technische universität dortmund

fakultät für mathematik

Prof. Dr. Ivan Veselić

Dr. Michele Serra

Synopse zur 10. Vorlesung – 13.09.2024

Vorkurs Mathematik

im Wintersemester 2024

Reflektion

Gibt es überhaupt vernünftige Alternativen zur Euklidischen Norm und zum Euklidischen Abstand?
In der Tat können Sie bei Google Maps, bei der Berechnung einer Route die

- schnellste Route,
- kürzeste Route,
- ökologischste Route,
- und bei manchen anderen Routenplanern andere Kriterien,

auswählen. Bei jeder Option wird die "optimale" Route berechnet, was aber optimiert wird, ist jedesmal ein anderer Abstandsbegriff oder eine andere Kostenfunktion. Ich kann z.B. die Fahrzeit minimieren wollen oder das verbrauchte Benzin.
Also gibt es auch im alltäglichen Lebensituationen, wo ich zwischen verschiedenen Begriffen von "Abstand" bei einer Optimierungsaufgabe wählen kann.
Ganz ähnlich gibt es auch im Vektorraum $\mathbb{R}^3$ oder $\mathbb{R}^n$ unterschiedliche Abstandsbegriffe.

Beispiel (ℓ_1-Norm)
Wir setzen für $\vec{v} \in \mathbb{R}^n$

$$\|\vec{v}\|_1 = N_1(\vec{v}) := |v_1| + |v_2| + ... + |v_n| = \sum_{j=1}^{n} |v_j| \,.$$

Dann gilt offensichtlich:

$$N_1(\vec{v}) \geq 0 \text{ für jedes } \vec{v} \in \mathbb{R}^n.$$

Ist $\lambda \in \mathbb{R}$ beliebig, gilt auch

$$\begin{aligned} N_1(\lambda \cdot \vec{v}) &= \sum_{j=1}^{n} \underbrace{|\lambda \cdot v_j|}_{=|\lambda| \cdot |v_j|} \\ &= \sum_{j=1}^{n} |\lambda| \cdot |v_j| = |\lambda| \sum_{j=1}^{n} |v_j| = |\lambda| \cdot N_1(\vec{v}) \end{aligned}$$

Für ein beliebiges Paar $\vec{v}, \vec{w} \in \mathbb{R}$ gilt

$$N_1(\vec{v}+\vec{w}) = \sum_{i=1}^{n} \underbrace{|v_i + w_i|}_{\leq |v_i|+|w_i|} \leq \sum_{i=1}^{n} (|v_i| + |w_i|) = \sum_{i=1}^{n} |v_i| + \sum_{i=1}^{n} |w_i| = N_1(\vec{v}) + N_1(\vec{w})$$

Insgesamt haben wir nachgewiesen, dass die Abbildung N_1 alle Anforderungen erfüllt, die für eine Norm gelten müssen. Die Notation $\|\cdot\|_1$ ist üblicher als $N_1(.)$, deshalb werde ich sie ab jetzt benutzen.
Nun werden wir zeigen, dass

$$d_1(\vec{v}, \vec{w}) = \|\vec{v} - \vec{w}\|_1$$

alle Eigenschaften einer Metrik erfüllt.

$$\text{“nicht negativ”}: \quad d_1(\vec{v}, \vec{w}) \geqslant 0$$

da $\|\vec{u}\|_1 \geqslant 0$ für alle $\vec{u} \in \mathbb{R}^n$ also auch für $\vec{u} = \vec{v} - \vec{w}$

$$\begin{aligned} \text{“symmetrisch”}: \quad d_1(\vec{v}, \vec{w}) &= \|\vec{v} - \vec{w}\|_1 \\ &= \|\underbrace{(-1)}_{=\lambda}(\vec{w} - \vec{v})\|_1 = |\lambda| \ \|\vec{w} - \vec{v}\|_1 \\ \lambda = -1 \Rightarrow |\lambda| = 1 & \\ &= 1 \cdot d_1(\vec{w}, \vec{v}) \end{aligned}$$

