29 Kleines Wörterbuch der linearen Algebra (und etwas mehr)

Körper sind kompliziert zu definieren (9 Axiome). Wir haben:

K1 Eine Menge M, abgeschlossen unter zwei Operationen $+, \cdot$

K2,3 Es gibt für $+, \cdot$ jeweils ein neutrales Element, genannt $0, 1$

K4,5 Es gibt für $+, \cdot$ und alle $x \in M - \{0\}$ jeweils ein inverses Element, genannt jeweils $-x$, $1/x$, so dass $x + (-x) = 0$, $x \cdot (1/x) = 1$

K5-9 $\cdot, +$ sind kommutativ und assoziativ

K10 $(a + b) \cdot c = (a \cdot c) + (b \cdot c)$ f.a. $a, b, c \in M$

Es reicht zu wissen dass es drei bekannte Körper gibt,

1. die rationalen,
2. die reellen und
3. die komplexen Zahlen

sowie Körper, die mittels algebraischen Konstruktion (z.B. Restklassen) aus diesen konstruiert werden.

Wir bleiben hier bei den reellen Zahlen, von daher gibt es keine weiteren Probleme. Wenn wir Vektoren oder Matrizen betrachten, dann sind diese immer über Körper konstruiert, d.h. die einzelnen Komponenten sind normalerweise reelle Zahlen. Man spricht in diesem Zusammenhang auch oft Skalaren, was nichts anderes ist als eine reelle Zahl, allgemeiner: ein Objekt aus dem Körper, über den wir Vektoren, Matrizen etc. konstruieren.

Vektoren und Vektorraum, konkret : Sei K ein Körper, $n \in \mathbb{N}$ eine beliebige natürliche Zahl. Ein Vektorraum V ist eine Teilmenge

$$V \subseteq K^n,$$

die geschlossen ist unter gewissen Operationen, die wir gleich definieren. Ein Vektor

$$\vec{v} \in V$$

ist also für uns einfach ein Tupel aus n reellen Zahlen. Wir definieren das Skalarprodukt

$$\lambda \cdot (v_1, ..., v_n), \text{ wobei } \lambda \in \mathbb{R},$$

mit

(384) $\lambda \cdot (v_1, ..., v_n) = (\lambda v_1, ..., \lambda v_n)$

Die Addition zweier Vektoren ist definiert durch:

(385) $(v_1, ..., v_n) + (w_1, ..., w_n) = (v_1 + w_1, ..., v_n + w_n)$

Ein Vektorraum ist nun wie folgt definiert:

1. V ist abgeschlossen unter $+$ und $\cdot$ mit allen Skalaren.
2. V enthält 0_V, den sog. Nullvektor so dass $\vec{v} + 0_V = \vec{v}$.
3. V enthält für jeden Vektor $\vec{v}$ ein inverses Element $-\vec{v}$, so dass $\vec{v}+(-\vec{v}) = 0_V$. Natürlich ist leicht zu sehen dass $-\vec{v} = (-1) \cdot \vec{v}$.

Der wichtigste Begriff der einfachen Vektorrechnung ist der der **linearen Abbildung**: eine lineare Abbildung ist eine Funktion $f : V \to V$ so dass

(386) $f(\vec{v}) = (\lambda \cdot \vec{v}) + \vec{w}$

wobei λ ein Skalar ist, $\vec{w} \in V$. Falls V ein Vektorraum ist, f eine lineare Abbildung, dann ist auch $f[V] = \{f(\vec{v}] : \vec{v} \in V\}$ ein Vektorraum.

