmatrix}$, dann erhält man für die Wahrscheinlichkeit, dass eine 1 auftritt den Wert a^2. Somit ist $a = \sqrt{\frac{1}{6}}$. Das Skalarprodukt entspricht der Messung und ergibt die zu erwartende Wahrscheinlichkeit, dass eine 1 auftritt. Man nennt die Bildung des Skalarproduktes die Reduktion der Wellenfunktion. Diese Reduktion wird durch das Messen vollzogen: Die Messung wird durch den Tisch ausgelöst, der den Flug des Würfels beendet und ihn zur Ruhe bringt, so dass man die Augenzahl ablesen kann.

sei bloss ein Modell zur Berechnung der Wahrscheinlichkeiten. Wir müssten nicht darüber nachdenken, ob eine solche Überlagerung in der Realität existiert, es genüge, wenn uns das Modell die richtigen Wahrscheinlichkeiten liefert. Diese Überzeugung wird oft auch FAPP-Ansatz genannt: Für Alle Praktischen Probleme (eigentlich: for all practical purposes).

2. Die Mehrere-Welten-Theorie stützt sich auf den Gedanken des Wohnzimmers im schwarzen Kasten (siehe § 3.5.3). Wenn das Lichtteilchen eingetreten ist, wissen wir nicht, ob der Beobachter tot ist, oder ob er noch lebt. Wenn er lebt, hat er vielleicht geheiratet und Kinder gezeugt. Wenn er tot ist, hat er keine Nachkommen. Es würden mit den zwei Zuständen des Lichtteilchens im Laufe der Zeit zwei ganz unterschiedliche Welten entstehen. Herausfinden, welche Welt realisiert ist, würde man nur, wenn man in den Kasten hineinschaut. Aber bei den meisten quantenmechanischen Abläufen schaut niemand in den Kasten. Natürlich wissen die Viele-Welten-Theoretiker auch, dass wir Menschen uns über unsere Wahrnehmungen austauschen können und uns deshalb gewiss sind, in der gleichen Welt zu leben. Aber wie reduzieren sich die vielen Welten der Quantenmechanik?
3. Der Ansatz des naiven Realismus geht davon aus, dass die erwähnte Überlagerung eine eigene Realität aufweist. Es muss sich um Objekte handeln, die in der wahrgenommenen äusseren Realität vorkommen. Es sind nicht bloss Konstrukte in uns drinnen, in unserem Geist. Rovelli ist der Meinung, dass viele Physiker diesem naiven Realismus anhängen (Vgl. Rovelli 1997, S. 19, Fussnote 11). Zudem ist das Messen ein Problem: Es kann doch nicht sein, dass die äussere Realität davon abhängt, ob ein Beobachter sie betrachtet, ob er eine Messung durchführt (Vgl. z.B. Penrose 1994, S. 307-316, Friebe 2018, S.66 ff).

Die meisten Physikerinnen und Physiker gehen davon aus, dass die Natur entscheidet, ob das Lichtteilchen am Fenster spiegelt oder hindurchtritt. Es gibt in der Realität keine Überlagerung von Zuständen. Eine Überlagerung von Zuständen und eine unentschiedene Situation gibt es nur im Geist des Betrachters, weil ihm die Information fehlt. Wenn nun Menschen ihre geistigen Bilder und Modelle mit der Realität gleichsetzen, sprechen einige Philosophen wie z.B. Metzinger von naiven Realisten. Solange wir nicht über uns kritisch nachdenken, sind wir alle naive Realisten.

10.1.8 Natürliche Informationsbeschränkung

Die Frage, ob man aus einem System beliebig viel Information gewinnen kann, hängt mit dem Konzept von so genannten hidden variables zusammen. Beim schwarzen Kasten und Schrödingers Katze können wir deren Wirken nachvollziehen. Der Beobachter ist tot, wenn das Photon reflektiert wurde. Nun «weiss» aber das Lichtteilchen nicht, ob es reflektiert wird, es trägt intern keinen Schalter mit, der bei „On" reflektieren bedeutet und bei „Off", nicht-reflektieren. Dieser Schalter wäre eine hidden variable. Selbst wenn es solche Variablen gäbe, könnten wir sie nicht entdecken. Ich will diese Aussage am Atomkern begründen:

Bei Atomkernen hat man ein Modell mit einer Art hidden variables aufgebaut: das Quarkmodell mit den so genannten Gluonen. Die Quarks sind die Grundbestandteile der Protonen und Neutronen und die Gluonen sind der Klebstoff, der sie aneinanderbindet. Dieses Modell leistet zwar gute Dienste, wir können aber seine Einzelteile nicht gesondert untersuchen. Wenn wir in den Atomkern eindringen wollen, um ihn zu analysieren, brauchen wir so viel Energie, dass sofort neue Atomkerne entstehen.

Zur Untersuchung verwendet man Elektronen oder Protonen, die man auf den Kern schiesst. Wie beim Lichtteilchen und der Fensterscheibe, erwartet man, dass sie von den Strukturen des Kerns reflektiert oder abgelenkt werden. Damit die Testteilchen, die Protonen und Elektronen, auch nur in die Nähe des Atomkerns kommen können, brauchen sie etwa so viel Energie, wie ein Kernbestandteil selber aufweist. Statt dass die Testteilchen nun reflektiert oder abgelenkt werden, verschmelzen sie mit den Kernbestandteilen und bilden neue Teilchen. Diese stieben auseinander; sie bilden so genannte Jets. Aus der Anzahl und Natur der Jets kann man Rückschlüsse auf die Quarks ziehen: Man kann Quarks und Gluonen aber nicht unbeeinflusst untersuchen. Wenn die Energie der Probeteilchen gleich gross ist wie die der untersuchten Teilchen, ist deshalb eine natürliche Grenze der Untersuchungsmöglichkeit erreicht. Unsere Erkenntnisfähigkeit ist dadurch beschränkt. Wir sagen, die Messung (zer-)störe das System. Man kann schon komplizierte Modelle bauen, aber was nützen sie, wenn sie keine praktischen Auswirkungen haben oder prinzipiell unzugänglich sind?

Eine weitere Beschränkung des Informationsgewinns ergibt sich aus der unter 10.1.5 dargestellten Heisenbergschen Unschärfe. Sie legt fest, wie genau Paare von so genannten komplementären Messgrössen wie Energie und Zeit oder Impuls und Ort gemessen werden können. Auch dadurch ist die Informationsmenge beschränkt. Allerdings ergeben sich aus dieser

Unschärfe auch neue Phänomene, die ganz ungewöhnliche Effekte hervorrufen: Innerhalb einer kleinen Zeiteinheit Δt kann eine Energiemenge ΔE entstehen. Es ist möglich, dass kurzzeitig ein Teilchen-Antiteilchen-Paar aufscheint, das nach der Zeit Δt wieder verschwindet. Damit wird das Nichts, das Vakuum, einer der interessantesten Zustände der Physik: Es kann durch solche Teilchenpaare bevölkert sein. Sie leben zwar nur kurz, haben aber dennoch messbare Auswirkungen.

10.1.9 Quantengravitation

Heute gibt es Ansätze, die postulieren, dass es keine Distanzen geben kann, die genauer sind als 10^{-35} m, die so genannte Planck'sche Länge. Man geht davon aus, dass der Raum körnig ist, allerdings mit sehr, sehr kleiner Korngrösse. Wenn man der Welt eine solche Hypothese zu Grund legt, kann man zwar das Digitalisierungsproblem lösen, man handelt sich aber eine neue Schwierigkeit ein, die man schon in der Quantisierung der Energie antrifft: Die Information ist dann beschränkt. Mit dieser Gesetzlichkeit kommen wir in der Physik noch nicht gut zurecht.

10.2 Standardmodell der Teilchenphysik

10.2.1 Symmetrien

Das so genannte Standardmodell leitet alle Eigenschaften der uns bekannten Teilchen aus der Grundstruktur des Raumes, oder präziser gesprochen der vierdimensionalen Raum-Zeit, her. Einige Leserinnen und Leser werden sich fragen, wie diese Herleitung zu Stande kommt. Man betrachtet die so genannten Symmetrien des Raumes: Man *dreht* zum Beispiel den Raum und weiss, dass die Physik dann gleichbleibt: Das ist eine Symmetrie. Eine solche Rotation erleben wir jeden Tag auf unserer Erde: Die Physik ist am Abend noch genau gleich wie am Morgen, obwohl die Erde sich gedreht hat. Nun identifiziert man die Operation, die eine Drehung erzeugen kann. Es ist eine Rotation, die die Erde um einen Winkel dreht. Zu ihr gehört eine der genannten fundamentalen Eigenschaften. Bei der Drehung ist es der Drehimpuls. Eine weitere Symmetrie ist das Verschieben eines physikalischen Systems zum Beispiel von Zug nach Zürich: In Zürich läuft das System immer noch gleich wie in Zug. Dies ist eine weitere Symmetrie. Für diese Ortsveränderung braucht es eine Geschwindigkeit oder einen Impuls, der das System horizontal bewegt. Die Geschwindigkeit bewirkt eine Ortsveränderung, deshalb ist die Geschwindigkeit oder allgemeiner der

Impuls eine weitere Eigenschaft, die ein System haben kann. Man könnte auch links und rechts vertauschen, dann nennt man die Eigenschaft Parität. Nicht alle Teilchen haben aber eine Parität, genauso wie nicht alle eine Ladung haben: Parität haben nur die Teilchen, die unter einer Vertauschung von Links und Rechts gleichbleiben. Teilchen der schwachen Wechselwirkung (Radioaktivität) verfügen nicht über diese Symmetrie. Sie verändern sich, wenn man tauscht.

