# Inhaltsverzeichnis


**2 Learning from Data . . . . . 7**
1 Aus der Kursbeschreibung . . . . . 7
2 Vektoren . . . . . 7
2.1 Vektor- & Unterraum . . . . . 7
2.2 Span . . . . . 7
2.3 Basis & Dimension . . . . . 7
2.4 Lineare Unabhängigkeit . . . . . 8
2.5 Norm und Skalarprodukt . . . . . 8
3 Lineare Abbildungen . . . . . 9
3.1 Matrix . . . . . 9
3.2 Matrix-Vektor-Multiplikation . . . . . 9
3.3 Matrix-Matrix-Multiplikation . . . . . 9
4 Analysis . . . . . 10
5 Wahrscheinlichkeitstheorie . . . . . 12
6 Machine Learning . . . . . 13
6.1 Klassifikation und Regression . . . . . 13
6.2 Support Vector Machines . . . . . 13
7 Neuronale Netze . . . . . 14
7.1 Einführung . . . . . 14
7.2 Deep Learning . . . . . 15
7.3 Convolutional Neural Networks . . . . . 16
7.4 SVMs als Neuronale Netze . . . . . 16
7.5 Lineare Regression . . . . . 16

# Learning from Data

## Mathematische Theorien und Methoden für das digitale Zeitalter

***Titelbild fehlt***
***--***
***Bitte zusenden!***

## 1 Aus der Kursbeschreibung

*Philipp Moritz KL, Fanny Yang KL*

Test.

Jeder Kurs beginnt mit einem Teil „1 Aus der Kursbeschreibung«, entnommen aus dem Programmheft. Hier sollte nur der den Inhalt des Kurses beschreibende Teil der Kursbeschreibung erscheinen, nicht die Erwartungen an die Teilnehmenden.

Gegen Ende des Kurses können die Teilnehmenden je nach Interesse weitere Themen diskutieren. Es kann zum Beispiel das $H_2{}^+$-Ion als ein

## 2 Vektoren

In der linearen Algebra nehmen Vektoren eine zentrale Rolle ein. Ein Vektor wird in der Schule als eine Menge an Pfeilen gelehrt, die parallel, gleichgerichtet und gleich lang sind. Betrachten wir jedoch nicht nur den dreidimensionalen Raum, sondern $\mathbb{R}^n$, so ist die Vorstellung eines Pfeils nicht immer möglich . Dies bedingt die Notwendigkeit einer anderen Definition.

Ein *Vektor* ist eine Aufzählung von Objekten und beschreibt eine Verschiebung. Eine solche Aufzählung entspricht der Definition eines *Tupels*. Entscheidend bei Tupeln ist die Reihenfolge der Objekte, die auch mehrfach vorkommen können.

$$\text{n-Tupel } (a_1, a_2, \ldots, a_n)$$

### 2.1 Vektor- & Unterraum

Vektoren bilden die Elemente eines *Vektorraumes V*. Addieren wir zwei Vektoren eines Vektorraumes oder multiplizieren wir sie mit einem Skalar, so ist die Summe bzw. das Produkt ebenfalls ein Element des Vektorraumes.

$$\begin{aligned}(1)\ & u, v \in V,\ u + v \in V \\ (2)\ & u \in V, \lambda \in \mathbb{R},\ \lambda u \in V\end{aligned}$$

Ein *Unterraum* ist eine Teilmenge eines Vektorraumes. Es gelten für sie die oben genannten Eigenschaften eines Vektorraumes.

$$\begin{aligned}U \subseteq V, &\text{wenn gilt:} \\ (1)\ & u, v \in U,\ u + v \in U \\ (2)\ & u \in U, \lambda \in \mathbb{R},\ \lambda u \in U\end{aligned}$$

### 2.2 Span

Alle möglichen Vektoren eines Vektorraumes werden durch den sogenannten *Span* dargestellt. Der Span ist die Menge aller möglichen Linearkombinationen der Basisvektoren.

$$\begin{aligned}\text{Span } (V_1, \ldots, V_k) = \{U \in V : U = \lambda_1 v_1 + \cdots + \lambda_k v_k\} \\ \text{mit } \lambda_1, \ldots, \lambda_k \in \mathbb{R}\end{aligned}$$

### 2.3 Basis & Dimension

Die Menge der Basisvektoren wird *Basis* eines Vektorraumes genannt. Jeder Vektorraum besitzt eine *Dimension p*, die durch die Anzahl der Basisvektoren bestimmt wird.

**Übung**
Beweise, dass ein Span immer ein Vektorraum ist.
Beweis Wir nehmen an:

$$U \in Span(V_1, \dots, V_k)$$
$$V \in Span(V_1, \dots, V_k)$$

$$U = \lambda_1 V_1 + \cdots + \lambda_k V_k$$
$$V = \alpha_1 V_1 + \cdots + \alpha_k V_k$$

Additionsregelung bei Vektorräumen (s.(1))

$$\begin{aligned} W_1 &= U + V \\ &= \lambda_1 V_1 + \cdots + \lambda_k V_k + \alpha_1 V_1 + \cdots + \alpha_k V_k \\ &= (\lambda_1 \alpha_1) V_1 + \cdots + (\lambda_k \alpha_k) V_k \end{aligned}$$

Multiplikationsregelung bei Vektorräumen (s.(2))

$$W_2 = \lambda_1 V_1 + \cdots + \lambda_k V_k$$
$$\lambda W_2 = (\lambda \lambda_1) V_1 + \cdots + (\lambda \lambda_k) V_k$$

□

## 2.4 Lineare Unabhängigkeit

$v_1, \dots, v_k$ mit $v_i \in V$ sind *linear unabhängig*, falls $\lambda_i v_i + \cdots + \lambda_k v_k = 0$ nur für $\lambda_i = \cdots = \lambda_k = 0$ gilt.

## 2.5 Norm und Skalarprodukt

Definition 1 $\|v\|$ ist eine Norm im Vektorraum $V$ falls

1. a) $\|v\| \geq 0$ für alle $v \in V$
   b) $\|v\| = 0$ nur für $v = 0$
2. $\|u + v\| \leq \|u\| + \|v\|$ für alle $v \in V$
3. $\|\lambda v\| = |\lambda| \|v\|$ für alle $\lambda \in \mathbb{R}$ und $v \in V$

Eine Norm kann nie negativ sein, ist die Norm 0 bedeuted das, dass der Vektor der Nullvektor ist. Wenn man zwei Vektoren addiert, ist die Norm des resultierenden Vektors genau die Summe der Normen der beiden Ausgangsvektoren. Außerdem ist die Norm eines Vektors, der mit einem reelen Faktor multipliziert wurde gleich dem Produkt aus der Norm des Vektors und dem Betrag des Faktors.