Linear unabhängig Ein Vektor $\vec{v}$ ist **linear abhängig** von einer Menge $\{\vec{w}_1, ..., \vec{w}_i\}$, falls es $\lambda_1, ..., \lambda_i \in \mathbb{R}$ gibt so dass

(387) $\vec{v} = \lambda_1 \vec{w}_1 + ...\lambda_i \vec{w}_i$

Man nennt das eine **Linearkombination** von $\vec{w}_1, ..., \vec{w}_i$. Eine Menge $\{\vec{w}_1, ..., \vec{w}_i\}$ ist **linear unabhängig**, falls $\vec{w}_1$ nicht linear abhängig ist von $\{\vec{w}_2, ..., \vec{w}_i\}$ (die Auswahl von $\vec{w}_1$ ist beliebig und spielt keine Rolle). Was man wissen muss:

- Eine Menge von n linear unabhängigen Vektoren der Dimension m spannt einen n-dimensionalen Teilraum des Vektorraums $V \subseteq \mathbb{R}^m$ auf.
- Es gibt jeweils nur maximal m linear unabhängige Vektoren der Dimension m. Falls also bei 1 gilt: $n > m$, dann können die Vektoren gar nicht linear unabhängig sein!

Hyperebenen stellen eine Generalisierung von Ebenen (im 3-dimensionalen Raum) auf beliebige Dimensionen dar. Eine normale Ebene ist spezifiert durch einen **Stützvektor** $\vec{s}$ und zwei linear unabhängige **Richtungsvektoren** $\vec{r}_1, \vec{r}_2$. Wichtig ist: die Richtungsvektoren müssen

1. Ungleich 0_V sein (sonst haben sie keine Richtung!)
2. Sie müssen linear unabhängig sein – sonst ist die Ebene unterdeterminiert!

Ein Punkt p liegt auf der Ebene, falls er sich darstellen lässt als

(388) $p = \lambda_1\vec{r}_1 + \lambda_2\vec{r}_2 + \vec{s}$ für $\lambda_1, \lambda_2 \in \mathbb{R}$

Diese Definition lässt sich leicht verallgemeinern: man nimmt, für einen Raum $\mathbb{R}^n$, einfach den Stützvektor $\vec{s} \in \mathbb{R}^n$ und ebenso $n-1$ linear unabhängige Richtungsvektoren $\vec{r}_1, ..., \vec{r}_{n-1}$ (ungleich 0). Technisch gesehen ist die Hyperebene eine Menge von Punkten:

(389) $H = \{\vec{s} + \lambda_1\vec{r}_1 + ... + \lambda_{n-1}\vec{r}_{n-1} : \lambda_1, ..., \lambda_{n-1} \in \mathbb{R}\}$

Norm : definiert auf einem Vektorraum V ist das eine Funktion

$$\| - \| : V \to \mathbb{R}_0^+$$

es werden also beliebige Vektoren auf einen nicht-negativen Wert abgebildet. Zusätzlich muss $\| - \|$ noch folgende Bedingungen erfüllen f.a. $\vec{v} \in V$, $\lambda \in \mathbb{R}$.

1. $\|\vec{v}\| = 0 \Rightarrow \vec{v} = 0_V$
2. $\|\lambda \cdot \vec{v}\| = |\lambda| \cdot \|\vec{v}\|$, wobei $|\lambda|$ der Betrag ist
3. $\|\vec{v} + \vec{w}\| \leq \|\vec{v}\| + \|\vec{w}\|$

Die intuitivste Norm ist die *euklidische*, die jedem Vektor seine **Länge** zuweise (wenn wir einen Vektor als eine Linie vom Ursprung auf seine Koordinaten (im n-dimensionalen Raum) auffassen. Diese Norm basiert auf einer Verallgemeinerung des Satz des Pythagoras:

$$\|(v_1, ..., v_n)\| = \sqrt{v_1^2 + ... + v_n^2}$$

In dieser geometrischen Interpretation wird Bedingung 3 zur **Dreiecksungleichung**: in jedem rechwinkligen Dreieck ist die Länge der Hypothenuse geringer als die Summe der Länge der Katheten. Es gibt aber noch viele weitere Normen, z.B. die sog. p-Norm, wobei $p \geq 1$ eine reelle Zahl ist:

$$\|(v_1, ..., v_n)\|_p = (\sum_{i=1}^{n} |v_i|^p)^{\frac{1}{p}}$$

Für $p = 1$ vereinfacht sich das zu

$$\|(v_1, ..., v_n)\|_1 = \sum_{i=1}^{n} |v_i| \tag{390}$$

Das ist die sog. Manhatten-Norm, weil man immer rechtwinklig um die Blocks fahren muss – das gibt im 2-dimensionalen Fall also die kürzeste Strecke in Manhattan an.

