Aufgabe 1

Die Menge der n-dimensionalen Vektoren $\mathbb{R}^n$ wird zu einem metrischen Raum, wenn die Entfernung zwischen zwei Vektoren $\vec{a}$, $\vec{b} \in \mathbb{R}^n$ durch

$$d(\vec{a}, \vec{b}) = |\vec{a} - \vec{b}|$$

definiert wird (siehe Vorlesung). Bei diesem Abstandsbegriff spricht man von der euklidischen Metrik; anschaulich ist hierbei die Länge der „Luftlinie“ gleich der Entfernung zwischen zwei Punkten.

In einer Stadt mit rechtwinklig verlaufenden Straßenzügen bietet sich aber ein anderer Abstandsbegriff an:

$$d(\vec{a}, \vec{b}) = \sum_{i=1}^{n} |a_i - b_i| = |a_1 - b_1| + \ldots + |a_n - b_n|.$$

(a) Machen Sie sich klar, was diese Definition anschaulich im zwei- und dreidimensionalen Raum bedeutet.

(b) Machen sie sich anschaulich im zweidimensionalen Raum klar, daß $d(\vec{a}, \vec{b})$ alle drei Eigenschaften einer Metrik (siehe Vorlesung) besitzt.

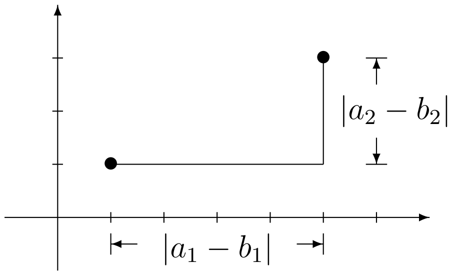
(c) Zeigen Sie formelmäßig, daß durch die Definition von $d(\vec{a}, \vec{b})$ ein metrischer Raum entsteht, d.h. daß die drei Eigenschaften einer Metrik erfüllt sind. (Hinweis: Für reelle Zahlen gilt die Dreiecksungleichung $|x + y| \leq |x| + |y|$. Was folgt damit für $|a_i - b_i| = |a_i - c_i + c_i - b_i|$?)

Lösung:

(a) Die Formel liefert die Länge des tatsächlich zurückgelegten Weges, wenn man sich beim Marsch vom Anfangs- zum Endpunkt nur parallel zu den Koordinatenachsen bewegt. Idealisiert wäre das die Weglänge bei rechtwinklig verlaufenden Straßenzügen.

(b) Wir betrachten nacheinander die drei Eigenschaften einer Metrik.

1. Es muß $d(\vec{a}, \vec{b}) \in \mathbb{R}$ gelten, mit $d(\vec{a}, \vec{b}) \geq 0$ und $d(\vec{a}, \vec{b}) = 0 \Leftrightarrow \vec{a} = \vec{b}$.
   Da $d(\vec{a}, \vec{b})$ die Summe der positiven waagrechten Entfernung und der positiven senkrechten Entfernung ist, gelten die Bedingungen.

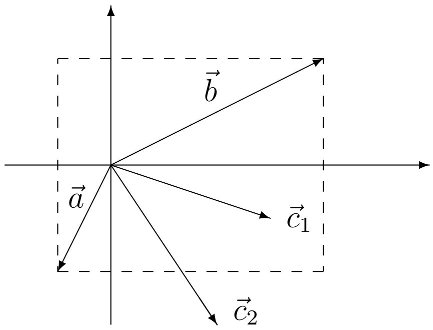


2. Es muß $d(\vec{a}, \vec{b}) = d(\vec{b}, \vec{a})$ gelten.

   Das ist unmittelbar klar, da der Weg vom Anfangs- zum Endpunkt entlang der skizzierten Verbindung genauso lang ist wie der Weg in der umgekehrten Richtung.