10.2.2 Symmetrie und Wellenfunktion

Leserinnen und Leser, die sich die Vertiefungen zur Quantenmechanik bis zur imposanten Formel für die Wellenfunktion (§ 10.1.4) angeeignet haben, können nun diese Symmetrieüberlegung an der Wellenfunktion (WF) nachvollziehen. Die WF lautet:

$$\psi(t) = r * e^{-i\frac{E}{\hbar}*t}$$

Wenn zwei Physiker den Nullpunkt ihrer Uhren unterschiedlich gestellt hätten (Δt), dann müsste die Physik gleich sein. Wie reagiert die Wellenfunktion nun auf eine solche Symmetrie-Transformation? Sie erzeugt eine Phasendifferenz:

$$\Delta\varphi = \frac{E}{\hbar} * \Delta t$$

Die Zeitdifferenz der Wellenfunktion ist für die Physik nicht erheblich, wohl aber der Faktor mit dem sie multipliziert wird. Oder anders gesagt: Die Frequenz ist entscheidend. Man kann eine (ebene) Welle irgendwo anschauen, der Abstand zwischen den Wellenbuckeln ist immer gleich gross. (Weinberg 1993, S. 145): „*In der heutigen Physik (Quantenmechanik) definieren wir die Energie eines Systems als Phasenänderung der Wellenfunktion, wenn wir den Gang unserer Uhren um 1 s verstellen.*“ Wie üblich ist die Planck‘sche Konstante nur ein Umrechnungsfaktor der Energie von der natürlichen Einheit (Schwingungen pro Sekunde) auf die üblichen Einheiten wie Joule oder Nm oder eV. Diese Definition hat alle Eigenschaften einer Energie. Insbesondere deren Invarianz unter Zeit-Transformation. Diese Symmetrie ist sogar der Grund, *warum* es eine Energie gibt.

Eine gleiche Symmetrie ergibt sich auch daraus, dass es unerheblich ist, ob ich die Physik in Zug oder Zürich betreibe. Die entsprechende Phasenänderung ist dann proportional zum Impuls eines Objektes. Dadurch definieren wir den Impuls.

$$\Delta\varphi = \frac{p}{\hbar} * \Delta x$$

10.2.3 Felder statt Kräfte

Viele Menschen – und ich lange Zeit auch – denken sich die Welt aufgebaut aus Teilchen, die über Kräfte miteinander wechselwirken. Diese mechanistische Sicht ist überholt. Seit Einstein und Maric 1905 forderten, dass die Physik so genannt «relativistisch invariant» sein muss, dominieren die Symmetriebetrachtungen die Modelle der Physik. Im Fall der Relativitäts-Symmetrie heisst dies: Die Gesetze, die ein System bestimmen, müssen genau gleich sein, ob ich die Welt von einem ruhenden oder einem gleichförmig bewegten Bezugssystem aus betrachte. Die Gesetzte dürfen nicht davon abhängen, ob ich mit dem System mitfahre – indem ich in ihm ruhe – oder ob ich es von aussen betrachte und es an mir vorbeifliegt. Dies ist einerseits eine Folge davon, dass sich nichts schneller als mit Lichtgeschwindigkeit bewegen kann und andererseits muss sich nun aber eine Kraft *ausbreiten*, sie kann nicht instantan wirken: Da die Sonne acht Minuten von uns entfernt ist, würden wir eine Änderung ihrer Anziehungskraft erst nach acht Minuten bei uns auf der Erde feststellen. Diese Forderung kann nur eine Kraft leisten, die als so genanntes Feld formuliert wird. Ich will dies im Folgenden anhand der Darstellung von Steven Weinberg im Buch: *Der Traum von der Einheit des Universums* nachvollziehen.

Der Feldbegriff ist nicht so schwierig zu verstehen. Fragen Sie einen Maturanden, was sein Gewicht in Newton sei: Er sagt: «m mal g». Zum Beispiel 75 kg mal 10 Newton pro Kilogramm. Dieses kleine g ist ein «Gravitationsfeld». Warum? Das Gravitationsgesetz ist zunächst vollständig gleichberechtigt in den beiden Massen (M: Erdmasse und m: Masse des Maturanden).

$$F_G = G\frac{M * m}{r^2}$$

In der Praxis erleben wir dies aber überhaupt nicht so. Wenn wir zu Boden fallen und hart aufschlagen, empfinden wir uns als Opfer und die Erde als Verursacher der Anziehungskraft. Sehr oft erzeugt die grosse Masse (M) eine Kraft und die kleine Masse (m) erleidet sie. Man teilt das Gravitationsgesetz deshalb in einen Täter und in ein Opfer auf:

$$F_G = G\frac{M * m}{r^2} = g(r) * m$$

$$g(r) = G\frac{M}{r^2} \quad (= g = 10\,N/kg, auf\ der\ Erdoberfläche)$$

Damit lässt sich die Ursache einer Beeinflussung an jedem Punkt im Raum festlegen. Ein Kraftfeld ist eine gerichtete Grösse, sie zeigt beim Beispiel der Erdanziehung zum Mittelpunkt der Erde. Zudem kann sich diese

Beeinflussung ausbreiten, wenn die felderzeugende Masse z.B. etwas verschoben wird. Das Feld ist dann auch eine *zeitabhängige* Grösse.

10.2.4 Energiebündel statt Teilchen

Auch das Konzept von Teilchen als Punktmassen erfährt eine Verallgemeinerung in der modernen Physik. Wegen der Quantenmechanik beschreibt man Teilchen mit Pfeilen oder so genannten Materiewellen, die zu Wellenpaketen gebündelt sind. Im Kapitel 10.1 haben wir diese Pfeile eingeführt und bis zu einem so genannten Wellenpaket weiterentwickelt (§ 10.1.5). Weinberg sagt über das Gravitationsfeld (Weinberg 1993, S. 148): „Auch im Gravitationsfeld treten die Energie und der Impuls in Bündeln auf, den Gravitonen, die sich ebenfalls wie Teilchen ohne Massen verhalten."

Massereiche Teilchen wie Elektronen können nun auch als Bündel von Energie und Impuls aufgefasst werden, die sich in unterschiedlichen Feldern bewegen. Weinberg schreibt den Ursprung (1929) dieser Idee Heisenberg, Pauli, Jordan und Wigner zu. Ich würde de Broglie noch hinzufügen. Weinberg führt dann weiter aus (ebd.): «Die Kraft zwischen Photon und Elektron kann daher auch mit einem Austausch von Elektronen beschrieben werden. Die Unterscheidung zwischen Kraft und Materie verschwindet weitgehend. Jedes Teilchen kann die Rolle eines Testkörpers einnehmen oder es kann die Rolle des Austauschteilchens spielen, das die Kraft vermittelt." Wir sprechen dann von Feldern, die sich wie Wellen verhalten und Energiebündeln, die als Wellenpakete dargestellt sind. Damit werden Teilchen und Kräfte symmetrisch. Allerdings ist zu viel Symmetrie auch in der Physik nicht attraktiv.

10.2.5 Spontane Symmetriebrechung

Genauso wie im alltäglichen Leben wäre auch in der Physik eine vollständig symmetrische Welt eintönig und uninteressant. Die Tatsache, dass ein Gesicht dann individuell und unvergleichlich ist, wenn die linke und die rechte Hälfte sich geringfügig unterscheiden, findet ihre Entsprechung in der so genannten Symmetriebrechung. Wir erleben Prozesse, die Symmetrien aufheben täglich mit: Wasserdampf kondensiert zu Regen, Wasser erstarrt zu Eis usw. Ein wegweisendes Beispiel stellt die Magnetisierung eines Eisenstücks dar: Man kann sich jedes einzelne Eisenatom als kleinen Stabmagneten vorstellen. Wenn das Eisen sehr heiss ist, werden die Eisenatome so stark durcheinandergeschüttelt, dass die Stabmagnete in alle Richtungen zeigen. Würde man sich innerhalb des Eisens vom Punkt A um eine bestimmte Strecke nach B verschieben, dann

träfe man überall die gleiche Konfiguration an: Die Magneten zeigen in alle Richtungen. Der Zustand des Eisens ist symmetrisch. Er ist überall gleich.

Kühlt das Eisen ab, schüttelt die Temperatur die kleinen Magnete weniger stark durcheinander und ihre gegenseitige Beeinflussung – sie richten sich gerne wie der Nachbar aus – beginnt eine Wirkung zu entfalten: Es bilden sich Bezirke, in denen die Magnete alle in eine Richtung zeigen. Man nennt sie Weiss'sche Bezirke. Würde man sich jetzt von A nach B bewegen, dann könnte es sein, dass bei A die Magnete nach Norden ausgerichtet sind, während sie bei B nach Süden zeigen. Der Zustand des Eisenstücks ist nicht mehr symmetrisch. Die anfängliche Symmetrie ist gebrochen. Dieses Verständnis von Symmetrie widerspricht unserem intuitiven Empfinden: Wir würden Bezirke, in denen die Magneten alle gleich ausgerichtet sind, als geordnet und damit symmetrisch anschauen, während wir den ungeordneten Zustand bei hoher Temperatur als unsymmetrisch wahrnehmen.