Welchen Unterschied machen unterschiedliche Normen? Wir betrachten das in der Ebene, weil hier die Dinge am Einfachsten sein. Man nehme den Einheitskreis um den Ursprung. Mit dem Satz des Pythagoras lässt sich leicht ableiten: sowohl $(0, 1)$ als auch $(\sqrt{0.5}, \sqrt{0.5})$ sind Vektoren, die vom Ursprung auf den Einheitskreis zeigen. Sie sind also gleich lang, rein geometrisch. (beachte: $0.5 < \sqrt{0.5} < 1$!) Es gilt auch ganz offensichtlich auch:

$$\|(\sqrt{0.5}, \sqrt{0.5})\|_2 = \|(0, 1)\|_2 = 1 \tag{391}$$

Die euklidische Norm in der Ebene ist also ein Modell der geometrischen Länge. Dagegen gilt:

$$\begin{aligned}
\|(\sqrt{0.5}, \sqrt{0.5})\|_3 &= (\sqrt{0.5}^3 + \sqrt{0.5}^3))^{1/3} \\
&= (0.5 \cdot \sqrt{0.5} + 0.5\sqrt{0.5})^{1/3} \\
&= ((\sqrt{0.5} + \sqrt{0.5}) \cdot 0.5)^{1/3} \\
&= ((2 \cdot \sqrt{0.5}) \cdot 0.5)^{1/3} \\
&= \sqrt{0.5}^{1/3} \\
&< 1
\end{aligned}$$

Dagegen:

$$\begin{aligned}
\|(0, 1)\|_3 &= (0 + 1)^{1/3} \\
&= 1
\end{aligned}$$

Was kann man also sagen? Normen mit Grad $p > 2$ bestrafen einzelne große Komponenten stärker als gleich verteilte Komponenten! Kurz: Ausreißer werden bestraft!

Die Manhattan-Norm macht übrigens genau das Gegenteil:

(392) $\|(\sqrt{0.5}, \sqrt{0.5})\|_1 = \sqrt{0.5} + \sqrt{0.5} > 1 = \|(0, 1)\|_1$

Hier belohnen wir ungleichmäßige Abweichungen (d.h. in möglichst wenigen Komponenten). Das kann durchaus Sinn machen, wenn eine einzelne Komponente gut vernachlässigt werden kann!

Metrik Darauf basiert der Begriff der Metrik; jede Norm induziert eine Metrik d mittels

(393) $d(\vec{v}, \vec{w}) = \|\vec{v} - \vec{w}\|$

wobei natürlich

(394) $\vec{v} - \vec{w} = (v_1 - w_1, ..., v_i - w_i)$

Es ist nicht schwer zu sehen dass gilt:

- $d(x, y) \geq 0$ (positiv)
- $d(x, y) = d(y, x)$ (symmetrisch)
- $d(x, y) \leq d(x, z) + d(z, y)$ (Dreiecksungleichung)

Die **euklidische Distanz** ist definiert durch die euklidische Norm, mit

(395) $d_2(\vec{v}, \vec{w}) = \|\vec{v} - \vec{w}\|_2$

Für die verschiedenen Distanzen gilt dann natürlich genau dasselbe wie für die verschiedenen Normen: je höher das p in d_p, desto stärker wird die Abweichung in einzelner Komponenten (gegenüber verteilter Abweichung in vielen Komponenten) bestraft. Neutral ist dabei die 2!