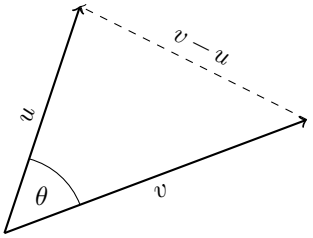
Definition 2 $f : V \times V \to \mathbb{R}$ ist ein Skalarprodukt wenn

1. $\langle \alpha u + \beta v, w \rangle = \alpha \langle u, w \rangle + \beta \langle v, w \rangle$ (Bilinearität)
2. $\langle u, v \rangle = \langle v, u \rangle$ für alle $u, v \in V$
3. $\langle v, v \rangle \geq 0$ mit $\langle v, v \rangle = 0$ nur für $v = 0$

Üblicherweise wird das Skalarprodukt mit der folgenden Formel berechnet.

$$\langle u, v \rangle = u_1 v_1 + \cdots + u_n v_n$$

An der ausgeschriebenen Form lässt sich überprüfen, ob es sich per Definition dabei wirklich um ein Skalarprodukt handelt. In der ausgeschriebenen Form wird aus $\langle \alpha u + \beta v, w \rangle$ dabei $(\alpha u_1 + \beta v_1) w_1 + \cdots + (\alpha u_n + \beta v_n) w_n$. Durch Ausmultiplizieren der Produkte erhalten wir die Form $\alpha u_1 w_1 + \beta v_1 w_1 + \cdots + \alpha u_n w_n + \beta v_n w_v$. Wenn wir daraus nach zwei Summen sortieren und die Summanden mit $\alpha$ von denen mit $\beta$ trennen können, wir die beiden Faktoren ausklammern und gelangen zu der aus der Definition geforderten Form $\alpha \langle u, w \rangle + \beta \langle v, w \rangle$.
Die zweite Bedingung können wir mit dem Kommutativgesetz beweisen, wenn wir dieses auf die ausgeschriebene Form anwenden. Für die dritte Bedingung ergibt sich $v_1 v_1 + \cdots + v_n v_n = v_1^2 + \cdots + v_n^2$. Da durch das Quadrieren keiner der Summanden kleiner als 0 werden kann und auch nur für $v_i = 0$ genau 0 werden kann, ist auch diese Bedingung erfüllt.




Da grundsätzlich $\langle v, v \rangle = \|v\|^2$ gilt, kann über die Abbildung auch die Aussage getroffen werden, dass $\|v - u\|^2 = \langle v - u, v - u \rangle$. Mit der Bilinearität des Skalarproduktes und der eben getroffenen Aussage können wir die Gleichung zu

$$\|v - u\|^2 = \|v\|^2 - 2 \langle v, u \rangle + \|u\|^2$$

umformen. Zusätzlich folgt aus dem Kosinussatz, dass $\|v - u\|^2 = \|u\|^2 + \|v\|^2 - 2\|u\|\|v\| \cos\theta$. Setzt man beide Terme gleich, gelangt man über einige wenige Umformungen zu der Form $\langle u, v \rangle = \|u\|\|v\| \cos\theta$, was ebenfalls eine Möglichkeit ist, das Skalarprodukt darzustellen.
Aus dieser Gleichung ergibt sich direkt die Cauchy-Schwarz Ungleichung. Da der Betrag des Kosinus sich nur zwischen 0 und 1 bewegt, gilt $\langle u, v \rangle \leq \|u\|\|v\|$.
Eine häufig genutzte Norm ist die Euklidische Norm, welche mit $\|v\|_2 = \sqrt{v_1^2 + \cdots + v_n^2}$ definiert ist. Mit den zuvor getroffenen Aussagen können wir beweisen, dass es sich bei der Euklidischen Norm wirklich um eine Norm handelt.
Die erste und die dritte Bedingung aus der Definition für eine Norm zeigen sich aus der ausgeschriebenen Form der Euklidischen Norm. Da alle

Komponenten dabei quadriert werden und die Wurzel gezogen wird muss die erste Bedingung stimmen. Ein Faktor $\lambda$, der ebenfalls quadriert in der Wurzel steht kann aus der Summe ausgeklammert werden und als Faktor vor die Wurzel gelangen. Der Betrag ergibt sich dabei daraus, dass $\lambda$ nach dem Vorgang nur positiv sein kann, auch wenn es zuvor negativ war.
Um zu beweisen, dass $\|u+v\| \leq \|u\| + \|v\|$ formen wir diese Ungleichung zu der bekannten Cauchy-Schwarz Ungleichung um. Hierzu quadrieren wir zunächst beide Seiten, ersetzen die quadrierten Normen durch Skalarprodukte und erhalten

$$|\langle u+v, u+v\rangle| \leq \langle u,u\rangle + 2\|u\|\|v\| + \langle v,v\rangle$$

Die linke Seite der Ungleichung kann aufgrund der Bilinearität als $|\langle u,u\rangle| + 2|\langle u,v\rangle| + |\langle v,v\rangle|$ geschrieben werden. Dadurch können wir auf beiden Seiten $|\langle u,u\rangle|$ und $|\langle v,v\rangle|$ subtrahieren und kommen so auf $\|u+v\| \leq \|u\| + \|v\|$, womit bewiesen ist, dass die Euklidische Norm eine Norm ist.

# 3 Lineare Abbildungen

In diesem Abschnitt werden wir die elementarste Form einer Funktion kennenlernen, die einen Vektorraum auf einen anderen Vektorraum abbildet. Diese Funktionen nennt man *lineare Abbildungen*.

Definition 3 Eine Funktion $f : U \to V$ ist linear, wenn folgende Bedingungen gelten:
1.
$$f(u+v) = f(u) + f(v) \forall\ u, v \in U$$

2.
$$f(\lambda v) = \lambda f(v)\ \forall\ v \in U\ und\ \lambda \in \mathbb{R}$$

Das besondere an solchen Abbildungen ist, dass sie von *Matrizen* der Form $A \in \mathbb{R}^{n\times m}$ definiert werden, wobei $n$ die Dimension von $U$ und $m$ die Dimension von $V$ ist.

## 3.1 Matrix

Eine Matrix ist eine spezielle Anordnung von $n \cdot m$ Zahlen in $n$ Zeilen und $m$ Spalten. Man kann eine Matrix aber auch als eine Zusammenfassung von $m$ Vektoren mit $n$ Komponenten ansehen, dabei bildet ein Vekotor eine Spalte. Die Komponente in der $i$-ten Zeile und $j$-ten Spalte wird als $a_{ij}$ geschrieben und die ganze Spalte mit $a_j$ notiert.