Ein solches magnetisierbares Eisenstück wird mit einem berühmten Modell in der statistischen Physik beschrieben: dem so genannten Ising-Modell. Die durch dieses Beispiel dargestellte Symmetriebrechung leitet auch die Theoriebildung in der Elementarteilchenphysik. Sie erfordert folgende Voraussetzungen:

- Es muss ein (Kraft-)Feld geben.
- Die Teilchen (Atome) müssen sich an dieses Feld koppeln.
- Durch die Symmetriebrechung erhalten die Teilchen unterschiedliche Eigenschaften.

Aus unserer Erfahrung wissen wir, dass ein Eisenstück durch ein äusseres Magnetfeld magnetisiert werden kann. Dies ist aber keine notwendige Voraussetzung für eine Symmetriebrechung. Es braucht prinzipiell ein (Kraft-) Feld – es muss nicht von aussen wirken. Es genügt, wenn zwei Atome oder Teilchen sich an ein solches Feld koppeln oder sich durch dieses Feld gegenseitig beeinflussen können.

10.2.6 Formalisierung: Gruppentheorie

Bisher haben wir kaum auf den Formalismus zurückgegriffen, in dem das Standardmodell formuliert ist. Die beschriebenen Symmetrieoperationen, z.B. eine Rotation oder eine Translation, sind so genannte Koordinatentransformationen, die aber den Abstand zwischen zwei Punkten konstant halten müssen. Abstände werden allgemein durch eine so genannte Metrik gemessen. Solche Operationen führen dazu, dass die inverse Transformation, angewendet auf die Transformation selbst, die

Einheit ergibt, oder allgemeiner: Dass die inverse, angewendet auf die transformierte Metrik, wieder die ursprüngliche Metrik ergibt. Es gibt eine enge Analogie dieser Transformationen zur Rotation im 3-dimensionalen Raum. Eine Rotation R und ihre Inverse ergeben die Einheitsmatrix **1**: R^T***1***R = **1.** Sie werden *orthogonal* genannt und mit O(3) bezeichnet. Weil O(3) die Umkehrung der Richtung (Parität) als Symmetrie enthält, schliesst man diese aus, indem man fordert, die Determinante müsse 1 sein, und redet dann von der speziellen orthogonalen Gruppe SO(3). Der Begriff Determinante ist ein Fachbegriff aus der so genannten linearen Algebra oder populärer der Matrizenrechnung.

Wenn solche Transformationen nicht auf Vektoren (Geschwindigkeiten oder Ortsbezeichnungen) wirken, sondern auf Materiewellen, dann müssen sie auch die Wahrscheinlichkeit invariant lassen. Man spricht dann von unitären Transformationen (SU(3)). Aus diesen Gründen sind Gruppentheorie und die Darstellung von Gruppen der theoretische Mechanismus, auf dem das Standardmodell fusst.

11 Vertiefungen zu Konzepten der Informatik

11.1 Informationsbegriff von Shannon

Man kann den Informationsbegriff von Shannon noch mathematischer fassen. Dabei wollen wir nicht die Anzahl Möglichkeiten zu Grunde legen, sondern die Wahrscheinlichkeit, mit der ein Ereignis auftritt. Wir gehen wieder von einem Beispiel aus.

11.1.1 Beispiel: Überraschungseffekt

Der Informationsgehalt kann auch als Grösse der Überraschung geschrieben werden. Wir nehmen an, wir hätten eine unfaire Münze, die in 95 % aller Fälle einen Kopf zeigt und nur in 5 % eine Zahl. Wir wären ziemlich überrascht, wenn sie bei einem Wurf die Zahl zeigen würde, und wenig überrascht, wenn der Kopf oben liegen würde. Der Informationsgehalt misst diese Überraschung. Bei «Zahl» ist die Berechnung nicht schwierig: In 5 von 100 Fällen kommt «Zahl» oder in einem von 20 Fällen. Die Anzahl Bit, die man für 20 Möglichkeiten braucht, ist etwas mehr als 4: 16 Fälle wären 2^4 und 32 Fälle wären 2^5. Bei «Kopf» ist die Berechnung schwieriger: auf 100 Fälle gibt es 95 günstige. Auf einen günstigen Fall gibt es nur 1.05 mögliche. Die Berechnung der Anzahl Bit durch die Anzahl Möglichkeiten ist in diesem Fall undurchsichtig. Wir wollen die Anzahl Bit oder die Informationsmenge mit Wahrscheinlichkeiten berechnen. Wir haben gesagt, man könne eine Wahrscheinlichkeit als Anzahl der günstigen Fälle / Anzahl der möglichen Fälle definieren.

$$p = \frac{günstige\ Fälle}{alle\ möglichen\ Fälle}$$

Nehmen wir zur Vereinfachung an, es gebe nur einen günstigen Fall. Dann entspricht p gerade dem Kehrwert aller möglichen Fälle, die im Bolzmannschen Entropiebegriff mit W bezeichnet wurden. Nun haben wir gesagt, der Shannon‘sche Informationsbegriff zähle, wie viele Bits nötig seien, um diese Anzahl möglicher Fälle zu beschreiben. Im Zweiersystem sind die Bits die Hochzahl, wenn man die möglichen Fälle mit einer Zweierpotenz zu fassen sucht.

$2^1 = 2$
$2^2 = 4$
$2^3 = 8$
$2^4 = 16$

Als Faustregel kann man sich merken: $2^{10} = 1000$; dies wäre ein Kilo-Bit.

Logarithmus

Es gibt nun eine Funktion, die ermittelt, wie gross ein solcher Exponent ist, wenn man die Zahl kennt, die als Hochzahl dargestellt werden sollte. Allgemein nennt man diese Funktion Logarithmus. Viele Leute kennen den 10-er Logarithmus, der log heisst und der den Exponenten für eine Zehnerpotenz beschreibt. Z.B wollen wir 4500 als $10^{\text{hoch irgendwas}}$ schreiben. Bei 1000 wäre dies drei und bei 10'000 wäre der Exponent 4. Mit dem Taschenrechner bekommen wir als Exponenten: 3.65. Man kann diese Berechnung als beeindruckende Formel schreiben:
Exponent = log(4500)
Hätte man als Basis das Zweiersystem, so würde man $\log_2(4500)$ oder lb(4500) schreiben. Lb steht für «Logarithmus binär», da das Zweiersystem Binärsystem heisst. Die Shannon'sche Informationsmenge (Inf) lässt sich deshalb schreiben als
Inf(4500) = lb (4500)= 12.1
Um 4500 Möglichkeiten darzustellen, braucht man 13 Bit. Wenn es nun eine einzige günstige Möglichkeit unter diesen 4500 gibt, dann wäre deren Wahrscheinlichkeit 1/4500. Also könnten wir die Anzahl der Möglichkeiten W als 1/p beschreiben:
Inf(4500)=lb(4500)=lb(1/p)
Da eine Potenz unter einem Bruch als negative Potenz geschrieben werden kann, ist der Informationsgehalt eines Ereignisses mit Wahrscheinlichkeit p: Inf(p) = -lb(p)
Mit dieser Erkenntnis können wir nun das Münzenbeispiel berechnen:
Inf(Kopf) = - lb(0.95) = 0.074
Inf(Zahl) = - lb(0.05) = 4.32
Mit Möglichkeiten gesprochen: Bei «Zahl» gibt es sehr viele Möglichkeiten und nur wenige günstige, nämlich 5 auf 100 oder 1 auf 20. Bei «Kopf» gibt es fast keine Möglichkeiten und die Informationsmenge ist annähernd null. Lb(1) = null. Wenn wir nun verschiedene Ereignisse hätten, die mit den Wahrscheinlichkeiten p_1, p_2, p_3 .. auftreten, dann hätten wir einen mittleren Informationsgehalt von:
Inf(versch. Ereignisse) = -(p_1*lb(p_1) + p_2*lb(p_2) + p_3*lb(p_3) + …)
Wir könnten uns nun noch überlegen, wie gross der *durchschnittliche* Überraschungseffekt ist, der so genannte Erwartungswert (E):
E(Überraschung)= 0.95 *0.074 + 0.05*4.32 =0.07 + 0.216 = 0.286

Die unfaire Münze generiert Information, die im Durschnitt 0.286 Bit liefert. Der Informationsgehalt ist gering, weil dieser Durschnitt durch den Fall mit hoher Wahrscheinlichkeit dominiert ist. In einer solchen Situation,

gibt es wenig Möglichkeiten – verglichen mit dem günstigen Fall. Bei kleinen Wahrscheinlichkeiten existieren neben dem günstigen Fall viele weitere Möglichkeiten.

Der Informationsgehalt, gemessen mit der Entropie, kann nun zur Optimierung der Informationsübertragung genutzt werden. Rob DiPietro stellt in seinem Blogbeitrag ein schönes Beispiel dafür dar: [15] Wir denken uns zwei Freunde, die wie die Kinder Autos zählen. Der eine steht auf einer Autobahnbrücke und gibt dem anderen den Autotyp durch, der gerade vorbeifährt. Dabei kostet jedes Bit, das er dem Zählenden sendet, z.B. 0.1 Fr. Ein Autotyp, der oft vorkommt, wie z.B. ein Toyota Camry, soll mit einem teuren und seltenen Wagen, wie z.B. einem Tesla der S-Klasse, verglichen werden. Was ist nun die kostengünstigste Übertragung des Zählens? Der Informationsgehalt in Form der Entropie gibt uns sofort die Antwort: Der Autotyp mit der grösseren Häufigkeit sollte mit wenig Bit übertragen werden, während der mit dem seltenen Auftreten mehr Bits beanspruchen darf. Um mit diesem Denken in Entropie und Bits etwas vertraut zu werden, stellen wir uns vor, der Toyota sei 128-mal häufiger als der Tesla. Wie viel Bit ist dann der Informationsunterschied? Angenommen die Wahrscheinlichkeit des Toyota sei $128*p_{Tesla}$, der Informationsgehalt für den Toyota ist dann:

$Inf(Toyota)=-lb(128*p_{Tesla})$

Nun müssen wir etwas mit Logarithmen rechnen. Dabei ist es günstig, sich den Logarithmus als eine Abkürzung für «Exponent» vorzustellen. Beim Potenzrechnen wissen die meisten Leute, dass man bei einer Multiplikation von Potenzzahlen die Exponenten addiert.

