Data Mining Tutorial

Hausaufgabe Distanzfunktionen

Erich Schubert, Arthur Zimek

Ludwig-Maximilians-Universität München

2013-04-26 — KDD Übung

Distanzfunktionen

- 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"

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

Eine kleine Erinnerung ...

Ein Beispiel ist kein Beweis.

Eine kleine Erinnerung ...

Ein Beispiel ist kein Beweis.

... aber ein Gegenbeispiel wiederlegt!

Eine kleine Erinnerung ...

Ein Beispiel ist kein Beweis.

... aber ein Gegenbeispiel wiederlegt!

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

Hausaufgabe – Summe der Differenzen

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

Hausaufgabe – Summe der Differenzen

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

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

Hausaufgabe – Quadrierte Euklidische Distanz

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

Hausaufgabe – Quadrierte Euklidische Distanz

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

Reflexiv, symmetrisch und strikt: offensichtlich.
Aber Dreicksungleichung?

Hausaufgabe – Quadrierte Euklidische Distanz

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

Reflexiv, symmetrisch und strikt: offensichtlich.
Aber Dreicksungleichung?

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

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

Hausaufgabe – Quadrierte Euklidische Distanz

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

Reflexiv, symmetrisch und strikt: offensichtlich.
Aber Dreicksungleichung?

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: $0, 1, 2$)

Hausaufgabe – Projizierte Euklidische Distanz

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

Hausaufgabe – Projizierte Euklidische Distanz

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

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

Hausaufgabe – Projizierte Euklidische Distanz

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

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

Aber nicht strikt: die Dimension n wird ignoriert!

Hausaufgabe – Konkordanz

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

Hausaufgabe – Konkordanz

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

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

Hausaufgabe – Hamming-Distanz

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

Hausaufgabe – Hamming-Distanz

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

Diskordanz auf binären Vektoren.
“Anzahl der gesetzten Bits nach einer XOR-Verknüpfung der beiden Vektoren”.
Wichtige Metrik aus der Informationstheorie.
Reflexivität, Strikheit, Symmetrie sind offensichtlich.

Hausaufgabe – Hamming-Distanz

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

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

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

Hausaufgabe – Hamming-Distanz

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

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

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, x_i) + d(y_i, x_i) &\geq d(x_i, x_i) \\ 0 + 0 &\geq 0 \end{aligned}$$

Hausaufgabe – Hamming-Distanz

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

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

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, x_i) + d(x_i, z_i) &\geq d(x_i, z_i) \\ 0 + 1 &\geq 1 \end{aligned}$$

Hausaufgabe – Hamming-Distanz

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

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

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, y_i) + d(y_i, x_i) &\geq d(x_i, x_i) \\ 1 + 1 &\geq 0 \end{aligned}$$

Hausaufgabe – Hamming-Distanz

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

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

$$\begin{aligned} d(x_i, y_i) + d(y_i, z_i) &\geq d(x_i, z_i) \\ d(x_i, y_i) + d(y_i, y_i) &\geq d(x_i, y_i) \\ 1 + 0 &\geq 1 \end{aligned}$$

Hausaufgabe – Hamming-Distanz

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

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

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

Hausaufgabe – Hamming-Distanz

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

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

Distanzfunktionen – weitere Beispiele

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

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

Diese kommen im nächsten Kapitel, Clusteranalyse!

Distanzfunktionen – weitere Beispiele

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: $d(x, y) := x^T My$ für positiv-definite (i.d.R. symmetrische) Matrix M.

Auch eine Art "Vorverarbeitung":
eine passende Distanzfunktion auswählen!