6 Distanzfunktionen

1. Eigenschaften und Klassifikation

2. Distanzfunktionen auf Punkten
   - Minkowski-Distanzfunktion L_m
   - Gewichtete Minkowski-Distanzfunktion L_m^w
   - Quadratische Distanzfunktion d_q
   - Quadratische Pseudo-Distanzfunktion
   - Dynamical-Partial-Semi-Pseudo-Distanzfunktion
   - Chi-Quadrat-Semi-Pseudo-Distanzfunktion
   - Kullback-Leiber-Abstandsfunktion
   - Bahattacharyys-Abstandsfunktion

6 Distanzfunktionen (2)

3. Distanzfunktionen auf Binärdaten

4. Distanzfunktionen auf Sequenzen
   - Earth-Mover-Distanzfunktion
   - DFT-L_2-Distanzfunktion
   - Editierdistanz

5. Distanzfunktionen auf allgemeinen Mengen
   - Bottleneck-Distanzfunktion
   - Distanzfunktion über das Volumen der symmetrischen Differenz
   - Hausdorff-Distanzfunktion
   - Fréchet-Distanzfunktion

Einführung

- paarweiser Vergleich der Feature-Werte von Medienobjekten
- hier die häufigsten Distanzfunktionen analysiert nach Eigenschaften
- Eigenschaften nutzbar zur Konfiguration eines MMDBS bzgl. Suchszenario
- Distanzen auf Punkten, Binärdaten, Sequenzen und allgemeinen Mengen

6.1 Eigenschaften und Klassifikationen

- Abbildung Feature-Werte zweier Medien-Objekte auf nichtnegative, reelle Zahl
- Distanzwert 0 bedeutet maximale Ähnlichkeit
- Invarianz einer Distanzfunktion
  → also Unabhängigkeit bzgl. Operation
  g: $d(g(o_1), g(o_2)) = d(o_1, o_2)$
  – Translation
  – Skalierung
  – Rotation

Formale Eigenschaften einer Distanzfunktion

binäre Funktion $d(o_1, o_2)$ mit $d : O \times O \longrightarrow \mathbb{R}_0^+$ und

- Selbstidentität (Si): $\forall o \in O : d(o, o) = 0$
- Positivität (Pos): $\forall o_1 \neq o_2 \in O : d(o_1, o_2) > 0$
- Symmetrie (Sym):
  $\forall o_1, o_2 \in O : d(o_1, o_2) = d(o_2, o_1)$
- Dreiecksungleichung (Dreieck):
  $\forall o_1, o_2, o_3 \in O : d(o_1, o_3) \leq d(o_1, o_2) + d(o_2, o_3)$

Klassifikation anhand Erfüllung der Eigenschaften

Klasse	Si	Pos	Sym	Dreieck
Distanzfunktion	✓	✓	✓	✓
Pseudo-Distanzfunktion	✓	–	✓	✓
Semi-Distanzfunktion	✓	✓	✓	–
Semi-Pseudo-Distanzfunktion	✓	–	✓	–

Beispiele von Distanzfunktionen

- absoluter Betrag der Differenz zweier reeller Zahlen

  $$d_{abs} : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}_0^+, d_{abs}(r_1, r_2) \mapsto |r_1 - r_2|$$

- euklidsche Distanzfunktion d_{L_2} auf Punkten p_i der Menge $\mathbb{R}^n$

$$d_{L_2} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_0^+, d_{L_2}(p_1, p_2) \mapsto \sqrt{\sum_{i=1}^{n} (p_1[i] - p_2[i])^2}$$

Beispiel einer Pseudo-Distanzfunktion

indiskrete Pseudo-Distanzfunktion, die jedem Elementepaar aus $O \times O$ den Wert 0 zuweist:

$$d_{indiskret} : O \times O \longrightarrow \mathbb{R}_0^+, d_{indiskret}(o_1, o_2) \mapsto 0$$

(Funktion ist praktisch sinnlos)

