

# Data Mining Tutorial

## Hausaufgabe Distanzfunktionen

Erich Schubert, Arthur Zimek

Ludwig-Maximilians-Universität München

2014-04-25 — KDD Übung

- Reflexiv: "Distanz zu sich selbst ist 0"

$$x = y \quad \Rightarrow \quad d(x, y) = 0$$

- Symmetrisch: "Reihenfolge der Parameter ist egal"

$$d(x, y) = a \quad \Leftrightarrow \quad d(y, x) = a$$

- Strikt: "Nur identische Elemente haben Distanz 0"

$$d(x, y) = 0 \quad \Rightarrow \quad x = y$$

- Dreiecksungleichung:  
"Der direkte Weg ist nie länger als ein Umweg"

$$d(x, y) \leq d(x, z) + d(z, y)$$

## Ein Beispiel ist kein Beweis.

Ein Beispiel ist kein Beweis.

... aber ein Gegenbeispiel widerlegt!

Ein Beispiel ist kein Beweis.

... aber ein Gegenbeispiel widerlegt!

Wenn ihr etwas beweisen sollte, dann immer für *alle*  
Situationen, oder eben ein Gegenbeispiel.  
Ein Positivbeispiel bringt nichts!

$$d(x, y) = \sum_{i=1}^n (x_i - y_i)$$

$$d(x, y) = \sum_{i=1}^n (x_i - y_i)$$

$d((0), (1)) = -1$  – darf aber nicht negativ werden!  
Gegenbeispiel!

$$d(x, y) = \sum_{i=1}^n (x_i - y_i)^2$$

$$d(x, y) = \sum_{i=1}^n (x_i - y_i)^2$$

Reflexiv, symmetrisch und strikt: offensichtlich.  
Aber Dreiecksungleichung?

$$d(x, y) = \sum_{i=1}^n (x_i - y_i)^2$$

Reflexiv, symmetrisch und strikt: offensichtlich.  
Aber Dreiecksungleichung?

Wie sieht es aus mit:  $o = (0, 0)$ ,  $p = (1, 0)$ ,  $q = (2, 0)$ ?

$$d(o, q) = 4 \qquad d(o, p) + d(p, q) = 1 + 1 = 2$$

$$d(x, y) = \sum_{i=1}^n (x_i - y_i)^2$$

Reflexiv, symmetrisch und strikt: offensichtlich.  
Aber Dreiecksungleichung?

Wie sieht es aus mit:  $o = (0, 0)$ ,  $p = (1, 0)$ ,  $q = (2, 0)$ ?

$$d(o, q) = 4 \quad \not\leq \quad d(o, p) + d(p, q) = 1 + 1 = 2$$

“Quadrierte Euklidische Distanz” ist nicht metrisch!  
(1-dimensionales Gegenbeispiel: 0, 1, 2)

$$d(x, y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2}$$

$$d(x, y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2}$$

Reflexiv und symmetrisch: offensichtlich.  
Dreiecksungleichung per Cauchy-Schwarzscher  
Ungleichung.

$$d(x, y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2}$$

Reflexiv und symmetrisch: offensichtlich.  
Dreiecksungleichung per Cauchy-Schwarzscher  
Ungleichung.

Aber nicht strikt: die Dimension  $n$  wird ignoriert!

$$d(x, y) = \sum_{i=1}^n \begin{cases} 1 & \text{iff } x_i = y_i \\ 0 & \text{iff } x_i \neq y_i \end{cases}$$

$$d(x, y) = \sum_{i=1}^n \begin{cases} 1 & \text{iff } x_i = y_i \\ 0 & \text{iff } x_i \neq y_i \end{cases}$$

$d$  ist nicht reflexiv. Damit für uns keine Distanz oder Metrik.

$$d(x, y) = \sum_{i=1}^n \begin{cases} 1 & \text{iff } x_i \neq y_i \\ 0 & \text{iff } x_i = y_i \end{cases}$$

$$d(x, y) = \sum_{i=1}^n \begin{cases} 1 & \text{iff } x_i \neq y_i \\ 0 & \text{iff } x_i = y_i \end{cases}$$

Diskordanz auf binären Vektoren.