3. Es muß $d(\vec{a}, \vec{b}) \leq d(\vec{a}, \vec{c}) + d(\vec{c}, \vec{b})$ gelten.

   Dies kann man sich anschaulich an einer Skizze klar machen. Es gilt „=“, falls $\vec{c}$ im Rechteck oder auf dem Rand liegt. Es gilt „<“, falls $\vec{c}$ außerhalb des Rechtecks liegt.



(c) Es sei $\vec{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ und $\vec{b} = (b_1, \ldots, b_n) \in \mathbb{R}^n$. Wir zeigen, daß $d(\vec{a}, \vec{b}) = \sum_{i=1}^{n} |a_i - b_i|$ die drei Eigenschaften einer Metrik erfüllt.

1. Daß immer $d(\vec{a}, \vec{b}) \in \mathbb{R}$ ist und $d(\vec{a}, \vec{b}) \geq 0$ gilt ist klar, da Beträge summiert werden. Ferner gilt

$$\begin{aligned} d(\vec{a}, \vec{b}) = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{n} |a_i - b_i| = 0 \\ &\Leftrightarrow \quad |a_i - b_i| = 0 \quad \text{für alle } i \\ &\Leftrightarrow \quad a_i = b_i \quad \text{für alle } i \\ &\Leftrightarrow \quad \vec{a} = \vec{b}. \end{aligned}$$

2. Es ist $d(\vec{a}, \vec{b}) = d(\vec{b}, \vec{a})$, da für alle i gilt: $|a_i - b_i| = |b_i - a_i|$.

3. Mit der Dreiecksungleichung folgt $|a_i - c_i + c_i - b_i| \leq |a_i - c_i| + |c_i - b_i|$. Damit können wir die dritte Eigenschaft einer Metrik nachweisen:

$$\begin{aligned}
d(\vec{a}, \vec{b}) &= \sum_{i=1}^{n} |a_i - b_i| \\
&= \sum_{i=1}^{n} |a_i - c_i + c_i - b_i| \\
&\leq \sum_{i=1}^{n} (|a_i - c_i| + |c_i - b_i|) \\
&= \sum_{i=1}^{n} |a_i - c_i| + \sum_{i=1}^{n} |c_i - b_i| \\
&= d(\vec{a}, \vec{c}) + d(\vec{c}, \vec{b}).
\end{aligned}$$

Aufgabe 2

Mit Hilfe der Vektorrechnung kann man versuchen, die inhaltliche Ähnlichkeit von Dokumenten zu erfassen. Als Beispiel betrachten wir drei Texte, die aus Fabeln von Aesop stammen. Die Tabelle enthält alle Substantive, die in den drei Texten vorkommen, sowie die Häufigkeiten mit denen die Substantive in den einzelnen Texten auftreten. Insgesamt haben wir 26 Substantive.

Die drei Texte sollen nun durch 26-dimensionale Vektoren $\vec{v}_1$, $\vec{v}_2$ und $\vec{v}_3$ beschrieben werden, in denen die Häufigkeiten der Substantive stehen. Als Spaltenvektoren dargestellt hat man nichts anderes als die entsprechenden Spalten der Tabelle.

(a) Der Abstand $d(\vec{a}, \vec{b})$ zweier n-dimensionaler Vektoren wird bei der euklidischen Metrik durch

$$d(\vec{a}, \vec{b}) = |\vec{a} - \vec{b}|$$

definiert.

Berechnen Sie mit der euklidischen Metrik die Entfernungen $d(\vec{v}_1, \vec{v}_2)$ und $d(\vec{v}_1, \vec{v}_3)$ sowie $d(\vec{v}_2, \vec{v}_3)$. Welche beiden Texte sind am „ähnlichsten“ im Sinne dieser Metrik?

(b) Berechnen Sie den Winkel zwischen den Vektoren $\vec{v}_1$ und $\vec{v}_2$, den Winkel zwischen $\vec{v}_1$ und $\vec{v}_3$, und schließlich den Winkel zwischen $\vec{v}_2$ und $\vec{v}_3$. Wie verhält es sich mit der „Ähnlichkeit“ der drei Texte, wenn man den Winkel zwischen den Vektoren als Maß nimmt, also von größerer Ähnlichkeit bei kleinerem Winkel spricht?