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1m} \\ a_{21} & a_{22} & ... & a_{2m} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nm} \end{pmatrix} = (a_1, a_2, ..., a_m)$$

Da wir nun wissen was eine Matrix ist, führen wir zwei neue Operationen ein.
1. Matrix-Vektor-Multiplikation
2. Matrix-Matrix-Multiplikation

## 3.2 Matrix-Vektor-Multiplikation

Bei einer Matrix-Vektor-Multiplikation wird eine Matrix $A \in \mathbb{R}^{n\times m}$ mit einem Vektor $v \in \mathbb{R}^m$ multipliziert. Dabei wird jede Zeile der Matrix mit dem Vektor skalarmultipliziert; deshalb ist es wichtig, dass die Matrix so viele Spalten wie der Vektor Dimensionen hat. Eine Multiplikation zwischen Matrizen und Vektoren, bei denen die Anzahl nicht übereinstimmt, ist nicht definiert.

$$\begin{aligned} Av &= \begin{pmatrix} a_{11} & a_{12} & ... & a_{1m} \\ a_{21} & a_{22} & ... & a_{2m} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nm} \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ ... \\ v_m \end{pmatrix} \\ &= \begin{pmatrix} a_{11} \times v_1 + a_{12} \times v_2 + ... + a_{1m} \times v_m \\ a_{21} \times v_1 + a_{22} \times v_2 + ... + a_{2m} \times v_m \\ ... \\ a_{n1} \times v_1 + a_{n2} \times v_2 + ... + a_{nm} \times v_m \end{pmatrix} \end{aligned}$$

*2.1: Matrix mit Vektor multiplizieren*

Allgemein lässt sich für die Matrix-Vektor-Multiplikation sagen

$$(Av)_i = \sum_{j=1}^{n} a_{ij} v_j$$

## 3.3 Matrix-Matrix-Multiplikation

Eine Matrix-Matrix-Multiplikation ist nichts anderes als mehrere Matrix-Vektor-Multiplikationen hintereinander. Haben wir zwei Matrizen $A \in \mathbb{R}^{n\times m}$ und $B \in \mathbb{R}^{m\times p}$ und berechnen $AB$ nehmen wir jede Spalte von $B$ als Vektor und multiplizieren ihn mit $A$ wie oben beschrieben. Die Vektoren die man als Ergebnisse erhält, fasst man wieder in einer Matrix zusammen mit der Dimension $\mathbb{R}^{n\times p}$. Allgemein können wir sagen:

$$(AB)_{il} = \sum_{n=i}^{m} a_{ij} \times b_{jl}$$

**Übung**
Beweise, dass eine Komposition $h(x) = f \circ g(x)$ aus den linearen Abbildungen $f$ und $g$, auch eine lineare Abbildung ist.
Beweis Da $f$ und $g$ linear sind gelten die Bedin-

gungen aus Definition 3:

$$\begin{aligned} h(u+v) &= f \circ g(u+v) \\ &= f(g(u+v)) \\ &= f(g(u)+g(v)) \\ &= f(g(u))+f(g(v)) \\ &= f \circ g(u) + f \circ g(v) = h(u)+h(v) \end{aligned}$$

Des Weiteren gilt:

$$\begin{aligned} h(\lambda v) &= f \circ g(\lambda v) \\ &= f(g(\lambda v)) \\ &= f(\lambda g(v)) \\ &= \lambda f(g(v)) \\ &= \lambda f \circ g(v) = \lambda h(v) \end{aligned}$$

Da beide Bedingungen aus Definition 3 für $h$ erfüllt sind, ist auch eine Komposition aus zwei linearen Abbildungen linear. □

## 4 Analysis

Analysis ist ein wichtiges Teilgebiet der Mathematik. Sie ist die Grundlage für Optimierung und somit auch von großer Bedeutung beim Maschinellen Lernen.
Im Kurs war für uns besonders die Stetigkeit und Differenzierbarkeit von ein- und mehrdimensionalen Funktionen wichtig.
Um uns Klarheit üuber diesen Bereich der Mathematik zu verschaffen, lösen wir - unter anderem - die folgenden Aufgaben.

**Aufgabe 1:**
Aufgabenstellung: Beweisen Sie, dass eine Funktion $f : \mathbb{R} \to \mathbb{R}$, welche differenzierbar an der Stelle $x \in R$ ist auch stetig an der Stelle $x \in R$ ist.

Lösung:
Zuerst überlegen wir, welche Bedingungen bereits in der Aufgabenstellung gegeben sind.
Eine Funktion ist an der Stelle x differenzierbar, wenn gilt:

$$f(x+h) = f(x) + l_x(h) + r(h) \text{ mit } \lim_{h \to 0} \frac{r(h)}{h} = 0 \tag{2.1}$$

Diese Funktion wäre außerdem stetig, wenn $\lim_{w \to x} f(w) = f(x)$.
Da die Funktion $f(x+h)$ äquivalent zur Funktion $f(w)$ sein soll, ersetzen wir in der Bedingung für die Stetigkeit $f(w)$ mit der gesamten Funktion von $f(x+h)$. Somit gilt:

$$\lim_{h \to 0} f(x) + l_x(h) + r(h) = f(x).$$

Damit die Bedingung der Stetigkeit erfüllt ist, muss der Grenzwert dieser Funktion gleich dem Funktionswert sein. Um diese Bedingung zu erfüllen, müssen wir beweisen, dass die Teile der Funktion $l_x(h)$ und $r(h)$ gegen Null gehen. Wir beginnen mit dem Teil $r(h)$. Würde $r(h)$ nicht gegen Null gehen, so würde $\frac{r(h)}{h}$ nicht gegen Null gehen. Die Funktion $r(h)$ muss gegen Null gehen, da die Bedingung der Differenzierbarkeit gilt. Jetzt müssen wir zeigen, dass $l_x(h)$ auch gegen Null geht. Dieser Teil der Funktion kann auch als $f'(x) \cdot h$ dargestellt werden. Da wir annehmen, dass h gegen Null geht und somit ein Faktor von $l_x(h)$ Null ist, wird die Funktion Null. Damit ist bewiesen, dass $\lim_{h \to 0} f(x) = f(x)$.