“Anzahl der gesetzten Bits nach einer XOR-Verknüpfung der beiden Vektoren”.

Wichtige Metrik aus der Informationstheorie.

Reflexivität, Striktheit, Symmetrie sind offensichtlich.

Beweis der Dreiecksungleichung durch Fallunterscheidung auf den einzelnen Stellen (Dimensionen):

$$\begin{aligned}
 d(x, y) + d(y, z) &= \sum_i^n d(x_i, y_i) + \sum_i^n d(y_i, z_i) \\
 &= \sum_i^n (d(x_i, y_i) + d(y_i, z_i)) \\
 ?? \quad \sum_i^n d(x_i, z_i) &= d(x, z)
 \end{aligned}$$

Kernidee: wenn  $\leq$  für jeden Summanden gilt, gilt es auch insgesamt! Also Fallunterscheidung.

Beweis der Dreiecksungleichung durch Fallunterscheidung  
auf den einzelnen Stellen (Dimensionen):

A)  $x_i = y_i \wedge y_i = z_i$ :

$$d(x_i, y_i) + d(y_i, z_i) \geq d(x_i, z_i)$$

$$d(x_i, x_i) + d(y_i, x_i) \geq d(x_i, x_i)$$

$$0 + 0 \geq 0$$

Beweis der Dreiecksungleichung durch Fallunterscheidung auf den einzelnen Stellen (Dimensionen):

B)  $x_i = y_i \wedge x_i \neq z_i$ :

$$d(x_i, y_i) + d(y_i, z_i) \geq d(x_i, z_i)$$

$$d(x_i, x_i) + d(x_i, z_i) \geq d(x_i, z_i)$$

$$0 + 1 \geq 1$$

Beweis der Dreiecksungleichung durch Fallunterscheidung auf den einzelnen Stellen (Dimensionen):

C)  $x_i = z_i \wedge x_i \neq y_i$ :

$$d(x_i, y_i) + d(y_i, z_i) \geq d(x_i, z_i)$$

$$d(x_i, y_i) + d(y_i, x_i) \geq d(x_i, x_i)$$

$$1 + 1 \geq 0$$

Beweis der Dreiecksungleichung durch Fallunterscheidung auf den einzelnen Stellen (Dimensionen):

D)  $x_i \neq y_i \wedge y_i = z_i$ :

$$d(x_i, y_i) + d(y_i, z_i) \geq d(x_i, z_i)$$

$$d(x_i, y_i) + d(y_i, y_i) \geq d(x_i, y_i)$$

$$1 + 0 \geq 1$$

Beweis der Dreiecksungleichung durch Fallunterscheidung auf den einzelnen Stellen (Dimensionen):

E)  $x_i \neq y_i \wedge y_i \neq z_i \wedge x_i \neq z_i$ :

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ 1 + 1 &\geq 1 \end{aligned}$$

Beweis der Dreiecksungleichung durch Fallunterscheidung  
auf den einzelnen Stellen (Dimensionen):  
Und damit gilt insgesamt:

$$\begin{aligned}d(x, y) + d(y, z) &= \sum_i^n d(x_i, y_i) + \sum_i^n d(y_i, z_i) \\&= \sum_i^n (d(x_i, y_i) + d(y_i, z_i)) \\&\geq \sum_i^n d(x_i, z_i) = d(x, z)\end{aligned}$$

Ein paar weitere Beispiele für Distanzfunktionen (für zwei Mengen  $X, Y \subseteq \mathbb{R}^n$ ), basierend auf einer bestehenden Distanzfunktion  $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_0^+$ :

- ▶  $\text{single-link}(X, Y) = \min_{x \in X, y \in Y} d(x, y)$
- ▶  $\text{average-link}(X, Y) = \frac{1}{|X| \cdot |Y|} \cdot \sum_{x \in X, y \in Y} d(x, y)$
- ▶  $\text{complete-link}(X, Y) = \max_{x \in X, y \in Y} d(x, y)$

Diese kommen im nächsten Kapitel, Clusteranalyse!