(c) Um die Entfernung eines Vektors $\vec{a} = (a_1, \ldots, a_n)$ von einem Vektor $\vec{b} = (b_1, \ldots, b_n)$ festzulegen, muß man nicht zwingend die euklidischen Metrik

Substantive	Text 1	Text 2	Text 3
Abend			1
Bauer	1		1
Besitzer		1	
Bitte	1		
Esel	3	3	3
Fell		1	
Freund			1
Futter			1
Hartherzigkeit		1	
Haut		1	
Herr		1	
Herz			1
Kost			1
Kraft	1		
Last	1	3	2
Markt	1		
Morgen			1
Mühe		1	
Pferd	3	2	
Rat			1
Schläge			1
Strecke	1		
Tag			3
Teil	2	1	
Tod		1	
Ziege			2

verwenden. Mit der Definition

$$d(\vec{a}, \vec{b}) = |a_1 - b_1| + \ldots + |a_n - b_n|$$

erhält man eine andere Metrik, d.h. einen anderen Abstandsbegriff. Berechnen Sie mit dieser Metrik die Entfernungen $d(\vec{v}_1, \vec{v}_2)$, $d(\vec{v}_1, \vec{v}_3)$ sowie $d(\vec{v}_2, \vec{v}_3)$. Was folgt für die „Ähnlichkeit“ der Texte im Sinne dieser Metrik?

Die verwendeten Texte stammen aus Fabeln von Aesop[1]. Text 1 besteht aus den drei ersten Absätzen der Fabel „Das Pferd und der Esel“. Text 2 besteht aus dem vierten und fünften Absatz der gleichen Fabel. Text 3 enthält die ersten drei Absätze der Fabel „Der Esel und die Ziege“.

[1]Quelle: Projekt Gutenberg-DE, http://gutenberg.spiegel.de/

Text 1:

> Ein Bauer trieb ein Pferd und einen Esel, beide gleichmäßig beladen, zu Markte. Als sie schon eine gute Strecke vorwärts gegangen waren, fühlte der Esel seine Kräfte abnehmen. „Ach“, bat er das Pferd kläglich: „Du bist viel größer und stärker als ich, und doch hast du nicht schwerer zu tragen, nimm mir einen Teil meiner Last ab, sonst erliege ich.“
>
> Hartherzig schlug ihm das Pferd seine Bitte ab: „Ich habe selbst meinen Teil, und daran genug zu tragen.“
>
> Keuchend schleppte sich der Esel weiter, bis er endlich erschöpft zusammenstürzte.

Text 2:

> Vergeblich hieb der Herr auf ihn ein, er war tot. Es blieb nun nichts weiter übrig, als die ganze Last des Esels dem Pferde aufzupacken, und um doch etwas von dem Esel zu retten, zog ihm der Besitzer das Fell ab und legte auch dieses noch dem Pferde oben auf.
>
> Zu spät bereute dieses seine Hartherzigkeit. „Mit leichter Mühe“, so klagte es, „hätte ich dem Esel einen kleinen Teil seiner Last abnehmen und ihn vom Tode retten können. Jetzt muß ich seine ganze Last und dazu noch seine Haut tragen.“

Text 3:

> Ein Bauer hatte einen Esel und eine Ziege. Weil nun der Esel sehr viel arbeiten und große Lasten tragen mußte, erhielt er ein reichlicheres und besseres Futter als die Ziege.
>
> Diese beneidete den Esel, und um ihn um die bessere Kost zu bringen, oder doch wenigstens ihm Schläge einzutragen, sprach sie eines Tages zu ihm:
>
> „Höre, lieber Freund! Oft schon habe ich dich von Herzen bedauert, daß du Tag für Tag die schwersten Lasten tragen und vom Morgen bis Abend arbeiten mußt; ich möchte dir wohl einen guten Rat geben.“

Lösung: Bei dieser Aufgabe sind die Berechnungen nicht schwierig, erfordern aber etwas Platz, da die Ausgangsvektoren 26 Koordinaten haben. Deshalb sind die Teilrechnungen in einer Tabelle zusammengefaßt.