**Aufgabe 3:**
Bei der nachfolgenden Aufgabe soll bewiesen werden, dass es immer möglich ist für eine differenzierbare Funktion f einen Skalar $\lambda$ zu finden, damit folgende Aussage gilt:

$$(1) f(x + \lambda \nabla f(x)) \leq f(x)$$

Manche von Ihnen werden sich jetzt fragen, was das Zeichen $\nabla$ bedeutet. Es steht für den Gradienten, welcher im Grunde die Ableitung einer Funktion darstellt. Dadurch lässt sich die Ungleichung (1) in folgende Form umwandeln:

$$f(x) + Df(x)h + r(h) \leq f(x)$$

Hierbei ist h = $\lambda\nabla$ f(x) und r(h) das Restglied .
Das Ganze kann man noch weiter umschreiben, indem man das Skalarprodukt bildet:

$$f(x) + \lambda < \nabla f(x), \nabla f(x) > + r(h) \leq f(x)$$

Anschließend kann man von der Regel Gebrauch machen, dass sich ein Skalarprodukt, das in der Form $< u, u >$ vorliegt, in die Ausprágung der Euklidischen Norm von $||u||^2$ umgewandelt werden kann.
Daraus folgt:

$$f(x) + \lambda \cdot ||\nabla f(x)||^2 + r(h) \leq f(x)$$

Um nun die obige Ungleichung (1) zu beweisen, muss man zeigen, dass
$\lambda||\nabla f(x)||^2 + r(h) \leq 0$ ist.
Dafür wird $\lambda$ ausgeklammert. Anschließend erhält man:

Da $\|\nabla f(x)\|$ quadriert wird, wird dieser Term automatisch positiv. Da man den gesamten Term aber negativ haben will, wählt man $\lambda <$ o. Schließlich kennt jeder die Regel, dass ein Produkt negativ wird, falls eine ungerade Zahl von Faktoren negativ ist. Da wir $\lambda$ negativ gewählt haben, muss $\|\nabla f(x)\|^2(2)\lambda(\|\nabla f(x)\|^2 + \frac{r(h)}{\lambda}$ positiv sein.
Bei $\|\nabla f(x)\|^2\|$ stellt dies aus den eben genannten Gründen kein Problem dar. Deswegen muss nur noch gezeigt werden, dass es mindestens einen Wert von $r(h)$ gibt, der größer oder gleich 0 ist. Ansonsten würde der Wert in der Klammer negativ werden können und damit die unsprüngliche Aussage (1) widerlegen.
Damit wir zeigen können, dass $r(h)$ >= o ist, formten wir zuerst den Term (2) um:

$\lambda(\|\nabla f(x)\|^2 + \frac{r(h)}{\lambda})$

$\Leftrightarrow \lambda(\|\nabla f(x)\|^2 + \frac{r(\lambda\nabla f(x))}{\lambda})$

Bei dieser Umformung wurde h wieder mit $\lambda\nabla f(x)$ ersetzt.

$\hookrightarrow \lambda\|\nabla f(x)\| \cdot (\|\nabla f(x)\| + \frac{r(\lambda\nabla f(x))}{\lambda\|f(x)\|})$

Durch diese Umformungen schlussfolgerten wir, dass für die Funktion $h(x)$ folgendes gelten muss:
(3) $h(\lambda) = \|\nabla f(x)\| + \frac{r(\lambda\nabla f(x))}{\lambda\|\nabla f(x)\|} >= o$

Da in $\|\nabla f(x)$ kein $\lambda$ steht ist dieser Term konstant. Gleichzeitig wird durch das Bilden der euklidischen Norm festgelegt, dass ein Wert größe/gleich null vorliegt.
Deswegen muss nur noch gezeigt werden, dass dder zweite Summand ebenfalls positiv werden kann. Dazu bildeten wir den Grenzwert.
$\lim_{h\to o} \frac{r(\lambda\nabla f(x))}{\lambda\|\nabla f(x)\|} = \frac{r(\lambda\nabla f(x))}{\|\lambda\nabla f(x)\|}$

Man kann $\lambda\nabla f(x)$ wieder durch h ersetzen, woraus folgt:
$\lim_{h\to o} \frac{r(h)}{h}$

Der Grenzwert des Restbetrages strebt gegen null. Daraus lässt sich schließen, dass es einen Punkt gibt, ab welchem die Ungleichung der Funktion $h(\lambda)$ (3) erfüllt ist und damit auch die anfängliche Aussage (1).

**Aufgabe 6:**

In der letzten Aufgabe soll die Produktregel bewiesen werden.
Dafür waren zwei Funktionen gegeben: Einmal die Funktion f, die vom $\mathbb{R}^m$ in den $\mathbb{R}$ abbildet und die Funktion g, die ebenfalls vom $\mathbb{R}^m$ in den $\mathbb{R}$ abgebildet wird.
Es sind also die Funktion f(x) mit der Ableitung $f(x+h) = f(x) \cdot Df(x) \cdot h \cdot r_f(h)$ und die Funktion g(x) mit der Ableitung $g(x+h) = f(x) \cdot Dg(x) \cdot h \cdot r_g(h)$ gegeben.
Die Produktregel besagt, dass für die Funktion $h(x) = f(x) \cdot g(x)$ die Ableitung $Dh(x) = f(x) \cdot Dg(x) + Df(x) \cdot g(x)$ ist.
Damit dies bewiesen werden kann schreiben wir zuerst die Ableitung von $h(x+v)$ als Kombination aus den Ableitungen von $f(x)$ und $g(x)$ auf.

$$\begin{aligned} h(x+v) = (f(x) &+ Df(x) \cdot v \\ &+ r_g(h)) \cdot (g(x) \\ &+ Dg(x) \cdot v + r_g(v) \end{aligned} \tag{2.2}$$

Dieser Term wird ausmultipliziert und wir erhalten:

$$\begin{aligned} h(x+v) = f(x) &\cdot g(x) \\ &+ f(x) \cdot D_g(x) \cdot v \\ &+ f(x) \cdot r_g(v) \\ &+ Df(x) \cdot v \cdot g(x) \\ &+ Df(x) \cdot v \cdot D_g(x) \cdot v \\ &+ Df(x) \cdot v \cdot r_g(v) \\ &+ r_g(v) \cdot g(x) \\ &+ r_g(x) \cdot Dg(x) \cdot v \\ &+ r_g(v) \cdot r_g(v) \end{aligned} \tag{2.3}$$

Von den ganzen Summanden müsssen letztendlich alle eliminiert werden, bis auf

$$f(x) \cdot g(x) + f(x) \cdot Dg(x) \cdot v + Df(x) \cdot v \cdot g(x) \tag{2.4}$$

Dafür müssen wir zeigen, dass die übrigen Summanden gegen Null gehen. Um das zu erreichen bilden wir den Grenzwert:

Damit die Summanden einzeln äufgesplittet"werden können, wenden wir die Dreiecksungleichung an:

Nun können wir von jedem Summanden einzeln den Limes bilden, um zu zeigen, dass der gesamte Grenzwert gegen Null geht. Dies ist aber nur der Fall, falls der Limes jedes Summanden gegen Null geht. Wir zeigen das exemplarisch für

$$\lim_{v\to o} \frac{\|Df(x) \cdot r_g(v)\|}{\|v\|} \tag{2.5}$$

Df(x) ist eine Matrix mit einer Zeile und m Spalten in der jeweils c steht. Daraus folgern wir:

$$\lim_{v\to o} \frac{m \cdot c \cdot \|v\| \cdot \|r_g(v)\|}{\|v\|} \tag{2.6}$$

$$m \cdot c \cdot \lim_{v\to o} \frac{\|v\| \cdot \|r_g(v)\|}{\|v\| = o} \tag{2.7}$$

Damit haben wir gezeigt, dass der Summand gegen Null geht. Das Selbe lässt sich auch mit den anderen Summanden zeigen und auch diese gehen alle gegen Null.
Die Produktregel ist damit bewiesen.

## 5 Wahrscheinlichkeitstheorie

1933 veröffentlichte der russische Mathematiker Andrey Kolmogorov sein Buch *Foundations of the Theory of Probability*, in dem er die drei Axiome der Wahrscheinlichkeitstheorie aufstellte. Es existiert ein Wahrscheinlichkeitsraum, der sich aus den drei Elementen $\Omega, \mathcal{F}, \mathbb{P}$ zusammensetzt.

1. $\Omega$ ist eine Menge mit endlicher Anzahl an Elementen. $|\Omega| = n$
2. $\mathcal{F}$ ist eine Menge der Ereignisse ḍ und eine Teilmenge von $\Omega$. $\mathcal{F} \subseteq \Omega$
3. $\mathbb{P}$ bezeichnet die Sicherheit der Wahrscheinlichkeit, das Wahrscheinlichkeitsmaß. Dieses weist jedem Ereignis in $\mathcal{F}$ eine reelle Zahl zu. $\mathbb{P} : \mathcal{F} \longrightarrow \mathbb{R}$

— **Kolmogorovs Axiome lauten:**

1. Die Wahrscheinlichkeit eines Ereignisses ist eine positive, reelle Zahl.
   $\mathbb{P}(\mathbb{E}) \geq 0$ für alle $\mathbb{E} \in \mathcal{F}$.
2. Die Menge aller Ergebnisse bezeichnet man als sicheres Ereignis, das die Wahrscheinlichkeit 1 hat. $\mathbb{P}(\Omega) = 1$
3. Die Wahrscheinlichkeit der Vereinigung von disjunkten Ereignisssen ist gleich der Summe der Wahrscheinlichkeiten der disjunkten Ereignisse.

$$\mathbb{P}(\cup_{i=1}^{m} \mathbb{E}_i) = \sum_{i=1}^{n} \mathbb{P}(\mathbb{E}_i)$$

für alle

$$\mathbb{E}_1, ..., \mathbb{E}_m \in \mathcal{F}$$

$\mathbb{E}_1, ..., \mathbb{E}_m$ ist disjunkt, falls $\mathbb{E}_i \cap \mathbb{E}_j = \emptyset$ für alle $i, j = 1, 2, ..., m$. Hierbei bezeichnet Ø ein unmögliches Ereignis $\emptyset \in \mathcal{F}$.
Im Gegensatz dazu gibt es das sichere Ereignis mit $\Omega \in \mathcal{F}$.

Die Wahrscheinlichkeit wird errechnet durch $\mathbb{P} = \frac{|\mathbb{E}|}{|\Omega|}$. Alternativ definiert man:

$p_i \geq 0$ für $i \in \Omega$, sodass $\sum_{i \in \Omega} p_i = 1$, $p_i = \mathbb{P}(i)$

— **Konsequenzen der Axiome:**

1. $\mathbb{P}(\emptyset) = 0$
2. Monotonie: $A \subseteq B$ daraus folgt: $\mathbb{P}(A) \leq \mathbb{P}(B)$
3. $0 \leq \mathbb{P}(E) \leq 1$
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

— **Zufallsvariablen**

In einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ beschreibt die Funktion $X : \Omega \longrightarrow \mathbb{R}^k$ einen k-dimensionalen Zufallsvektoren. Im Fall von $k = 1$ redet man von einer Zufallsvariablen.

— **Erwartungswert**

Der Erwartungswert einer Zufallsvariablen beschreibt die Zahl, die die Zufallsvariable im Mittel annimmt.

$$\mathbb{E}[X] = \sum_{x \in \Omega} X(i) * \mathbb{P}(i) = \sum_{x \in \mathcal{X}} X * \mathbb{P}(X = x) wobei \mathcal{X}$$

ein Wertebereich von x ist

— **Varianz**

Die Varianz kennzeichnet die Ausdehnung einer Wahrscheinlichkeit. Die Varianz einer Zufallsvariable X.

$$X = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

— **Unabhängigkeit von Ereignissen**

1. Wenn für Ereignisse $E, F \in \mathcal{F}$ gilt $\mathbb{P}(E \cap F) = \mathbb{P}(E) * \mathbb{P}(F)$ dann sind E und F unabhängig.
2. Wenn für die Zufallsvariablen

$$X_1 : \Omega_1 \longrightarrow \mathbb{R}, \mathcal{F}_1, X_2 : \Omega_2 \longrightarrow \mathbb{R}, \mathcal{F}_2$$

gilt

$$\mathbb{P}(X_1 \in A \cap X_2 \in B) = \mathbb{P}(X_1 \in A * \mathbb{P} X_2 \in B)$$

für alle $A, B \subseteq \mathbb{R}$

— **Marginal bestimmte Wahrscheinlichkeiten**

$$\mathbb{P}(X_1 \in A) = \sum_{b \in \Omega_2} \mathbb{P}X_1 \in A \cap X_2 \in b)$$

— **Bedingte Wahrscheinlichkeit**

$$\mathbb{P}(X_1 \in A | X_2 \in B) = \frac{\mathbb{P}(X_1 \in A \cap X_2 \in B)}{\mathbb{P}(X_2 \in B)}$$

— **Summe von Zufallsvariablen**

$$X_1 : \Omega \longrightarrow \mathbb{R}, X_2 : \Omega \longrightarrow \mathbb{R}$$
$$Y = X_1 + X_2$$