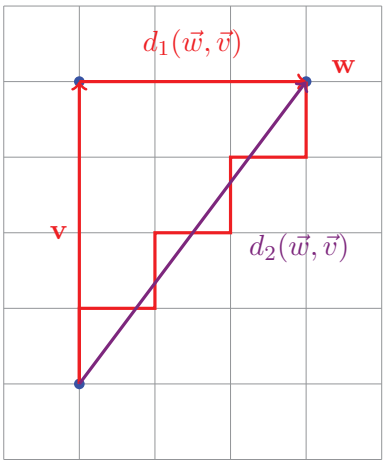
“Dreiecksungleichung”

$$\begin{aligned} d_1(\vec{v}, \vec{w}) = \|\vec{v} - \vec{w}\|_1 &= \|\vec{v} \ \underbrace{-\vec{u} + \vec{u}}_{=0 \ \text{addiert}} \ -\vec{w}\|_1 \\ &\leq \|\vec{v} - \vec{u}\|_1 + \|\vec{u} - \vec{w}\|_1 = d_1(\vec{v}, \vec{u}) + d_1(\vec{u}, \vec{w}) \end{aligned}$$

Reflektion

Wir haben in der vorletzten Zeile die Dreiecksungleichung für die Norm $\|.\|_1$ verwendet. Diese Eigenschaft haben wir ja kürzlich nachgerechnet und überprüft. Daher ist es legitim, sie zu verwenden, wenn wir etwas Neues nachrechnen wollen.
Der Trick mit der Addition der Null ist sehr häufig nützlich.

Macht dieser neue Abstand $d_1(.,.)$ überhaupt Sinn? “Manhattan Metrik”



Wir werden sehen, das der “Median” aus der Stochastik mit diesem Abstand in Verbindung steht.

Median

Definition

Sei X eine zufällige reellwertige Größe, d.h. eine "zufällige reelle Zahl" (Zufallsvariable, Zufallsgröße). Eine Zahl $m \in \mathbb{R}$ heißt Median der Zufallsgröße X, falls die Wahrscheinlichkeiten der beiden Ereignisse

X ist größergleich m und

X ist keinergleich m

jeweils mindestens $\frac{1}{2}$ sind.

Beispiel Sei der Datensatz $1, 2, 2, 3, 4, 7, 9$ gegeben. Ich kann ihn modellieren durch eine reellwertige Zufallsgröße, die die Werte $1, 3, 4, 7$ und 9 jeweils mit Wahrscheinlichkeit $\frac{1}{7}$ annimmt und den Wert 2 mit Wahrscheinlichkeit $\frac{2}{7}$. Das arithmetische Mittel des Datensatzes ist

$$c_0 = \frac{1}{7}\sum_{j=1}^{7} x_j = \frac{1+2+2+3+4+7+9}{7} = \frac{28}{7} = 4$$

Ist der Wert 4 auch ein Median von X bzw. dem Datensatz?
Um die Frage zu beantworten, überprüfen wir die beiden geforderten Bedingungen:

$$P(X \geq 4) := \text{Wahrscheinlichkeit, dass } X \geq 4 \text{ ist} = \frac{1}{7} + \frac{1}{7} + \frac{1}{7} = \frac{3}{7} < \frac{1}{2} \tag{8}$$

entsprechend den Daten $4, 7, 9$. Also ist die erste Bedingung verletzt und der Wert 4 kein Median.

Behauptung: 3 ist ein Median für diesen Datensatz, bzw. zu X.
Beweis. $P(X \geq 3) =$ Wahrscheinlichkeit, dass X einen Wert ≥ 3 annimmt

$$= \frac{1}{7} + \frac{1}{7} + \frac{1}{7} + \frac{1}{7} = \frac{4}{7} \geq \frac{1}{2}$$

$P(X \leq 3) =$ Wahrscheinlichkeit, dass X einen Wert ≤ 3 annimmt

$$= \frac{1}{7} + \frac{1}{7} + \frac{1}{7} + \frac{1}{7} = \frac{4}{7} \geq \frac{1}{2}$$

Beispiel

Betrachte nun die ZV Y, die die Werte $1, 3, 4, 7, 9, 52$ mit Wahrscheinlichkeit $1/8$ annimmt und den Wert 2 mit Wahrsch. $2/8$. Entspricht den Datensatz

$$1, 2, 2, 3, 4, 7, 9, 52$$

Arithmetisches Mittel ist: $c_0 = \frac{1}{8}\sum_{j=1}^{8} y_j = \frac{1+2+2+3+4+7+9+52}{8} = 10$

Frage: was ist das Median? Ist es wieder $m = 3$?
Test: Die erste Bedingung: $P(Y \geq 3) = 5/8 \geq 1/2$ ✓
Zweite Bedingung: $P(Y \leq 3) = 4/8 \geq 1/2$ ✓
Was ist der Grund dafür, dass der Median nicht mit dem Durchschnitt zusammenfällt? In beiden Datensätzen gibt es Ausreißer am oberen Ende der Daten. Diese "verschieben" den Durchschnitt stärker noch oben, als den Median.