Kosinus (zweier Vektoren) Im rechtwinkligen Dreieck ist der Kosinus eines Winkels $cos(\alpha)$ definiert als

(Länge der Ankathete)/(Länge der Hypothenuse)

also $cos(90°) = 0$, $cos(0°) = 1$, und alles andere liegt dazwischen. Man kann den Kosinus auch verallgemeinert definieren als Funktion $cos : \mathbb{R} \to \mathbb{R}$, und zwar einigermaßen kompliziert als unendliche Reihe:

$$cos(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{2n!} \tag{396}$$

Was uns insbesondere interessiert ist der Kosinus zweier Vektoren. Im 2-dimensionalen Raum lässt sich das natürlich schön veranschaulichen in dem wir einfach einen Vektor als Hypothenuse auffassen, von seinem Ende eine Linie ziehen so dass sie den (verlängerten) anderen Vektor im rechten Winkel schneidet, und dann die geometrische Definition anwenden.

Allgemeiner (d.h. in beliebigen Dimensionen) lässt sich der Cosinus wie folgt berechnen: wir haben

$$\vec{x}^\top \vec{y} = \|\vec{x}\|_2 \|\vec{y}\|_2 cos\sphericalangle(\vec{x}, \vec{y}) \tag{397}$$

Daraus folgt mittels Termumformung:

$$cos\sphericalangle(\vec{x}, \vec{y}) = \frac{\vec{x}^\top \vec{y}}{\|\vec{x}\|_2 \|\vec{y}\|_2} \tag{398}$$

Skalarprodukt : Bislang haben wir Multiplikation nur mit Skalaren ausgeführt; es gab noch keine Möglichkeit, zwei Vektoren miteinander zu multiplizieren. Das macht das Skalarprodukt $\bullet : V^2 \to \mathbb{R}$. Wir multiplizieren also Vektoren und bekommen eine reelle Zahl (einen Skalar) als Ergebnis. Es gibt nicht das eine Skalarprodukt, sondern mehrere Arten es zu definieren. Wir brauchen hier nur die geläufigste, nämlich das Standard-Skalarprodukt:

$$(a_1, ..., a_n)^\top (b_1, ..., b_n) = a_1 b_1 + a_2 b_2 + ... + a_n b_n \tag{399}$$

Wenn wir Vektoren als spezielle Matrizen betrachten, ist das Standard-Skalarprodukt nur das Produkt zweier Matrizen, nämlich der Transposition der ersten und der zweiten. Wichtig ist dass wir die erste Matrix transponieren, sonst wäre das Produkt undefiniert; weiterhin setzt diese Multiplikation voraus, dass beide Vektoren die gleiche Länge haben.

Hadamard Produkt Das ist eine Funktion $\odot : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, einfach definiert durch **punktweise Multplikation**: $(x_1, ..., x_n) \odot (y_1, ..., y_n) = (x_1 y_1, ..., x_n y_n)$. Es ist also parallel zur Vektoraddition, nur mit Multiplikation.

Matrix Matrizen generalisieren Vektoren in dem Sinne dass ein Vektor $\vec{v} \in \mathbb{R}^n$ ist, eine Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ eine kompliziertere Struktur ist, bzw. ein Vektor ist eine Matrix mit $m = 1$. Matrizen schreibt man wie folgt:

$$\begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{pmatrix}$$

NB: Zeilen zuerst, dann Spalten! Um ein Objekt an einer bestimmten Koordinate Matrix zu denotieren schreiben wir $\mathbf{A}_{ij}$; z.B. im obigen Beispiel $\mathbf{A}_{23} = b_3$.

Rechnen mit Matrizen ist etwas gewöhnungbedürftig; wir haben normalerweise die Addition und die Multiplikation. Addition ist im Prinzip dasselbe wie bei Vektoren. Nimm zwei Matrizen $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, dann haben wir

$$\mathbf{A} + \mathbf{B} \subseteq \mathbb{R}^{m \times n}, \text{ und } (\mathbf{A} + \mathbf{B})_{ij} = \mathbf{A}_{ij} + \mathbf{B}_{ij}.$$

Also ist die Addition **punktweise** definiert, genau dann wenn die beiden Matrizen dieselben Maße haben.