Die Spalten 1. bis 3. enthalten die Vektoren $\vec{v}_1$, $\vec{v}_2$ und $\vec{v}_3$:

1. der Vektor $\vec{v}_1 = (v_{1,1}, v_{1,2}, \ldots, v_{1,i}, \ldots, v_{1,26})$ als Spaltenvektor geschrieben,
2. der Vektor $\vec{v}_2 = (v_{2,1}, v_{2,2}, \ldots, v_{2,i}, \ldots, v_{2,26})$ als Spaltenvektor geschrieben,

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
Abend	0	0	1	0	1	1	0	0	0	0	1	1
Bauer	1	0	1	1	0	1	0	1	0	1	0	1
Besitzer	0	1	0	1	0	1	0	0	0	1	0	1
Bitte	1	0	0	1	1	0	0	0	0	1	1	0
Esel	3	3	3	0	0	0	9	9	9	0	0	0
Fell	0	1	0	1	0	1	0	0	0	1	0	1
Freund	0	0	1	0	1	1	0	0	0	0	1	1
Futter	0	0	1	0	1	1	0	0	0	0	1	1
Hartherzigkeit	0	1	0	1	0	1	0	0	0	1	0	1
Haut	0	1	0	1	0	1	0	0	0	1	0	1
Herr	0	1	0	1	0	1	0	0	0	1	0	1
Herz	0	0	1	0	1	1	0	0	0	0	1	1
Kost	0	0	1	0	1	1	0	0	0	0	1	1
Kraft	1	0	0	1	1	0	0	0	0	1	1	0
Last	1	3	2	4	1	1	3	2	6	2	1	1
Markt	1	0	0	1	1	0	0	0	0	1	1	0
Morgen	0	0	1	0	1	1	0	0	0	0	1	1
Mühe	0	1	0	1	0	1	0	0	0	1	0	1
Pferd	3	2	0	1	9	4	6	0	0	1	3	2
Rat	0	0	1	0	1	1	0	0	0	0	1	1
Schläge	0	0	1	0	1	1	0	0	0	0	1	1
Strecke	1	0	0	1	1	0	0	0	0	1	1	0
Tag	0	0	3	0	9	9	0	0	0	0	3	3
Teil	2	1	0	1	4	1	2	0	0	1	2	1
Tod	0	1	0	1	0	1	0	0	0	1	0	1
Ziege	0	0	2	0	4	4	0	0	0	0	2	2
Spaltensumme				18	39	35	20	12	15	16	23	25

3. der Vektor $\vec{v}_3 = (v_{3,1}, v_{3,2}, \ldots, v_{3,i}, \ldots, v_{3,26})$ als Spaltenvektor geschrieben.

Jede Vektorkoordinate wird durch zwei Indizes charakterisiert, der erste Index gibt die Nummer des Vektors an, der zweite Index die Position innerhalb des Vektors.

Die Spalten 4. bis 12. werden im folgenden in den Teilen der Aufgabe erläutert, in denen sie gebraucht werden.

(a) Wir verwenden die euklidische Metrik. Es soll der Wert

$$d(\vec{v}_1, \vec{v}_2) = |\vec{v}_1 - \vec{v}_2|$$

berechnet werden. Der Betrag eines Vektors ist die Wurzel aus der Summe der quadrierten Koordinaten

$$|\vec{v}_1 - \vec{v}_2| = \sqrt{(v_{1,1} - v_{2,1})^2 + (v_{1,2} - v_{2,2})^2 + \ldots + (v_{1,26} - v_{2,26})^2}.$$

Die quadrierten Koordinaten des Vektors $\vec{v}_1 - \vec{v}_2$ stehen in der 4. Spalte der Tabelle: Differenz aus Spalte 1 und Spalte 2 zum Quadrat, $(v_{1,i} - v_{2,i})^2$.