Annahme: $X_1$ und $X_2$ sind unabhängig.

$$Y = l = \cup_{i=-\infty}^{\infty} X_1 = 1 \cap X_2 = l - i$$

$$\mathbb{P}(Y = l) = \sum_{i=-\infty}^{\infty} \mathbb{P}(X_1 = i \cap X_2 = l - i) = \sum_{i=-\infty}^{\infty} \mathbb{P}(X_1 = l) * \mathbb{P}(X_2 = l - i)$$

— **Theorem: Liniarität des Erwartungswerts**

Angenommen, wir hätten eine Funktion
$f : \mathbb{R} \longrightarrow \mathbb{R}$ , für die gilt:

$$\int_{\mathbb{R}} f(x)dx = 1$$

dann könnten wir eine Wahrscheinlichkeitsverteilung folgendermaßen definieren:

$$\mathbb{P}(X \in A) = \int_{X \in A} f(x)dx$$

# 6 Machine Learning

Machine Learning ist ein Konzept, bei dem der Computer durch Algorythmen in der Lage ist Probleme selbstständig zu lösen. Dabei erkennt der Computer nach einer gewissen Lernphase Gesetzmässigkeiten, durch die er dann in der Lage ist, neue, ihm unbekannte Datensätze bezüglich eines bestimmten Problems zu lösen.

Dass der Computer die Gesetzmässigkeiten erkennen kann, werden verschiedene Systeme genutzt, doch zunächst einmal sollten wir uns mit den Begriffen Klassifikation und Regression vertraut machen.

## 6.1 Klassifikation und Regression

Bei Klassifikation und Regression handelt es sich um zwei verscheidene Arten von Problemen, beziehungsweise Möglichkeiten zur Problemlösung im Bereich Machine Learning.

### 6.1.1 Klassifikation

Bei der Klassifikation sorgt der Algorythmus am Ende dafür, dass der Datensatz klassifiziert wird. Das heißt, der Datensatz wird vom Computer in verschiedene Klassen eingeteilt. Die Klassen müssen vorher vom Mensch entschieden werden. Der Computer klassifiziert die Datensätze dann anhand der Attribute des entsprechenden Datensatzes. Die klassifizierten Datensätze enthalten dann meistens eine weitere Dimension, in der die Klassifizierung anhand einer Zahl gespeichert ist.

### 6.1.2 Regression

Bei der Regression geht es darum, den entsprechenden Datensätzen am Ende einen festen Wert zuzuordnen. Der Unterschied zur Klassifikation besteht darin, dass der zugeordnete Wert nicht für eine Klasse steht, sondern ein konkretes Ergebnis beinhaltet. Ein Beispiel wäre der Preis von einer Wohnung. Diese wurde mit vielen verscheidenen Attributen in die Datenbank eingespeichert (z. B. Größe, Lage ...) der Endwert wäre dann zum Beispiel der konkrete Preis und nicht eine Klassifizierung in „teuer«, „mittelteuer«, „billig«.

## 6.2 Support Vector Machines

Support Vector Machines oder kurz SVMs sind ein Konzept zum Lösen von Machine Learning Problemen. Dabei werden für alle Datensätze zürst gewisse Features festgelegt. Diese Features entsprechen den Attributen des Datensatzes. Je nach Problem kann es unterschiedlich viele Features geben, doch allgemein kann man sagen, dass, je mehr Features es sind die Genauigkeit des Programms genaür wird.

Diese Features werden normalerweise in einem Vektor für den entsprechenden Datensatz gespeichert. Dieser Vektor wird $x_i$ genannt wobei $i = 1, ..., n$ den Datensatz beschreibt mit $x \in \mathbb{R}^m$. Zusätzlich hat jeder Datensatz noch einen weiteren Wert, welcher das »Ergebnis« des Datensatzes beschreibt, der $y_i \in \mathbb{R}$ oder auch »Label« genannt wird.

Die Features und das Label zusammen beschreiben einen Punkt im n-dimensionalen Koordinatensystem, wobei $n = m+1$, da zu den m Dimensionen von x noch die eine Dimension von y dazukommt.

Durch die Features ist es dann möglich den Datensätzen einen bestimmten Ort im n-

dimensionalen Koordinatensystem zuzuordnen. Diese können dann auf Grund ihrer räumlichen Anordnung klassifiziert werden. Dabei soll der Abstand von beiden Gruppen von Punkten zu dem Trennobjekt maximal sein. Dadurch kann die Klassifizierung optimal durchgeführt werden. Andernfalls würden schon geringe Abweichungen von den Test-Datensätzen reichen, das ein Datensatz falsch klassifiziert würde. (**??**)

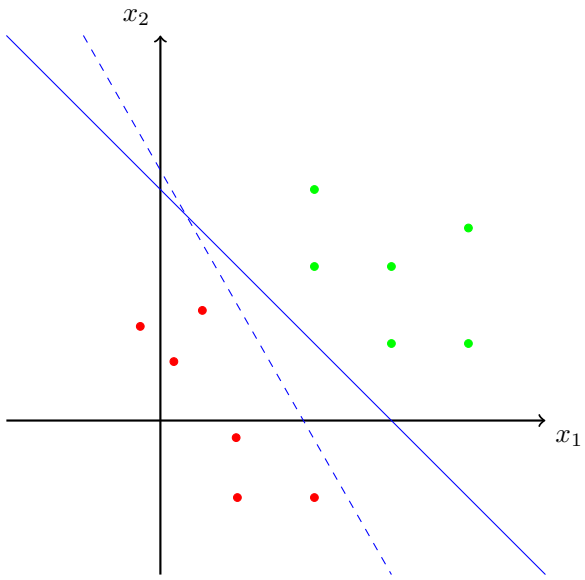



*Abb. 2.1: Da der Abstand zwischen den unterschiedlich klassifizierten Datensätzen maximiert werden soll, gilt die rote und nicht die blaue Linie als Trennelement.*

Um die Trennlinie zu optimieren, gibt es auch eine mathematische Beschreibung. Da sich Ebenen auch in der Form $\langle x, w\rangle = b$ beschreiben lassen, folgt, dass wir die Datensätze auch etwas anders betrachten können (**??**). Das daraus folgende mathematische Programm sieht wie folgt aus:

$$\begin{aligned} \text{maximiere}_{w,b} \quad & \frac{2}{\|(w)\|} \\ \text{sodass} \quad & y_i(\langle w, x_i\rangle - b) \geq 1 \quad \forall i = 1, ..., n; \\ & x \in \mathbb{R}^n; \quad y \in \{1; -1\} \end{aligned}$$