Beispiel

Konkrete Daten von Jahresgehälter, bei denen das Phänomen auftritt, dass Median $\neq$ Durchschnitt

Land	Jahr	Währung	Arithm. Mittel	Median	Kommentar
DEU	2024	EUR	50250	43750	Datenbank von Stellenausschreibung
USA	2023	USD	63795	59384	2023/24
UAE	2008	AED	90484	39000	4 AED$\simeq$ 1 EUR

Wir erinnern uns an die Lösung der Aufgabe 7: zu gegebenen Daten $(x_1, y_1), \ldots, (y_N, y_N)$ im $\mathbb{R}^2$ ist

$$g(x) = \frac{1}{N} \sum_{j=1}^{N} y_j,$$

die waagerechte Gerade, das "beste c" bei der der quadr. Fehler

$$Q(c) = \sum_{j=1}^{N} (c - y_j)^2 = d_2 \left(\begin{pmatrix} c \\ \vdots \\ c \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ c_N \end{pmatrix} \right)^2$$

am geringsten ist.
Dieser beste Wert $c_0 = \frac{1}{N} \sum_{j=1}^{N} y_j$ ist das arithmetische Mittel der $y_1, \ldots, y_N$.
Falls wir dagegen den Fehler in der Manhattan-Metrik d_1 minimieren, d.h. die Funktion

$$E \colon \mathbb{R} \to \mathbb{R}, \; E(c) = \sum_{j=1}^{N} |c - y| = d_1 \left(\begin{pmatrix} c_1 \\ \vdots \\ c \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ c \end{pmatrix} \right)$$

minimieren, dann erfüllt jeder Median m des Datensatzes $y_1, \ldots, y_N$ die Optimalitätsbedingung $E(m) = \min_{c \in \mathbb{R}} E(c)$.
Anders ausgedrückt: Für jeden Median m (des Datensatzes) und jede Zahl $c \in \mathbb{R}$ gilt

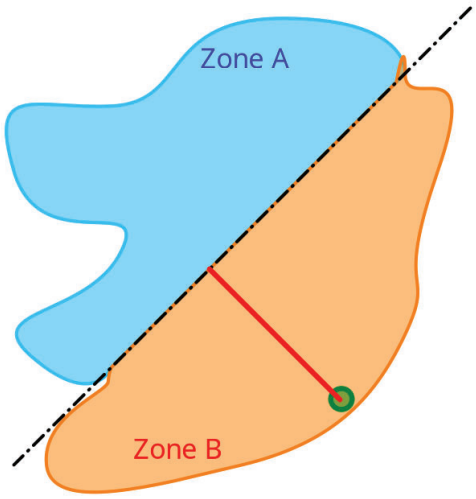
$$E(m) \leq E(c).$$

Reflektion

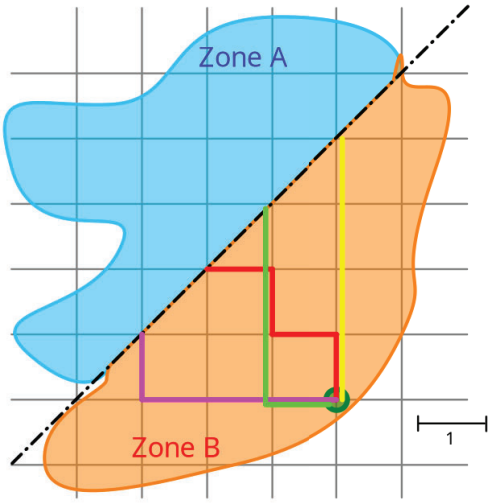
Haben auch etwas zu Frage 2 aus der Liste am Anfang des Vorkurses gelernt: Bei der "Fehlerbewertung" kann man tatsächlich statt der Potenz 2 auch die Potenz 1 verwenden (dann aber natürlich mit Beträge der einzelnen Differenzen). Dies liefert auch eine gewisse Bestapproximation aber im anderen Sinne und i.a. mit anderen Werten.

Frage: Anhand von ZV sieht man, dass der Median eines Datensatzes nicht eindeutig ist. Kann man das auch geometrich beleuchten?