Es gibt hunderte an Distanzfunktionen.

- ▶ Für Zeitreihen: DTW, EDR, ERP, LCSS, ...
- ▶ Für Text: Cosine und Normalisierungen davon
- ▶ Für Mengen – basierend auf Schnitt, Vereinigung, ...
- ▶ Für Clusters (z.B. single-link)
- ▶ Für Histogramme: histogram intersection, “Earth movers distance”, quadratische Formen
- ▶ Mit Normalisierung: Canberra, ...
- ▶ Quadratische Formen / Bilinearformen:  $d(x, y) := x^T M y$  für positiv-definite (i.d.R. symmetrische) Matrix  $M$ .

Auch eine Art “Vorverarbeitung”:  
eine passende Distanzfunktion auswählen!

Gegeben eine Pseudo-Metrik  $d$  auf der Menge  $A$ :  
 $d : A \times A \rightarrow \mathbb{R}_0^+$ .

Definiere die Äquivalenzrelation  $\sim$  so dass  
 $x \sim y \Leftrightarrow d(x, y) = 0$ .

Sei  $A^\sim$  die Menge der Äquivalenzklassen von  $A$  bzgl.  $\sim$ .

$$d^\sim : A^\sim \times A^\sim \rightarrow \mathbb{R}_0^+ \\ \text{mit } d^\sim(x^\sim, y^\sim) := d(x, y)$$

Eigenschaften?

Gegeben eine Pseudo-Metrik  $d$  auf der Menge  $A$ :  
 $d : A \times A \rightarrow \mathbb{R}_0^+$ .

Definiere die Äquivalenzrelation  $\sim$  so dass  
 $x \sim y \Leftrightarrow d(x, y) = 0$ .

Sei  $A^\sim$  die Menge der Äquivalenzklassen von  $A$  bzgl.  $\sim$ .

$$d^\sim : A^\sim \times A^\sim \rightarrow \mathbb{R}_0^+ \\ \text{mit } d^\sim(x^\sim, y^\sim) := d(x, y)$$

Eigenschaften? Wohldefiniert?

Zu zeigen: für alle  $z \in x^\sim$ ,  $w \in y^\sim$  gilt  $d(z, w) = d(x, y)$ .

Zu zeigen: für alle  $z \in x^\sim$ ,  $w \in y^\sim$  gilt  $d(z, w) = d(x, y)$ .

Da  $z \in x^\sim$  und  $w \in y^\sim$  haben wir per Definition:

$$\begin{aligned} z^\sim &= x^\sim \text{ und } d(z, x) = 0 \\ w^\sim &= y^\sim \text{ und } d(w, y) = 0 \end{aligned}$$

Zu zeigen: für alle  $z \in x^\sim$ ,  $w \in y^\sim$  gilt  $d(z, w) = d(x, y)$ .

Da  $z \in x^\sim$  und  $w \in y^\sim$  haben wir per Definition:

$$\begin{aligned} z^\sim &= x^\sim \text{ und } d(z, x) = 0 \\ w^\sim &= y^\sim \text{ und } d(w, y) = 0 \end{aligned}$$

Durch  $1 \times$  anwenden Dreiecksungleichung erhalten wir:

$$\begin{aligned} d(z, w) &\leq d(z, x) + d(x, w) \\ d(x, y) &\leq d(x, z) + d(z, y) \end{aligned}$$

Zu zeigen: für alle  $z \in x^\sim$ ,  $w \in y^\sim$  gilt  $d(z, w) = d(x, y)$ .

Da  $z \in x^\sim$  und  $w \in y^\sim$  haben wir per Definition:

$$\begin{aligned} z^\sim &= x^\sim \text{ und } d(z, x) = 0 \\ w^\sim &= y^\sim \text{ und } d(w, y) = 0 \end{aligned}$$

Durch  $2 \times$  anwenden Dreiecksungleichung erhalten wir:

$$\begin{aligned} d(z, w) &\leq d(z, x) + d(x, y) + d(y, w) \\ d(x, y) &\leq d(x, z) + d(z, w) + d(w, y) \end{aligned}$$

Zu zeigen: für alle  $z \in x^\sim$ ,  $w \in y^\sim$  gilt  $d(z, w) = d(x, y)$ .