Unter der Vorraussetzung, dass alle vorhandenen Datensätze richtig klassifiziert sind. Dieses Problem kann man aber genauso schreiben als:

$$\begin{aligned} \text{minimiere}_{w,b} \quad & \|(w)\| \\ \text{sodass} \quad & y_i(\langle w, x_i\rangle - b) \geq 1 \quad \forall i = 1, ..., n; \\ & x \in \mathbb{R}^n; \quad y \in \{1; -1\} \end{aligned}$$

Klassifikation:

$$\begin{aligned} y_i = &\{1 \text{ falls} \langle w, x_i\rangle - b \geq 1\} \\ &\{-1 \text{ falls} \langle w, x_i\rangle \leq 1\} \end{aligned}$$

Nachdem das Problem optimiert wurde, ist der Computer in der Lage, weitere Daten einzuordnen, vorrausgesetzt, es ist richtig optimiert. Dennoch kann es bei dieser Art von SVMs zu einigen Problemen kommen.

- Wenn die Datensätze nicht linear separierbar sind, d.h., es ist nicht möglich die beiden Datensätze mit einem linearen Element zu trennen
- Wenn die Daten nicht alle korrekt klassifiziert sind oder es Daten gibt, die in dem »Bereich« der anderen Seite liegen.

### 6.2.1 Soft Margin SVMs

Die Soft Margin SVMs können mit dem zweiten Problem umgehen. Sie sind im Prinzip wie herkömmliche SVMs aufgebaut, nur dass sie eine gewisse Fehlertoleranz besitzen. Diese kommt durch eine Veränderung an der Formel zustande:

$$\begin{aligned} \text{minimiere}_{w,b} \quad & \|w\| + C\sum_i z_i \quad \forall i = 1, ..., n \\ \text{sodass} \quad & y_i(\langle w, x_i\rangle - b) \geq 1 \quad \forall i = 1, ..., n; \\ & x \in \mathbb{R}^n; y \in \{1; -1\} \quad z_i \geq 0 \end{aligned}$$

Dabei steht $z_i$ für die Grösse des Fehlers des Punktes $x_i$ und C ist eine Konstante, deren grösse bestimmt, wie stark die Fehler gewichtet werden, da die Konstante mit der Summe der Fehler multipliziert wird. Da die gesamte Gleichung aber minimiert werden soll sorgt ein hohes C dafür, dass das Problem schwerer optimiert werden kann. Die Konstante muss vom Mensch selbst gewählt werden, je nachdem, wie schwer Fehler gewichtet werden sollen.

## 7 Neuronale Netze

### 7.1 Einführung

Inspiriert von unserem Verständnis, wie das menschliche Gehirn lernt, benutzen Neuronale Netze Lernalgorithmen, welche besonders für praktische Anwendungen geeignet sind. Dazu zählen Spracherkennung, Objekterkennung in Bildern und die Fähigkeit individuell passende Produkte vorzuschlagen, die dem Kunden gefallen könnten. Ein Neuronales Netz wird von mehreren Schichten

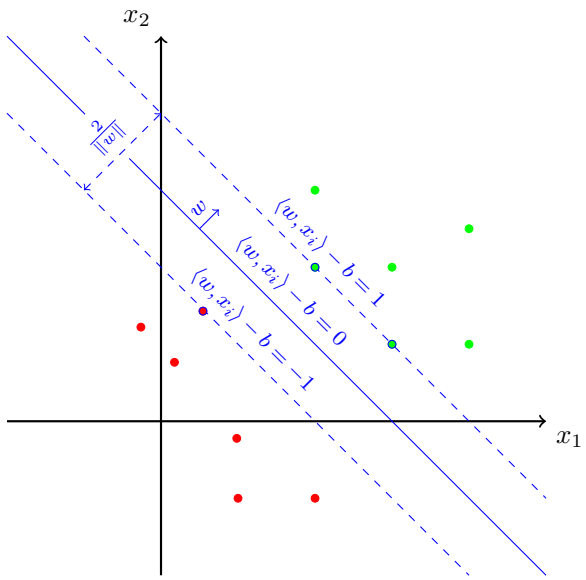



*Abb. 2.2: Links und rechts zur Trenngeraden befinden sich die parallelen Grenzen (gestrichelte Geraden). Ziel der Optimierung ist es, den Abstand zwischen den Grenzen zu maximieren, um den Normalenvektor w zu bestimmen.*

aufgebaut. Ausgangsschicht ist dabei, die Datenschicht, auf die ein oder mehrere hidden layer folgen. Als Ausgabewert erhält man schließlich einen Vektor, welcher die Wahrscheinlichkeitsverteilung darstellt. Mit anderen Worten, wie wahrscheinlich es ist, dass das Ausgangsobjekt zu einer bestimmten Klasse gehört.

## 7.2 Deep Learning

Mit Deep Learning beschreibt man die Neuronalen Netze, die über mehr als einen »versteckte Schicht« verfügen. Damit ist gemeint, dass sich zwischen Ein- und Ausgabeschicht weitere Schichten befinden.

Um die Anzahl an Parametern zu vergrößern, werden diese auch über nicht-lineare Funktionen miteinander verknüpft. Dadurch kann eine höhere Genauigkeit erzielt werden. Jedoch kann eine zu hohe Anzahl an Parametern auch dazu führen, dass das Netzwerk »overfitted«, d.h., zu sehr an den Trainingsdatensatz angepasst, wird. Dann kann das Netzwerk neue, fremde Daten nicht mehr korrekt klassifizieren.

Wie wir in der Abbildung 2.3 sehen können ist ein typisches Fully connected Neuronales Netz abgebildet. Als Basis findet man unten die Datenschicht mit ihren Eingabewerten in Form eines Vektors $x_1 ... x_i$. Diese Werte werden nun mit einem jeweiligen Faktor $w$ multipliziert. Hierbei handelt es sich um ein Skalarprodukt. Das Ergebnis wird