$Inf(Toyota)= -(lb(p_{Tesla}) + lb(128)) = - lb(p_{Tesla}) – 7$

Da p_{Tesla} klein ist, wird $–lb(p_{Tesla})$ eine grosse, positive Zahl sein. Von ihr werden 7 Bits abgezählt; der Informationsgehalt der häufigen Toyotas ist um 7 Bit kleiner als der der seltenen Teslas.

11.1.2 Arbeit bei Shannons Informationsmodell

Wie im Haupttext angedeutet erweitern Lyre und zuvor schon Haken nun das Boltzmann‘sche Informationsmodell der Thermodynamik. Sie unterscheiden zwischen Information, die man hat (man kennt den Bewegungszustand jedes Moleküls eines Gases ganz genau), und solcher, die man wissen könnte. Sie legen also auf die Unterscheidung zwischen potentieller und aktueller Information grossen Wert und reden explizit von

[15] Auffindbar unter: https://rdipietro.github.io/friendly-intro-to-cross-entropy-loss/#entropy (20.11.2020).

der Informationsentropie, um die potentielle Information immer als solche zu benennen (Vgl. Lyre 2002, S. 48). Sie stützen sich auf C.F. von Weizsäcker, der sagt:

> "Die Entropie (S) ist der Erwartungswert des Neuigkeitsgehalts eines noch nicht geschehenen Ereignisses, also dessen, was ich noch wissen könnte. /.../ S ist ein Mass einer definierten Art von Nichtwissen./... / Sie misst also, wie viel derjenige, der den Makrozustand kennt, noch wissen könnte, wenn er auch den Mikrozustand kennenlernte. /.../ Die Entropie nennen wir die im Makrozustand enthaltene potentielle Information. Sie ist am grössten im thermischen Gleichgewichtszustand. /.../ In ihm ist die aktuelle Information über die Mikrozustände am kleinsten."(Zitiert nach Lyre 2002, S.47).

Aus dieser Identifizierung von potentieller Information mit der Entropie lässt sich sofort der minimale Energieaufwand, das Wärmeäquivalent, für ein Bit ermitteln:

$\Delta E = k_B * T * \ln 2$

Dieses Wärmeäquivalent besagt, dass der Übergang von potentieller zu aktueller Information eine gewisse Energie benötigt.

11.2 NETtalk und überwachtes Lernen

Wir wollen das Lernen eines KNN, wie z.B. des Netzwerkes NETtalk, noch etwas genauer ins Auge fassen. Dazu müssen wir zwei Dinge verstehen:

- Wie leitet ein Neuron Signale weiter?
- Wie werden die Verbindungsstärken beim Lernen angepasst?

11.2.1 Das Schalten der Neuronen

Ein ganz simples Neuron würde alle eingehenden Signale gewichten, sie aufaddieren und dann an die nachfolgende Neuronenebene weitergeben. Wenn das eingehende Signal von einem Neuron XY der vorausgehenden Ebene z.B. 0.03 wäre und das Gewicht 0.15, dann würde dieser Eingang 0.0045 zur Summe beitragen. Als man die Idee des Neurons um die Mitte des letzten Jahrhunderts in die Programmierung von Computern einführte, erweiterte man diese einfache Idee um eine Schwelle. Wenn die Signale die Schwelle überschreiten, feuert das Neuron, falls nicht, bleibt es still. Damit wurde das Neuron zum binären Schalter. Diesen Schalter realisiert man mit einem so genannten Bias, einer Art Vorspannung oder Hürde. Wie wirkt sie?

Als Bias addieren Neuronen zur Summe der mit Gewicht versehenen eingehenden Signale einen konstanten Betrag hinzu. Falls der Bias positiv ist, hat diese Addition zur Folge, dass auch Neuronen mit schwachen Eingangssignalen den Schwellwert erreichen und ein Outputsignal abgeben

können. Dieser Bias kann auch negativ sein. Dann wirkt er als Hürde, die das Erreichen der Schwelle erschwert. Ein solches Neuron wirkt dämpfend. Von realen Neuronen wissen wir, dass sie auch schalten. Wie beim Bild mit den zwei Wassersäulen (§ 1.3.7) geht die Klappe bei einem bestimmten Druck auf und das Axon leitet. Ein biologisches Neuron gibt ab einer bestimmten Ladungsmenge ein Aktionspotential – einen Spannungsstoss – weiter. Er ist in seiner Höhe immer gleich, wird aber von einer Verbindung zu einem weiteren Neuron, einer Synapse, je nach deren individuellem Zustand weitergeleitet oder nicht.

Bei künstlichen Neuronen gibt es verschiedene Möglichkeiten, dieses Schalten zu simulieren. Meist spricht man von einer Aktivierungsfunktion, die schaltet. Eine einfache Aktivierung wirkt als Schalter, der bei einem bestimmten Schwellenwert einen Puls erzeugt. Damit wird wie beim realen Neuron ein fester Output erzeugt, der dann vom nächsten Neuron mit einem Gewicht abgeschwächt oder verstärkt wird. Das Gewicht entspricht also dem Wirken der Synapse.

Dieses Schalten des Neurons erzeugt einen digitalen Output: «Ein» oder «Aus»: Das Ausgangssignal ist binär. Dies vereinfacht die Situation einerseits, führt aber auch eine künstliche Schwierigkeit ein. Würde man eine Stufenfunktion zur Aktivierung wählen, dann hätte man eine künstliche Unendlichkeit eingeführt, wie beim Problem von Achilles: Der Puls, der dieses Schalten bewirkt, müsste unendlich gross sein, weil der Schalter in einer Zeitspanne von null Sekunden umgelegt werden muss. Technisch sagt man, die Funktion sei unstetig, sie sei nicht differenzierbar. Man wählt deshalb keinen abrupten Sprung, sondern eine allmählich ansteigende Funktion wie z.B. ein so genanntes Sigmoid.

$$f(x) = \frac{1}{1 + e^{-x}}$$

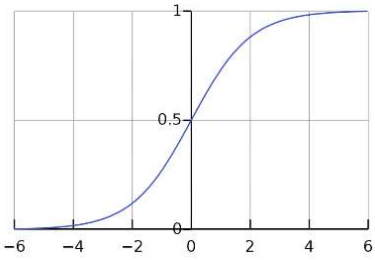


Fig. 11.2 a: Graph der Sigmoid-Funktion

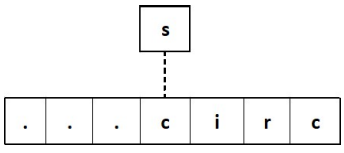
Diese Funktion schaltet in einem engen Bereich von 0 auf 1 und vermeidet den Fallstrick der Unendlichkeit. Auch diese Sigmoid- oder logistische Funktion ist nicht der Weisheit letzter Schluss. Aus zwei Gründen:

- Sie ist aufwändig zu berechnen
- Sie behindert das Lernen

Wenn ein künstliches Netzwerk trainiert wird, dann wird der Trainingsdatensatz Tausende von Malen abgearbeitet. Bei Verbindungszahlen zwischen den Neuronenebenen, die ebenfalls in die Tausende gehen, werden dabei sofort mehrere Millionen von Berechnungen der Sigmoidfunktion nötig. Zudem: Die starke Schwelle und die flachen Ausläufer dieser Funktion sind hinderlich. Bei so genannter Sättigung (Eingangssignal > +4 oder kleiner -4) ist das Ausgangssignal praktisch immer gleich; es bietet keinen Hebel mehr, um das Eingangssignal zu korrigieren. Wenn das Netzwerk nun lernt, sollte der Fehler aber auf das Eingangssignal übertragen werden können, so dass es verstärkt oder geschwächt werden kann. Dies ist bei einer Stufenfunktion unmöglich und bei einer Sigmoidfunktion nur in sehr beschränktem Masse machbar. Deshalb verwendet man – grob gesprochen – als Aktivierungsfunktion oft einen Multiplikator, der die Summe der Eingangssignale immer mit der gleichen Zahl multipliziert. Der Hebel, mit dem eine solche Funktion die Korrektur weitergibt, ist dann immer gleich stark. Eine Neuronenschicht mit derartiger Aktivierungsfunktion wird ReLU (Rectified Linear Unit) genannt.[16]

11.2.2 Lernen: die Verbindungsstärke anpassen

Wenn ein Netzwerk z.B. die Aussprache von «circuit» lernen muss, wird diese Buchstabenfolge dem Netzwerk eingegeben und am Ausgang sollte das Neuron mit dem s-Laut aktiviert werden. Da es sich um so genanntes überwachtes Lernen handelt, wird der Überwacher den Fehler bei der Ausgabe identifizieren. Wenn die Aussprache undeutlich ist, könnte beim Anfangs-c das s-Phon zu wenig aktiviert sein und das k-Phon zu stark. Deshalb müssten die Zuleitungen zum Output-Neuron für das s-Phon verstärkt werden.