Die Multiplikation ist etwas komplizierter. Nimm zwei Matrizen $\mathbf{A} \in \mathbb{R}^{m \times p}$, $\mathbf{B} \in \mathbb{R}^{p \times n}$, wobei wichtig ist das die beiden ps identisch sind. Wir definieren dann

$$(\mathbf{A} \cdot \mathbf{B})_{ij} = \sum_{k=1}^{p} (A_{ik} \cdot B_{kj}).$$

Diese Multiplikation ist natürlich *nicht* kommutativ; um das zu sehen, nehmen wir an $p = 1$, Matrizen $\mathbf{A} \in \mathbb{R}3 \times 1$, $\mathbf{B} \in \mathbb{R}^{1 \times 3}$; dann haben wir:

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \cdot \begin{pmatrix} b_1 & b_2 & b_3 \end{pmatrix} = \begin{pmatrix} a_1 \cdot b_1 & a_1 \cdot b_2 & a_1 \cdot b_3 \\ a_2 \cdot b_1 & a_2 \cdot b_2 & a_2 \cdot b_3 \\ a_3 \cdot b_1 & a_3 \cdot b_2 & a_3 \cdot b_3 \end{pmatrix}$$

Wenn wir das umdrehen, dann bekommen wir:

$$\begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \big((a_1 \cdot b_1) + (a_2 \cdot b_2) + (a_3 \cdot b_3)\big)$$

Also bekommen wir in diesem Fall eine 1×1-dimensionale Matrix. Es kann natürlich auch oft vorkommen dass $\mathbf{A} \cdot \mathbf{B}$ definiert ist, während $\mathbf{A} \cdot \mathbf{B}$ undefiniert bleibt. Ein wichtiges Konzept in dieser Hinsicht ist die **Einheitsmatrix**, definiert als:

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Diese Matrix ist also quadratisch, mit allen Werten 0, nur in der Diagonale hat sie den Wert 1. Einheitsmatrizen gibt es für jedes $n \in \mathbb{N}$, denn die Größe muss natürlich passen; wir denotieren die $n \times n$ Einheitsmatrix mit $\mathbf{1}_n$. Dann ist leicht zu sehen: Sei $\mathbf{A} \in \mathbb{R}m \times n$ eine beliebige Matrix. Dann ist

$$\mathbf{A} \cdot 1_n = 1_m \cdot \mathbf{A} = \mathbf{A}$$

Eigenvektor und Eigenwert Der **Eigenvektor einer Matrix** M ist ein ein Vektor $\vec{v}$ so dass gilt: $M\vec{v} = \lambda\vec{v}$ für einen Skalar λ. Also macht M mit $\vec{v}$ nichts anderes als den Vektor skalieren, anstatt seine Richtung zu ändern. λ nennt man dann den **Eigenwert**. Es ist ein schwieriges Problem, die Eigenvektoren und -werte einer großen Matrix auszurechnen, dass man im allgemeinen Fall nur näherungsweise lösen kann.

Gradienten Gradienten sind eine Generalisierung von Ableitungen für multivariate Funktionen. Ableitungen für Funktionen $f : \mathbb{R} \to \mathbb{R}$ sollten bekannt sein; intuitiv gibt der Wert der Ableitung $\frac{d}{d(x)}f$ an jedem Punkt den Grad an, in dem die Funktion zu- bzw. abnimmt. Wenn man nun eine multivariate Funktion hat, dann kann man sie nach jeder Variable *partiell ableiten.* Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine Funktion, $i \in \{1, ..., n\}$; dann haben wir mit

$$\frac{df}{dx_i} f(x_1, ..., x_n) \tag{400}$$

die Ableitung nach x_i, die uns angibt, wie die Steigung der Funktion in der i-ten Dimension verläuft; die anderen Variablen sind natürlich Parameter dieser Steigung: sie geben an wir groß sie ist an diesem Punkt im Raum. Geometrisch betrachtet: sei $f(x, y)$ eine Funktion, die für zwei Koordinaten einer Oberfläche die jeweilige Höhe liefert. Die Ableitung nach x gibt nun an, wie groß die Steigung ist, wenn ich in Richtung x gehe. Dasselbe Prinzip für höhere Dimensionen.