Da  $z \in x^\sim$  und  $w \in y^\sim$  haben wir per Definition:

$$\begin{aligned} z^\sim &= x^\sim \text{ und } d(z, x) = 0 \\ w^\sim &= y^\sim \text{ und } d(w, y) = 0 \end{aligned}$$

Durch  $2 \times$  anwenden Dreiecksungleichung erhalten wir:

$$\begin{aligned} d(z, w) &\leq d(z, x) + d(x, y) + d(y, w) = d(x, y) \\ d(x, y) &\leq d(x, z) + d(z, w) + d(w, y) = d(z, w) \end{aligned}$$

Zu zeigen: für alle  $z \in x^\sim$ ,  $w \in y^\sim$  gilt  $d(z, w) = d(x, y)$ .

Da  $z \in x^\sim$  und  $w \in y^\sim$  haben wir per Definition:

$$\begin{aligned} z^\sim &= x^\sim \text{ und } d(z, x) = 0 \\ w^\sim &= y^\sim \text{ und } d(w, y) = 0 \end{aligned}$$

Durch  $2 \times$  anwenden Dreiecksungleichung erhalten wir:

$$\begin{aligned} d(z, w) &\leq d(z, x) + d(x, y) + d(y, w) = d(x, y) \\ d(x, y) &\leq d(x, z) + d(z, w) + d(w, y) = d(z, w) \end{aligned}$$

$\Rightarrow$  Distanzberechnungen auf den Äquivalenzklassen wohldefiniert: egal welchen *Repräsentanten* wir wählen, es kommt das gleiche Ergebnis für  $d^\sim$  heraus.

Zu zeigen (Striktheit):  $d^{\sim}(a^{\sim}, b^{\sim}) = 0 \Rightarrow a^{\sim} = b^{\sim}$

Zu zeigen (Striktheit):  $d^{\sim}(a^{\sim}, b^{\sim}) = 0 \Rightarrow a^{\sim} = b^{\sim}$

$$d^{\sim}(a^{\sim}, b^{\sim}) = 0$$

$$\Rightarrow d(a, b) = 0$$

$$\Rightarrow a \sim b$$

$$\Rightarrow a^{\sim} = b^{\sim}$$

Zu zeigen (Striktheit):  $d^\sim(a^\sim, b^\sim) = 0 \Rightarrow a^\sim = b^\sim$

$$d^\sim(a^\sim, b^\sim) = 0$$

$$\Rightarrow d(a, b) = 0$$

$$\Rightarrow a \sim b$$

$$\Rightarrow a^\sim = b^\sim$$

Reflexivität, Symmetrie und Dreiecksungleichung folgen trivial aus den Eigenschaften von  $d$ !

$$\text{euclid}_{xy}((r_1, x_1, y_1), (r_2, x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

| Datensatz ID | $x$ | $y$ | Datensatz ID | $x$ | $y$ |
|--------------|-----|-----|--------------|-----|-----|
| 1            | 0   | 1   | 4            | 1   | 1   |
| 2            | 1   | 1   | 5            | 2   | 2   |
| 3            | 0   | 1   | 6            | 3   | 3   |

$$\text{euclid}_{xy}((r_1, x_1, y_1), (r_2, x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

| Datensatz ID | $x$ | $y$ | Datensatz ID | $x$ | $y$ |
|--------------|-----|-----|--------------|-----|-----|
| 1            | 0   | 1   | 4            | 1   | 1   |
| 2            | 1   | 1   | 5            | 2   | 2   |
| 3            | 0   | 1   | 6            | 3   | 3   |

Euklidische Distanz auf  $X \times Y$ . Metrik auf  $\mathbb{R}^2 \sim X \times Y$ ,  
aber nur eine Pseudo-metric auf  $\text{Datensatz ID} \times X \times Y$ .  
"Duplikate" haben eine Distanz von 0!