Die Spaltensumme der 4. Spalte ist also der Wert von

$$(v_{1,1} - v_{2,1})^2 + (v_{1,2} - v_{2,2})^2 + \ldots + (v_{1,26} - v_{2,26})^2,$$

und es folgt

$$\begin{aligned} d(\vec{v}_1, \vec{v}_2) &= |\vec{v}_1 - \vec{v}_2| \\ &= \sqrt{(v_{1,1} - v_{2,1})^2 + (v_{1,2} - v_{2,2})^2 + \ldots + (v_{1,26} - v_{2,26})^2} \\ &= \sqrt{18} \\ &\approx 4,24. \end{aligned}$$

Die 5. Spalte der Lösungstabelle enthält die Differenz aus Spalte 1 und Spalte 3 zum Quadrat, $(v_{1,i} - v_{3,i})^2$. Das sind die quadrierten Koordinaten des Vektors $\vec{v}_1 - \vec{v}_3$. Mit der Spaltensumme der 5. Spalte folgt

$$\begin{aligned} d(\vec{v}_1, \vec{v}_3) &= |\vec{v}_1 - \vec{v}_3| \\ &= \sqrt{(v_{1,1} - v_{3,1})^2 + (v_{1,2} - v_{3,2})^2 + \ldots + (v_{1,26} - v_{3,26})^2} \\ &= \sqrt{39} \\ &\approx 6,24. \end{aligned}$$

Die Werte der 6. Spalte sind die Differenzen aus Spalte 2 und Spalte 3 zum Quadrat genommen, $(v_{2,i} - v_{3,i})^2$. Entsprechend liefert die Spaltensumme der 6. Spalte

$$\begin{aligned} d(\vec{v}_2, \vec{v}_3) &= |\vec{v}_2 - \vec{v}_3| \\ &= \sqrt{(v_{2,1} - v_{3,1})^2 + (v_{2,2} - v_{3,2})^2 + \ldots + (v_{2,26} - v_{3,26})^2} \\ &= \sqrt{35} \\ &\approx 5,92. \end{aligned}$$

Fazit: Am „ähnlichsten“ im Sinne der euklidischen Metrik sind die Texte 1 und 2. Die „Entfernung“ von Text 1 zu Text 3 ist ungefähr gleich der „Entfernung“ von Text 2 zu Text 3 und deutlich größer als die „Entfernung“ von Text 1 zu Text 2.

(b) Der Winkel $\varphi_{1,2}$ zwischen den beiden 26-dimensionalen Vektoren $\vec{v}_1$ und $\vec{v}_2$ ist über das Skalarprodukt $\vec{v}_1 \cdot \vec{v}_2$ als

$$\varphi_{1,2} = \arccos\left(\frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1| \cdot |\vec{v}_2|}\right)$$

definiert.

Zur Berechnung von $\vec{v}_1 \cdot \vec{v}_2$ dient Spalte 7. der Lösungstabelle; hier stehen die Produkte aus der 1. und 2. Spalte. Damit ist die Spaltensumme gleich dem Skalarprodukt, also

$$\vec{v}_1 \cdot \vec{v}_2 = 20.$$

Quadriert man die Koordinaten von $\vec{v}_1$ und zählt die Quadrate zusammen, so erhält man die Zahl 28. Für die Quadratsumme der Koordinaten von $\vec{v}_2$ ergibt sich 30. Damit folgt

$$\begin{aligned}\varphi_{1,2} &= \arccos\left(\frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1| \cdot |\vec{v}_2|}\right) \\ &= \arccos\left(\frac{20}{\sqrt{28} \cdot \sqrt{30}}\right) \\ &\approx 0,81.\end{aligned}$$