Man leitet nun die Grösse des Fehlers auf die Neuronen der Mittelebene zurück, indem man deren Verbindungsstärke leicht anhebt. Allerdings dürfen diese Anpassungen nur ganz sachte ausgeführt werden: Wenn das s-Neuron 40 % zu wenig Signal bekommt, dann wird dieser Fehler mit einem

[16] Korrekterweise muss man erwähnen, dass eine derartige Aktivierungsfunktion auch alle negativen Werte auf null setzt.

Gewichtungsfaktor von einem Hundertstel bis zu einem Tausendstel geschwächt, bevor er auf die Neuronen der Mittelebene zurückübertragen wird. Der Faktor wird *Lernrate* genannt. Ihre Stärke ist ein so genannter System- oder Hyperparameter. Dieses Verfahren des Zurückspielens des Fehlers wird oft Backpropagation genannt.[17] Sehr raffinierte Netzwerke zeichnen sich durch äusserst ausgeklügelte Lernprozeduren aus. Diese Korrekturverfahren sind nicht ganze einfach zu verstehen, vor allem wenn das Netzwerk mehrere Mittelebenen umfasst, wie das bei so genanntem Deep-Learning der Fall ist. Sie bilden aber einen der Schlüssel für den Erfolg von KNN.

11.2.3 Mathematik des Korrekturverfahrens

Normalerweise wird dieses Korrekturverfahren mit einem sehr leistungsfähigen mathematischen Formalismus beschrieben, der für Laien eine grosse Hürde darstellt. Michael Nielsen erklärt diesen Mechanismus in einer gut lesbaren Einführung zu KNN in seinem beeindruckenden online-Buch *Neural Networks and deep learning* (Nielsen 2015). Ich will deshalb mit einfachen Worten einen Zugang ermöglichen: Zuerst muss man über den Fehler nachdenken. Beim Circuit-Beispiel ist der s-Laut zu wenig aktiviert. Wir sprachen von einem Fehler von 40 %. Für die Analyse genügt es, nur ein Trainingsbeispiel anzuschauen. Dem Netzwerk wird also ein einziges Mal circuit präsentiert. Der Fehler des s-Neurons wäre wie gesagt 40 %. Die Mathematiker betrachten nun alle Ausgangsneuronen. Der Überwacher gibt das Ziel vor: Es besteht aus einer Zahlenreihe von 79 Einträgen, bei der nur beim s-Laut eine 1 steht und bei allen übrigen eine Null. Diese Zahlenreihe ist ein Vektor und wird üblicherweise mit $\vec{\boldsymbol{y}}$ bezeichnet. Das Netzwerk selber liefert auch 79 Zahlen (Vektor $\vec{\boldsymbol{a}}$) für die Aktivierung jedes Lautes: Bei s würde nur 0.6 stehen und bei k wäre sie nicht null, sondern z.B. 0.3 für 30 %.

Nun kann man eine Fehlerfunktion, eine so genannte *Loss- oder Costfunktion* definieren. Sie besteht in vielen Fällen aus der Hälfte der Quadrate der einzelnen Abweichungen. Die Summe läuft über alle 79 Endneuronen.

[17] Der Begriff Backpropagation wird nicht einheitlich verwendet. Einige Autoren brauchen ihn in umfassendem Sinne, um den ganzen Korrekturprozess zu beschreiben. Andere setzen ihn nur in seiner ursprünglichen, eingeschränkten Bedeutung ein: als Technik zur schnelleren Berechnung der Korrekturrichtung, des so genannten Gradienten.

$$C = \frac{1}{2}\sum_j (y_j - a_j)^2 = \frac{1}{2} \parallel \vec{y} - \vec{a} \parallel^2$$

Der hintere Teil der Gleichung sagt, dass diese Fehlerfunktion eigentlich halb so viel beträgt, wie wenn man die Länge des Differenzvektors zwischen den idealen Werten $\vec{y}$ und den realen Werten $\vec{a}$ quadriert. Mathematiker berechnen den Anteil des Fehlers für den s-Laut, indem sie diese Fehlerfunktion nach a_s ableiten. Wobei a_s die Aktivierung des s-Lautes bedeutet:

$$\frac{dC}{da_s} = \frac{1}{2} * 2 * (y_s - a_s) = 100\,\% - 60\,\% = 40\,\%$$

Damit haben wir mit etwas Mathematik genau das erreicht, was wir oben schon mit einfachen Worten formulierten. Nun sollten wir diese 40 % auf den Bias, die Gewichte und die Zuleitungen verteilen. Die Aktivierung des s-Neurons der Endschicht, wir nennen sie etwas verallgemeinert l, setzt sich aus drei Komponenten zusammen: allen Zuleitungen aus den Neuronen der vorhergehenden Schicht (a^{l-1} genannt, bei NETtalk die Mittelschicht), den Gewichten, die diesen Zuleitungen gegeben wurden (w_{sk} genannt, wobei k das Neuron aus der vorhergehenden Schicht beschreibt) und dem Bias b_s. Der Einfachheit halber denken wir uns, dass wir den Fehler schön gerecht verteilen: Wenn der Bias 15 % ausmacht, dann trägt er 15 % des Fehlers. Ebenso die Gewichte beim s-Neuron und die Aktivierungsfunktion des Neurons einer Vorgängerschicht. Allerdings ist dies leichter gesagt als getan. Diese drei Komponenten werden ja dann noch mit der Aktivierungsfunktion umgewandelt. Wir nennen sie σ. Um diese Überlegungen mathematisch zu fassen, formulieren wir zuerst die Aktivierung des s-Neurons in der Schicht l.

$$a_s = \sigma\left\{\sum_k w_{sk} * a_k^{l-1} + b_s\right\}$$

Wie gesagt, dies ist die Aktivierung des s-Neurons in der Endschicht: Der ganz Input, inklusive Bias, wird durch die Aktivierungsfunktion hindurchgepresst und gibt dann den Ausgangswert des s-Neurons (60%).

Wenn die Aktivierungsfunktion eine blosse Multiplikation mit sagen wir 0.5 gewesen wäre (ReLU), könnten wir deren Wirkung leicht rückgängig machen und den Fehler (40 %) halbieren. Dann müssten die Werte innerhalb der geschweiften Klammer um total 20 % oder 1/5 nach oben korrigiert werden. In der Praxis wird bei einem Lernschritt aber bloss 1/100 oder 1/1000 korrigiert. Bei einem Hundertstel wäre die Korrektur noch 1/100 von 20 % oder eben 0.002; also müssten die Zahlen insgesamt um 2 Promille erhöht werden. Nun müssen wir diese 2 Promille auf die b_s, die w_{sk} und die a_k^{l-1} aufteilen. Da dies alles Zahlen sind, könnten wir

anteilmässig verteilen. Wir werden später zeigen, dass dieses naive Vorgehen dem Problem nicht gerecht wird.

Damit ist die Korrektur für die letzte Schicht erledigt. Wie wird aber die vorletzte (l-1) korrigiert? Indem das a_k^{l-1} weiterverfolgt wird. Der Fehler auf dieser Grösse muss zum k-ten Neuron auf der l-1-Schicht zurückverfolgt werden. Dieser (winzige) Anteil am Fehler kann aber genauso behandelt werden wie der 40 %-Fehler beim s-Neuron der Endschicht. All die Ideen, die wir entwickelt haben, werden auf die l-1-Schicht übertragen. Wenn wir nur *eine* Mittelschicht wie beim NETtalk hätten, müssten wir noch die a_k^{l-2} korrigieren und wären dann schon bei der Eingangsschicht.

Gradientenabstieg

Wir wollen nun begründen, warum das anteilsmässige Verteilen der Gewichte nicht richtig ist. Dazu stellen wir uns ein Neuron der Mittelschicht vor, das eine relativ starke Verbindung zum von uns betrachteten s-Phon der Endschicht hat. Zudem soll dieses Neuron aber eine ebenso starke Verbindung zum k-Phon der Endschicht haben – das ja beim ersten c von circuit in unserem Beispiel zu stark aktiviert ist. Wenn wir nun dieses Neuron zusätzlich stärken, dann wir auch die fehlerhafte Zuleitung zum k-Phon gestärkt!