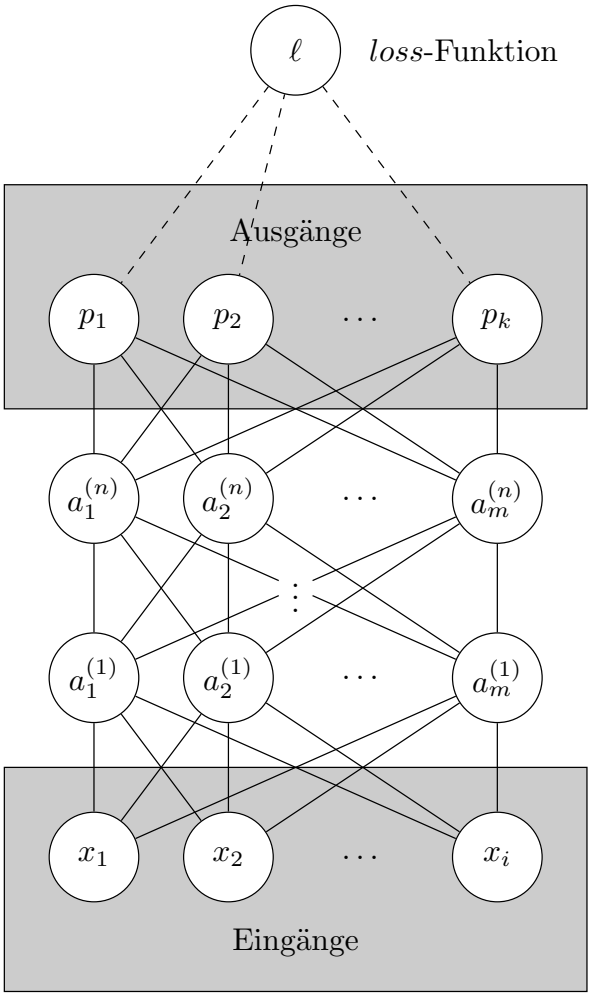



*Abb. 2.3: Ein vollständig verbundenes neuronales Netzwerk mit i Eingängen und k Ausgängen, bestehend aus n Schichten mit jeweils m »Neuronen«.*

nun im ersten hidden layer abgespeichert. Darauf wird eine nicht lineare Funktion angewendet und das Ganze wird erneut mit einem Faktor $w$ multipliziert. Dieser Vorgang wiederholt sich so lange bis uns schließlich ein Vektor mit seinen Komponenten $p_1 .. p_1$ zurückgegeben wird, welcher als Wahrscheinlichkeitsverteilung interpretiert wird. Anhand eines konkreten Beispiels würde es folgendes bedeuten: Nehmen wir an wir haben ein Bild und wollen ermitteln zu welcher Klasse Haus, Tisch oder Stuhl das daurauf abgebildete Objekt gehört. Unsere inputs wären demnach die Pixel des Bildes. Diese durchlaufen nun das Neuronale Netz und wir erhalten eine Wahrscheinlichkeitsverteilung als Rückgabewert, welche uns mitteilt, dass es am wahrscheinlichsten ist, dass das abgebildete Objekt ein Objekt der Klasse Stuhl ist. Demnach ist die Klassifizierung vollzogen.

$$a_j^{(k)} = f^{(k)}(\langle w_{ji}^{(k)}, a_i^{(k-1)} \rangle) = f^{(k)}(\sum_{i=1}^{m^{(k-1)}} w_{ji}^{(k)}, a_i^{(k-1)})$$

### 7.3 Convolutional Neural Networks

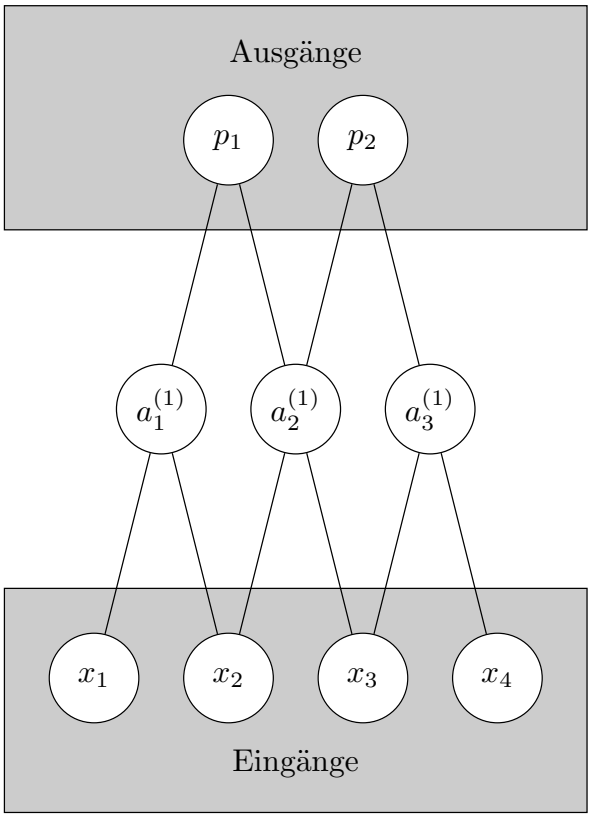



*Abb. 2.4: Ein Convolutional Neural Network (CNN) (dt.: »faltendes neurales Netzwerk«) mit vier Eingängen ($x_1, ..., x_4$) und zwei Ausgängen ($p_1, p_2$). Dazwischen befindet sich eine Schicht aus drei »Neuronen«, die als Filter wirkt.*

Bei den Convolutional Networks handelt es sich um ein Neuronales Netz mit einer vereinfachten Struktur. Diese kommt dadurch zu Stande, dass nun kein Fully Connected Neuronales Netzwerk mehr vorliegt. Das bedeutet nicht jeder der Knoten ist mit jedem verbunden, sondern nur jeweils die lokalen drei.

### 7.4 SVMs als Neuronale Netze

Wir haben bereits SVMs kennengelernt. Im folgenden wird verdeutlicht, wie man diese auch als Neuronales Netz ausdrücken kann. Charakteristisch für dieses Neuronale Netz ist, dass es nur einen Layer besitzt:

$$\min_{b,w,z} \|w\| + C \sum_{i=1}^{n} z$$

sodass:

$$y_i(\langle w, x \rangle - b) \geq 1 - z_i; z_i \geq 0$$
$$\Rightarrow z_i \geq 1 - y_i(\langle w, x \rangle - b) \text{ und } z_i \geq 0$$

$$\begin{aligned}
&\min \|w\| + C \cdot \sum_{i=1}^{n} \max(0, 1 - y_i(\langle w, x_i \rangle - b) \\
&= \|w\| + C \sum_{i=1}^{n} \ell(1 - y_i(\langle w, x_i \rangle) - b) \\
&= \|w\| + C \sum_{i=1}^{n} \ell(a_i)
\end{aligned}$$

### 7.5 Lineare Regression

Die lineare Regression kann auch in Form von Neuronalen Netzen ausgedrückt werden.

$$\min_{w,b} 0,5 \sum_{i=1}^{n} ((\langle w, x_i \rangle - b) - y_i)^2 a_i = \langle w, x_i \rangle - b - y_i$$