Beispiel: Tarifzonen im VRR. Angenommen es gibt zwei Tarifzonen: Zone A, Zone B. Ich habe ein Ticket für Zone A und befinde mich in Zone B. Ich will Zone A zu Fuß erreichen, um dort den Bus zu nehmen. Wie wähle ich den besten Fussweg zur Zone A?

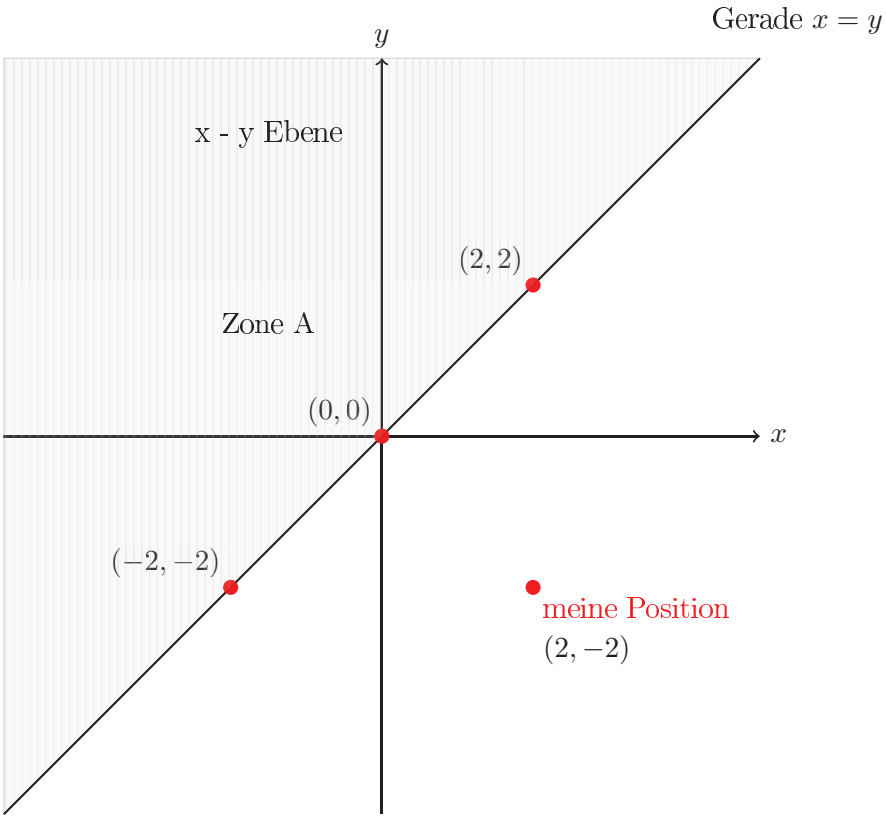


In der Euklidische Metrik ist die kürzste Strecke eindeutig. In der Realität ist sie aber wahrscheinlich keine Option!



In der Manhattan-Metrik gibt es mehrere minimale Routen der gleichen Länge: 4.

Hier Beispielrechnung mit Zahlenwerten



Euklidischer Abstand zwischen

$$(2,-2) \text{ und } (2,2) \text{ ist} \quad \sqrt{(2-2)^2+(-2-2)^2} = \sqrt{0+(-4)^2} = 4$$

$$(2,-2) \text{ und } (0,0) \text{ ist} \quad \sqrt{(2)^2+(-2)^2} = \sqrt{8} = 2\sqrt{2}$$

$$(2,-2) \text{ und } (-2,-2) \text{ ist} \quad \sqrt{(2-(-2))^2+(-2-(-2))^2} = \sqrt{(2+2)^2+0} = 4$$

Da

$$2\sqrt{2} < 2 \cdot 2 = 4$$

ist der *Euklidische* Abstand von $(2,-2)$ zu $(0,0)$ kürzer als die beiden anderen.
Wie sieht es mit dem d_1-Abstand aus?
Der d_1 -Abstand zwischen

$$\begin{aligned}
(2,-2) \text{ und } (0,0) \text{ ist} \quad & |2-2|+|-2-2| = 4 \\
(2,-2) \text{ und } (2,2) \text{ ist} \quad & |2-0|+|-2-0| = |2|+|2| = 4 \\
(2,-2) \text{ und } (-2,-2) \text{ ist} \quad & |2-(-2)|+|-2-(-2)| = |2+2|+|-2+2| = 4
\end{aligned}$$

Also ist der d_1-Abstand in allen drei Fällen gleich.