Das mathematische Verfahren, das heute angewendet wird, um die Gewichte anzupassen, heisst stochastischer Gradientenabstieg (stochastic gradient descent). Wenn wir das Lernen des KNN als Optimierungsprozess betrachten, der wie in § 11.4 beschrieben, den Abstieg von einer Passhöhe untersucht, dann ist der Gradient die Richtung des Abstiegs, mit der man am schnellsten Höhe verliert. Karim Raimi stellt in seinem Blog diesen Abstieg in vereinfachter Form dar: Er verwendet bloss zwei Eingangsneuronen und ein einzelnes Ausgangsneuron. Daran illustriert er das Lernen an 6 Übungsbeispielen und erklärt anschaulich die Begriffe des Gradentenabstiegs, warum er stochastisch genannt wird und was die Begriffe batch (einzelne Lerneinheit) und epoch (einmaliger Durchgang durch alle Lernbeispiele) in der Praxis bedeuten.[18]

[18] https://towardsdatascience.com/step-by-step-tutorial-on-linear-regression-with-stochastic-gradient-descent-1d35b088a843 (29.04.2021)

11.2.4 Der mathematische Formalismus am einzelnen Neuron

Wir haben uns das Leben massiv vereinfacht, indem wir die Sigmoid-Funktion durch eine blosse Multiplikation ersetzt haben. Nun wollen wir die Kraft des Formalismus wirken lassen, indem wir die Sigmoid-Funktion als Aktivierung zu Grunde legen, aber die Fehlerkorrektur nur an einem einzelnen Neuron untersuchen. Wir lassen uns von Michael Nielsens Vorschlag leiten und stellen uns ein Neuron vor, das eine *einzige* Zuleitung hat, deren Stärke x = 1 ist (Nielsen 2015, Kapitel 3, S.2 ff.). Das Neuron soll diese Verbindung mittels Gewicht, Bias und Sigmoid-Funktion auf null reduzieren. Dies ist ein Lernbeispiel für überwachtes Lernen und der Überwacher sagt, y sei null. Wir wollen nun die Anpassung des Gewichts und des Bias untersuchen. Die Sigmoidfunktion wird mit dem Input z gespeist:

$$z = w * x + b$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Wenn wir das Neuron z.B. mit einem Gewicht von 0.6 und einem Bias von 0.9 arbeiten liessen, dann würde der Output (σ(0.6*1 + 0.9)) den Wert 0.82 ergeben. Er ist weit von null entfernt und wir sollten nun die Gewichte korrigieren, indem wir die Kostenfunktion C(w,b) optimieren. Sie ist bei einem einzigen Neuron denkbar einfach:

$$C(w,b) = \frac{(y-a)^2}{2} = \frac{(y-\sigma(z))^2}{2}$$

Wobei nun für die Aktivierung a die Sigmoidfunktion σ(z) verwendet wurde. Die Ableitungen von C(w,b) nach w und b sind etwas tricky, weil die w und b im z verborgen sind und deshalb eine äussere Ableitung nach σ und eine innere Ableitung von σ nach z und dann nochmals eine innere Ableitung von z nach w oder b nötig ist. Wir müssen die sogenannte Kettenregel anwenden. Die äussere Ableitung von C nach σ ist nun:

$$\frac{dC}{d\sigma} = \frac{1}{2} 2 * (y-\sigma) * (-1) = (\sigma - y)$$

Die innere Ableitung nach σ führen wir noch nicht aus und schreiben einfach σ'. Die weitere innere Ableitung von z nach w gibt x und die von z nach b gibt 1. Zusammengesetzt bekommen wir das beeindruckende Resultat:

$$\frac{\partial C}{\partial w} = (\sigma - y) * \sigma' * x = \sigma * \sigma'$$

$$\frac{\partial C}{\partial b} = (\sigma - y) * \sigma' * 1 = \sigma * \sigma'$$

Wobei nun y = 0 und x = 1 gesetzt wurden. Jetzt sind wir fast fertig und könnten die Veränderungen von w und b numerisch berechnen. Es bleibt uns noch zu zeigen, dass die Ableitung von σ wieder lauter Sigmas ergibt. Man schreibt am besten

$$\sigma(z) = \frac{e^z}{(1 + e^z)}$$

und verwendet die Quotientenregel, so dass man – nach etwas Rechnerei – bekommt:

$$\sigma' = \sigma * (1 - \sigma)$$

Damit ergibt sich für beide Ableitungen: σ^2 - σ^3. Die numerische Auswertung kann man mit einem Excel-File ausführen und erhält:

w	**b**	**z**	**σ**	**σ^2-σ^3**	**Δz**	**z_{neu}**	**a, σ_{neu}**
0.600	0.900	1.500	0.818	0.122	0.183	1.317	0.789
2.000	2.000	4.000	0.982	0.017	0.069	3.931	0.981

Fig. 11.2 b: Numerische Auswertung der Korrektur an einem einzelnen Neuron

In der ersten Zeile sieht man, dass die Korrektur in die richtige Richtung geht, aber noch nicht so beeindruckend ist: 0.789 ist noch weit von null entfernt, aber besser als 0.818. Zu denken geben sollte uns die zweite Zeile. Die Gewichte sind hier weit von einem Optimum entfernt. Die Korrektur bewirkt nur 1 Promille Verbesserung. Nielsen lässt den Leser in seinem Onlinebuch mit einer Simulation dieser Fehlerkorrektur experimentieren. Sie zeigt, dass im Fall 2 das Netzwerk nur mühsam lernt, während es im ersten Fall 1 schnell zu einer guten Lösung gelangt.

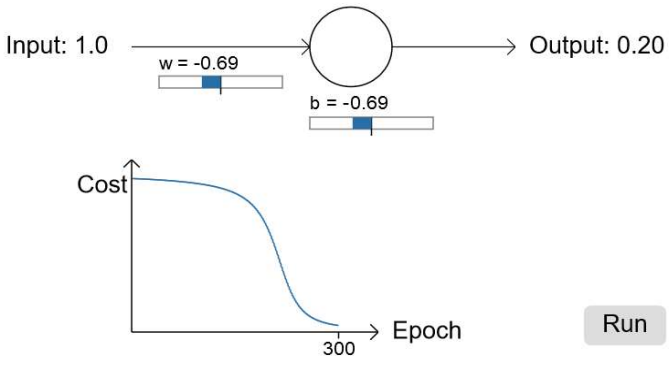


Fig 11.2 c: Schlechtes Lernen bei ungünstigen Anfangswerten w = 2 und b = 2

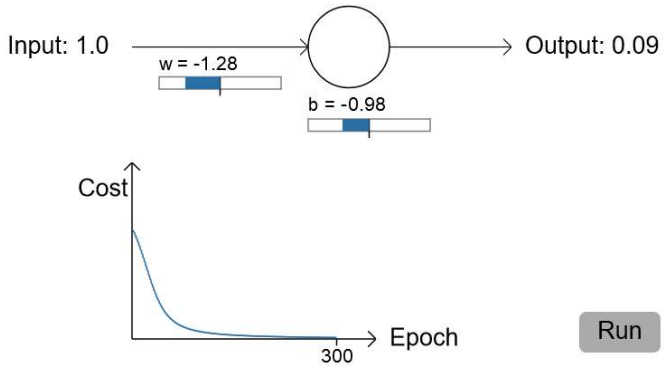


Fig. 11.2 d: Gutes Lernen bei günstigen Anfangswerten w=0.6 und b = 0.9

Wir stossen hier auf die bereits erwähnte Problematik der Sigmoid-Funktion. Bei grossen Werten (z > 4) ist sie flach und deren Ableitungen, die Steigungen, sind beinahe null. Sie kann deshalb den Fehler nicht effizient zurückspiegeln. Wenn nach unendlich langem Üben der Input z dann in die Nähe von 1 oder 2 kommt, dann wirkt die Fehlerkorrektur. Zu guter Letzt könnten wir nun die Fehlerfortpflanzung längs der Zuleitung ausrechnen, indem wir nach x ableiten:

$$\frac{dC}{dx} = (\sigma - y) * \sigma' * w = w * \sigma * \sigma'$$

Damit wissen wir, welche Korrektur auf das vorangehende Neuron übertragen wird: Mit dem Gewicht 0.6 und dem Bias 0.9 wäre das 0.6*0.122. Wenn wir nun bedenken, dass x eigentlich die Aktivierung des Neurons in der vorangehenden Schicht ist und sie aus $\sigma(w^{l-1}*x^{l-1} + b^{l-1})$ entstanden ist, dann können wir die Gewichte des Neurons dieses Layers anpassen. Statt das Korrekturgewicht wie oben zu berechnen, setzen Mathematiker dieses a^{l-1} anstelle von x ein und differenzieren nun nach w^{l-1} und b^{l-1}. Das ergibt natürlich imposant aussehende, komplizierte Funktionen. Die Differenzierung besteht aber nur in der fortwährenden Anwendung der Kettenregel. Die Backpropagation spannt somit eine Kette bis zu den Eingangsneuronen. Bitte beachten Sie, dass wir bei dieser elementaren Einführung die Lernrate nicht berücksichtigt haben; sie ist auf 1 gesetzt. Im Normalfall wäre sie eine kleine Zahl und würde die Korrekturen zusätzlich abschwächen.

Wenn man den oben dargestellten Formalismus von *einem* Neuron auf *alle* Neuronen einer Ebene ausweitet, dann ergeben sich für y, a und b lange Zahlenreihen; sie kann man als Vektoren auffassen. Die Gewichte dagegen werden zu rechteckigen Zahlenmustern, so genannten Matrizen, weil jedes einzelne Neuron j mit jedem Vorgängerneuron k verbunden sein kann. Ein einzelnes Element einer solchen Matrix würde dann als w_{jk} bezeichnet. Diese Verallgemeinerungen helfen beim Programmieren eines KNN enorm. Sie zu verstehen, vertieft aber das prinzipielle Verständnis der Fehlerkorrektur nicht.

Die Informatikerinnen und Mathematiker waren sehr erstaunt, dass diese Art des Lernens zu einem stabilen Resultat führt und KNN derart erfolgreich machte. Unter § 11.4 erkläre ich, was der Grund für diese Überraschung sein könnte.