In der 8. Spalte stehen die Produkte aus der 1. und der 3. Spalte, so daß die Spaltensumme das Skalarprodukt von $\vec{v}_1$ und $\vec{v}_3$ ist,

$$\vec{v}_1 \cdot \vec{v}_3 = 12.$$

Die 9. Spalte enthält die Produkte aus der 2. und der 3. Spalte, und mit der Spaltensumme folgt

$$\vec{v}_2 \cdot \vec{v}_3 = 15.$$

Für die Quadratsumme der Koordinaten von $\vec{v}_3$ erhält man 35.

Damit bekommt man für den Winkel $\varphi_{1,3}$ zwischen den Vektoren $\vec{v}_1$ und $\vec{v}_3$ den Wert

$$\begin{aligned}\varphi_{1,3} &= \arccos\left(\frac{\vec{v}_1 \cdot \vec{v}_3}{|\vec{v}_1| \cdot |\vec{v}_3|}\right) \\ &= \arccos\left(\frac{12}{\sqrt{28} \cdot \sqrt{35}}\right) \\ &\approx 1,18\end{aligned}$$

und für den Winkel $\varphi_{2,3}$ zwischen $\vec{v}_2$ und $\vec{v}_3$ den Wert

$$\begin{aligned}\varphi_{2,3} &= \arccos\left(\frac{\vec{v}_2 \cdot \vec{v}_3}{|\vec{v}_2| \cdot |\vec{v}_3|}\right) \\ &= \arccos\left(\frac{15}{\sqrt{30} \cdot \sqrt{35}}\right) \\ &\approx 1,09.\end{aligned}$$

Die Winkel sind in Bogenmaß angegeben.

Fazit: Die Ähnlichkeiten der Texte verhalten sich wie bei Teil (a) der Aufgabe.

(c) Es soll die Metrik

$$d(\vec{v}_1, \vec{v}_2) = |v_{1,1} - v_{2,1}| + \ldots + |v_{1,26} - v_{2,26}|$$

verwendet werden. In der 10. Spalte der Lösungstabelle stehen die Beträge der Differenzen der 1. und 2. Spalte. Die Spaltensumme der 10. Spalte ist also gleich dem gesuchten Wert

$$d(\vec{v}_1, \vec{v}_2) = |v_{1,1} - v_{2,1}| + \ldots + |v_{1,26} - v_{2,26}| = 16.$$

Die 11. Spalte liefert entsprechend

$$d(\vec{v}_1, \vec{v}_3) = |v_{1,1} - v_{3,1}| + \ldots + |v_{1,26} - v_{3,26}| = 23,$$

und die 12. Spalte gibt

$$d(\vec{v}_2, \vec{v}_3) = |v_{2,1} - v_{3,1}| + \ldots + |v_{2,26} - v_{3,26}| = 25.$$

Fazit: Auch im Sinne dieser Metrik sind die Texte 1 und 2 am „ähnlichsten“. Ebenso ist wie in (a) und (b) die „Entfernung“ von Text 1 zu Text 3 ungefähr gleich der „Entfernung“ von Text 2 zu Text 3 und deutlich größer als die „Entfernung“ von Text 1 zu Text 2.

Es gibt jedoch auch einen kleinen Unterschied zu (a) und (b). Sowohl bei (a) als auch bei (b) sind Text 1 und Text 3 am weitesten voneinander „entfernt“. Bei (c) jedoch liegen Text 2 und Text 3 „am weitesten auseinander“.

Wir sehen also, daß die Wahl des Abstandsmaßes einen Einfluß auf das Ergebnis haben kann. Welches Maß für die konkrete Aufgabenstellung am geeignetsten ist, muß eventuell mit Hilfe von Experimenten herausgefunden werden, falls es überhaupt eine eindeutige Antwort gibt.