Der Gradient von $f : \mathbb{R}^n \to \mathbb{R}$ ist eine Funktion $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, dessen Ausgabekomponenten 1-n durch die partiellen Ableitungen von f nach $x_1, ..., x_n$ berechnet werden, d.h.

$$(401)\ \nabla f(x_1, ..., x_n) = (\frac{df}{dx_1} f(x_1, ..., x_n), ..., \frac{df}{dx_n} f(x_1, ..., x_n))$$

Alternativ kann man auch die Einheitsvektoren zur Definition nutzen; nenne die n-dimensionalen Einheitsvektoren mit $i$ter Komponente 1 $1_1, ..., 1_n$; dann ist

$$(402)\ \nabla f(x_1, ..., x_n) = \frac{df}{dx_1} f(x_1, ..., x_n)1_1 + ... + \frac{df}{dx_n} f(x_1, ..., x_n)1_n$$

Die geometrische Interpretation des Gradienten ist folgende: der Gradient einer Funktion gibt einen Vektor, und der zeigt in diejenige Richtung, in der die Funktion am schnellsten steigt. Also im Fall $f : \mathbb{R}^n \to \mathbb{R}$: die Richtung in der Ebene, in die man gehen muss um am schnellsten Höhe zu gewinnen.

Es ist klar, dass man, um ein *lokales* Maximum einer Funktion zu finden, einfach nur dem Gradienten "nachgehen" muss (mit einer gewissen *Lernrate* ϵ; wir haben das Maximum gefunden, falls der Gradient 0 an allen Stellen ist.

Lernrate Eine kleine Zahl $l > 0$, die besagt, wie wir den Wert der Funktion verändern. Was normalerweise wissen ist: wir kennen den Wert $f(x)$, und wir wissen dass für ein $\epsilon > 0$, $f(x+\epsilon) < f(x)$, oder $f(x-\epsilon) < f(x)$. Was wir nicht kennen ist ϵ. Deswegen legen wir die Lernrate l fest, die besagt, wie gross die Schritte sein sollen, die wir in die richtige Richtung gehen. Wenn l zu klein ist, brauchen wir länger (mehr Schritte) als nötig; wenn l zu groß ist, dann kann es sein dass wir den optimalen Punkt verpassen.

(Einfaches) Perzetron Ein einfaches Perzeptron hat die Form

$$(403)\ P(\vec{x}) = g(M\vec{x} + \vec{a})$$

wobei $M\vec{x} + \vec{a}$ ein lineares Modell, g eine nichtlineare Funktion. g kann im erweiterten Fall auch die Form haben $g_1 \circ g_2$, z.B. $softmax \circ ReLU$. Entscheidend ist dass es nur ein lineares layer gibt.

Multilayer Perzeptron (MLP), auch genannt feedforward network
Ein MLP ist eine Funktion

(404) $MLP = P_1 \circ ... \circ P_n$

also eine Verkettung von Perzeptronen. Man nennt MLP auch feedforward Network; sie sind die einfachsten nichttrivialen neuronalen Netze. Das Beispiel aus 404 hat $n - 1$ *hidden layer*; also hat ein einfaches Perzeptron kein hidden layer. Das offene layer in MLP wäre dasjenige von P_1.

Wenn man von einem Eingabelayer spricht, dann meint man nur soz. dummy-Neuronen, in welche die Werte der Eingabe geschrieben werden; in der funktionalen Schreibweise fällt dieses layer ersatzlos weg.