11.2.5 Hyperparameter-Tuning

Auf Grund der Analyse, wie ein KNN lernt, sehen wir, dass der Überwacher zwei Möglichkeiten hat, das Netzwerk zu trainieren: Er kann die Verbindungsstärken (Gewichte und Bias) anpassen – er könnte aber auch die Struktur des Netzwerkes optimieren, indem er die Anzahl der Neuronen, den Grad der Fehlerabschwächung (Lernrate), die Übertragungsfunktion oder die Kostenfunktion verändert. Das Optimieren des Systems an sich nennt man *Hyperparameter-Tuning*. Lange Zeit wurde es von Hand gemacht. Die Strategien der fünften Schule, die evolutionären Algorithmen (§ 8.2), werden heute zum Teil für das Tuning des Systems eingesetzt. Eine solche Optimierung ist möglich, weil die Leistungsfähigkeit von Computern so extrem angewachsen ist.

11.3 SVM: Einfaches Beispiel

Eine SVM versucht wie gesagt, in einem Merkmalsraum Gebiete mit gleichen Objekten zu finden. Damit klassifiziert sie die Objekte. Dabei identifiziert der Algorithmus ein möglichst breites Band, das zwei Gebiete oder Kategorien voneinander trennt. Mit der Kenntnis von etwas Vektorgeometrie kann man eine SVM im einfachen Fall selber berechnen. Die folgende Darstellung lehnt sich an eine YouTube-Lektion an: https://www.youtube.com/watch?v=1NxnPkZM9bc.

Wir betrachten eine zweidimensionale Ebene mit drei Punkten: (1,1), (2,3) und (2,0). Ein Gebiet wäre von (1,1) und (2,0) bevölkert, das andere von (2,3). Die Gerade g würde die beiden Gebiete trennen. Sie wollen wir berechnen, so dass wir bei einem weiteren Punkt sofort sagen können, zu welchem Gebiet er gehört.

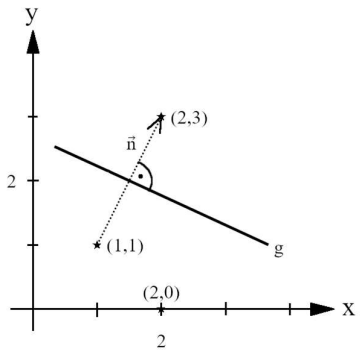


Fig. 11.3 a: Merkmalsraum

11.3.1 Hessesche Normalenform

Die trennende Gerade könnte in einem dreidimensionalen Raum eine Ebene oder in einem höherdimensionalen Raum eine Hyperfläche sein. Wir identifizieren sie mit ihrem Normalenvektor ($\vec{n}$: gepunktet) und geben deren so genannte Hessesche Normalform an. Diese Form beruht darauf, dass ein Vektor, der längs der Geraden oder in der Ebene liegt, keinen Schatten auf den Normalenvektor, der senkrecht auf der Ebene steht, werfen kann: Das Skalarprodukt mit ihm ist null. Ein Vektor längs der Geraden oder in der Ebene ist die Differenz zwischen dem allgemeinen Vektor $\vec{x}$ und dem Stützvektor $\vec{p}$ (Siehe Fig. 11.3 b). Also:

$\vec{n} * (\vec{x} - \vec{p}) = 0$ resp.

$\vec{n} * \vec{x} + \omega = 0$ (NF)

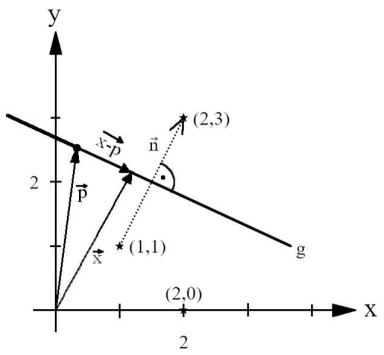


Fig. 11.3 b: Illustration der Geradengleichung

Wenn der Vektor $\vec{x}$ nicht auf einen Punkt der Geraden oder der Ebene zeigt, gibt die Hessesche Normalform seinen *Abstand* von der Geraden an; allerdings nur dann korrekt, wenn $\vec{n}$ die Länge 1 hat. Bei SVM ist es üblich, den minimalen Abstand mit -1 oder +1 zu bezeichnen. Zur Kategorie «unterhalb der Geraden» gehören deshalb alle Punkte mit Abstand grösser als -1.

Zuerst wollen wir $\vec{n}$ bestimmen, indem wir die Differenz zwischen (1, 1) und (2, 3) berechnen: Sie ist (1, 2). Der Normalenvektor schaut gegen den Punkt (2, 3). Damit wir den Abstand +1 oder -1 bekommen, setzen wir noch einen Faktor a zur Längenanpassung hinzu: $\vec{n} = a * (1\ , 2)$.

Nun können wir beide Punkte in die Normalform (NF) einsetzen und erhalten die Gleichungen:

(1 , 1): a + 2a + ω = -1

(2 , 3): 2a + 6a +ω = 1

Sodass a = 2/5 und ω= - 11/5 entstehen.

$\vec{n} = (\frac{2}{5}, \frac{4}{5})$ ist nun der Support-Vektor. Die Funktion g($\vec{x}$), die einen Punkt klassifiziert, heisst nun:

g($\vec{x}$)=2/5*x_1 + 4/5*x_2 – 11/5

Wir könnten testen, ob der Punkt (2,0) wirklich in der Minus-Kategorie liegt, also unterhalb der Geraden. Tatsächlich ist g(2,0) = 4/5-11/5 = - 6/5

Dieses Klassifizieren weiterer Punkte nennt man Generalisierung. Die Komponenten des Normalenvektors $\vec{n}$ nennt man in der KI die Gewichte und bezeichnet den Vektor als $\vec{w}$. Wir hätten die beiden Gebiete – oder die beiden Klassen – auch mit einer Ausgleichsgeraden trennen können. Diese in der Fachsprache lineare Regression genannte Methode kann man relativ leicht mit Excel berechnen. Man zeichnet die Punkte in einem Punktdiagramm und wählt dann

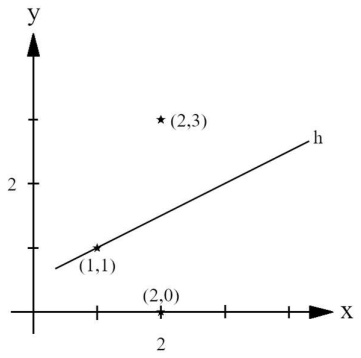


Fig. 11.3 c: Trendlinie

«Trendlinie hinzufügen». Im Gegensatz zu SVM unterscheidet die Trennlinie (h) die beiden Bereiche nicht gut. Der Punkt (1,1) liegt gar *auf* der Linie.

Mit Support Vector Machines kann man auch sehr verschachtelte Merkmalsräume analysieren und auftrennen. Russel und Norvig präsentieren ein Beispiel, bei dem die Objekte der einen Klasse (Punkte) einen Kreis abdecken, der von der zweiten Kategorie (Sterne) umgeben ist (Vgl. Russel 2012, S. 866). Ein Kreis ist ein nicht-lineares Objekt. Alle Punkte innerhalb eines Kreises um den Ursprung erfüllen aber die Bedingung
$x^2 + y^2 \leq r^2$

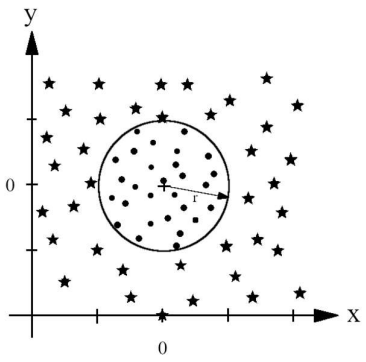


Fig. 11.3 d: Verschachtelter Merkmalsraum

Wir könnten allerdings die Eingabedaten umrechnen und sie in einem dreidimensionalen Raum mit erweitertem Koordinatensystem darstellen. Wir wählen als neue Achse $z= x^2 + y^2$. In diesem neuen Koordinatensystem sind die Objekte nun linear separierbar. Die Punkte aus dem Kreis liegen unten, nahe bei der x-y-Ebene, während die Punkte rundherum höher liegen. Dieses Umrechnen der Daten wird Kernel-Trick genannt. Die Umrechnungsfunktion heisst Kernel. Meist müssen zur Linearisierung aber höhere Dimensionen des Merkmalsraums gebildet werden.

11.3.2 Vor- und Nachteile von SVM

In unserem Beispiel haben wir es uns ein bisschen einfach gemacht. Wir gaben die Support-Vektoren vor; die Punkte (1,1) und (2,3), resp. deren Differenz ($\vec{n}$) mussten wir nicht automatisch aus den Daten heraussuchen. Zudem war die optimale Einpassung der Trennfläche denkbar einfach: Der Differenzenvektor war direkt der Normalenvektor; dies ist im Allgemeinen

nicht der Fall. Wenn wir dieses Problem auch maschinell bestimmen wollten, müssen wir etwas tiefer in die Mathematik eindringen und die Suche nach dem Normalenvektor resp. den Gewichten mit Lagrange-Multiplikatoren lösen. Damit stellt sich uns erneut eine Optimierungsaufgabe.

SVM haben den grossen Vorteil, dass nur wenige Punkte, die Support-Vektoren, betrachtet und umgerechnet werden müssen. Zudem trennen SVM die Kategorien mit einer linearen Funktion. Damit wird der Rechenaufwand massiv reduziert oder eine Klassifizierung überhaupt erst ermöglicht. Die unter Umständen komplizierte Struktur des höherdimensionalen Raumes geht in die Berechnung nicht ein: Der Klassifikator $g(\vec{x})$ besteht nur aus Produkten der Koordinaten der Supportvektoren.

11.4 Erfolgsstrategie: Keine Komplexitätsreduktion

Der Erfolg der Naturwissenschaften basierte lange auf der Strategie, Probleme auf wenige Wirkfaktoren zu reduzieren. In diesem Buch versuchte ich aufzuzeigen, dass sich heute eine zusätzliche Betrachtungsweise aufdrängt: das Denken in komplexen Systemen. Es stützt sich wesentlich auf höherdimensionale Räume. Sie sind sowohl Fluch als auch Segen.

Räume mit vielen Dimensionen bergen ungeahnte Schwierigkeiten. Pedro Domingos illustriert sie mit einer Orange (Vgl. Domingos 2015, S. 187). Denken Sie sich eine Apfelsine, deren Fruchtfleisch 90% des Radius einnimmt und deren Schale die restlichen 10 % bildet: Wie gross ist die Menge des Fruchtfleisches und wie gross die der Schale? 0.9^3, also 73% sind Frucht, die restlichen 27% sind Schale. Wenn wir nun eine 80-dimensionale Orange betrachten, dann verschwindet das Fruchtfleisch! 0.9^{80} ergibt ca. 0.0002; nur noch 0.2 Promille sind Fruchtfleisch. Die Schale nimmt überhand. Der KI-Forscher Richard Bellman nennt diese Beobachtung den *Fluch der Dimensionalität*.

Räume mit vielen Dimensionen bieten aber auch eine Fülle interessanter Möglichkeiten, wie wir sie bereits bei der Analyse von verschachtelten Merkmalsräumen durch eine SVM unter Paragraph 11.3 angetroffen haben. Auch bei KNN dringen wir in Räume mit höheren Dimensionen vor. Wir hatten dies bei der Analyse des Merkmalsraums von NETtalk und den Elman-Netzen schon einmal angesprochen. Bei künstlichen neuronalen Netzwerken kann man jedes Neuron als Maschine, als Prozess, betrachten, die sehr viele Inputdaten verarbeitet und sie an unzählige nachfolgende Maschinen weitergibt. Jedes dieser Neuronen stellt eine eigene Dimension

in einem Merkmalsraum dar. Auch unser Gehirn ist ein solches Gefüge von Millionen parallel geschalteter Prozesse. Deren Komplexität übersteigt die von künstlichen Netzwerken beträchtlich. Den Neurologen und Informatiker Terry Sejnowski beschäftigt die Wirkungsweise von KNN oder deep learning-Maschinen stark. Die Qualität, mit der künstliche neuronale Netzwerke Aufgaben lösen, ist auf der Grundlage unseres heutigen Kenntnisstandes unerklärlich. Der Titel seines Übersichtsartikels zu deep learning drückt diese Überraschung aus: *The unreasonable effectiveness of deep learning in artificial intelligence.* (Sejnowski 2020). Um diese unverschämte Leistungsfähigkeit neuronaler Netzwerke zu verstehen, müssen wir drei Aspekte bedenken:

1. Hoch-dimensionale Optimierungsprozesse verharren offenbar nicht in lokalen Minima
2. Regeln reduzieren komplexe Prozesse in vieldimensionalen Merkmalsräumen auf wenige Dimensionen
3. Die Leistungsfähigkeit künstlicher Intelligenz entstand erst, als man hochkomplexe Systeme nachbilden konnte

Räume mit mehr als drei Dimensionen sind für uns nur schwer verstehbar. Viele Mathematikerinnen und Mathematiker hielten die Optimierung eines mehrdimensionalen neuronalen Netzwerkes durch Lernsequenzen für unmöglich. Sie sagten vorher, dass ein Optimierungsprozess immer in einem lokalen Minimum hängen bleiben werde. Das war nicht der Fall und ist ein Hauptgrund für Sejnowskis Verwunderung. Um dies zu verstehen, denken wir nochmals über die Grundprinzipien eines Optimierungsprozesses nach, wie wir ihn unter § 8.2.1 mit der Fitnessfunktion zu illustrieren versuchten.

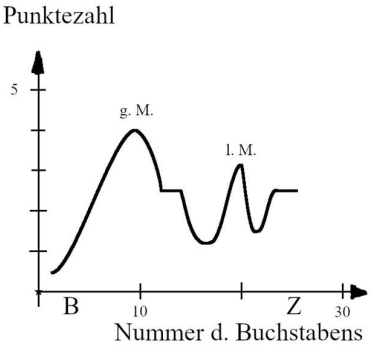


Fig. 8.2 a: Fitnessfunktion

Sie enthält globale und lokale Maximal- und Minimalwerte. Wir nehmen an, der Optimierungsprozess suche das globale Minimum. Dabei sollte er nicht in einem lokalen Minimum hängen bleiben. Ein solches lokales Minimum kann man sich als Bergsee vorstellen: Rund um ihn herum steigt das Gelände an, weiter entfernt würde es aber abfallen. Ein lokales Minimum könnte aber auch ein Bergpass sein: Auch da steigt das Gelände auf der Passhöhe teilweise steil an, aber es gibt eine Richtung in der es abfällt. Man spricht von einem Sattelpunkt. In einem höherdimensionalen Raum gibt es nun sehr viel mehr Sattelpunkte als Bergseen. Wieso? In ihnen gibt es viel mehr Dimensionen als im Dreidimensionalen: Eine Dimension braucht man in 3D für die Höhe, dann sind noch zwei Dimensionen übrig, um Bergseen oder Pässe zu konstruieren. In einem höherdimensionalen Raum hingegen lässt sich bei einem lokalen Minimum öfter eine Dimension finden, längs der es abwärts geht. Der Optimierungsprozess erkundet diese Richtung. Deshalb finden KNN erstaunlich oft eine gute Lösung.

Eine so genannte allgemeine Intelligenz versuchte bisher Regeln, nach denen Menschen geistige Tätigkeiten ausführen, zu identifizieren und damit Geist zu fassen und nachzubilden. Diese good old fashioned artifical intelligence (GOFAI) kam nicht wirklich zum Erfolg. Erst als Computer so leistungsfähig wurden, dass sie auch hochkomplexe Gefüge mit Tausenden von Dimensionen zu berechnen verstanden, ergaben sich Fortschritte. Sie sind mit dem genannten deep learning verbunden. Es braucht hochdimensionale Räume, um «intelligente» Leistungen nachzubilden. Das Einprogrammieren eines niedrigdimensionalen Regelwerks genügt nicht. Aus diesem Grunde ist es möglicherweise berechtigt, von einer neuen Zeit, dem *Informationszeitalter,* zu sprechen: *Die erfolgreiche Strategie, Naturereignisse auf Regeln zu reduzieren, wird durch die Nachbildung der Natur mit hochdimensionalen Gefügen ergänzt.*

12 Didaktische Fragen

12.1 Induktives Vorgehen

Lernenden zwei oder drei Beispiele zu präsentieren, aus denen sie dann selber eine Regel extrahieren, ist didaktisch sehr effizient. Dieses Vorgehen nennt man induktiv oder neudeutsch: bottom up. Wenn man die Regel schon hat und sie auf ein konkretes Beispiel anwenden will, dann nennt man das Vorgehen deduktiv (top down). Kindern fällt das Anwenden einer Regel schwer: Sie wissen zwar, dass man bei der Rechtschreibung fragen sollte, ob man «der», «die» oder «das» sagen kann, wenn man überlegt, ob man ein Wort gross schreibt. In der Hitze eines Diktates aber denken sie nicht an die Regeln. Allerdings fällt es auch den meisten Erwachsenen schwer, etwas «top down» zu lernen. In der Medizinausbildung gilt deshalb der Leitsatz: mehr Fälle, weniger Systematik. Computer als Lerner können ebenso eine Top down- oder eine Bottom up-Strategie verfolgen. Es ist aber sehr eindrücklich, dass die teilweise erstaunlichen Leistungen durch das Lernen an einer grossen Anzahl von Beispielen erzeugt werden. Nicht nur den Computern, auch den Schülerinnen und Schülern sollten wir Lehrer mehr Beispiele zur Verfügung stellen.

12.2 Erklären als iterativer Prozess

Kann sich ein Gehirn selbst erklären? Viele Philosophen setzen bei diesem Gedanken an, um die Kognitions- und Neurowissenschaften zu kritisieren: Ein Gehirn kann sich nicht selbst erklären! Ich betrachte dieses Argument als unzulässige Verabsolutierung. Die rationale Erklärung ist ein iterativer Prozess, der zu immer grösserer Genauigkeit gelangen kann: Der rationale Vorgang im Gehirn kann den rationalen Vorgang im Gehirn, sich selbst also, erklären, usw., usf. Es ist nicht nötig, dass dieser Prozess einen Abschluss findet. Es genügt, wenn er zu immer grösserer Genauigkeit führt. Es ist das Verdienst der Evolutionären Erkenntnistheorie, vertreten z.B. durch Gerhard Vollmer und Rupert Riedel, diesen Prozess im Detail erklärt und analysiert zu haben (Vollmer 1994).

12.3 Konstruktivismus

In diesem Text habe ich immer wieder versucht, einen schwierigen Begriff mit Beispielen zu erläutern. Selten habe ich einen Begriff einfach so definiert. Dieses didaktische Vorgehen wird oft auch Konstruktivismus

genannt. Es geht von der Überzeugung aus, dass ein Begriff aufgebaut werden muss: Er ist beim Lernenden nicht einfach vorhanden. In ihrem Buch *Where Mathematics Comes From* geben die beiden Autoren George Lakoff und Rafael Núñez eine überzeugende neurologische und sprachwissenschaftliche Begründung für dieses Vorgehen (Lakoff 2000).