Beispiel einer Semi-Distanzfunktion

Semi-Distanzfunktion d_{semi} auf der Menge {a, b, c}:

d_{semi}	a	b	c
a	0	1	3
b	1	0	1
c	3	1	0

Die Dreiecksungleichung ist nicht garantiert:

$$d_{semi}(a,c) \not\leq d_{semi}(a,b) + d_{semi}(b,c)$$
$$3 \not\leq 1+1$$

Weitere Eigenschaften von Distanzfunktionen

folgende Eigenschaften werden an konkreten Funktionen getestet:

- Invarianz bzgl.
  - Translation anhand Translationsobjekt *T*:
    $\forall o_1, o_2 : d(o_1, o_2) = d(o_1 + T, o_2 + T)$
  - Skalierung anhand Skalar *S*:
    $\forall o_1, o_2 : d(o_1, o_2) = d(o_1, S * o_2)$
  - Rotation anhand Rotationsobjekt *R:*
    $\forall o_1, o_2 : d(o_1, o_2) = d(R * o_1, R * o_2)$

Weitere Eigenschaften von Distanzfunktionen (2)

- Darstellung des Einheitskreises: alle Punkte $o \in O$, für die $d(z,o) = 1$ gilt (z ist Zentrum)

Distanzeigenschaften im Einheitskreis

verschiedene Eigenschaften sind graphisch aus Einheitskreis erkennbar:

- *Selbstidentität:* Zentrum liegt auf Kreis mit Radius 0.
- *Positivität:* Alle Punkte ungleich Zentrum liegen außerhalb des Kreises mit dem Radius 0

Distanzeigenschaften im Einheitskreis (2)

- *Translationsinvarianz:* Einheitskreis ändert Form nicht, wenn Zentrum verschoben wird
- *Symmetrie:* bei Translationsinvarianz und Symmetrie teilt Zentrum jede Diagonale zwischen zwei Randpunkten in genau zwei gleich lange Teile
- *Rotationsinvarianz:* Einheitskreis ist bzgl. Zentrum rotationssymmetrisch

6.2 Distanzfunktion auf Punkten

Datentyp: `array [1..n] (real)`

- Minkowski-Distanzfunktion L_m
- Gewichtete Minkowski-Distanzfunktion L_m^w
- Quadratische Distanzfunktion d_q
- Quadratische Pseudo-Distanzfunktion
- Dynamical-Partial-Semi-Pseudo-Distanzfunktion

6.2 Distanzfunktion auf Punkten (2)

- Chi-Quadrat-Semi-Pseudo-Distanzfunktion
- Kullback-Leibler-Abstandsfunktion
- Bhattacharyya-Abstandsfunktion

Minkowski-Distanzfunktion L_m

am häufigsten eingesetzte Distanzfunktion auf Punkten mit $m > 0$:

$$d_{L_m} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_0^+, d_{L_m}(p_1, p_2) \mapsto \left(\sum_{i=1}^{n} |p_1[i] - p_2[i]|^m \right)^{1/m}$$

$m = 1$: Manhattan-Distanzfunktion oder Blockdistanzfunktion
$m = 2$: euklidsche Distanzfunktion
$m = \infty$: Max-Distanzfunktion oder Tschebyscheff-Distanzfunktion

Sonderfall bei $m = \infty$:

$$d_{L_\infty} = d_{L_{max}}(p_1, p_2) \mapsto \max_{i=1}^{n} |p_1[i] - p_2[i]|$$

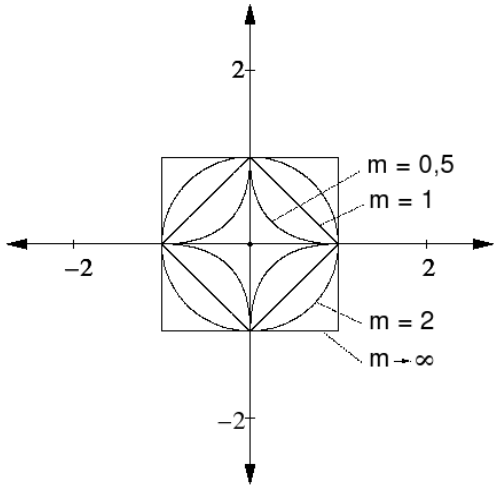
Translationsinvarianz

T sein ein n-dimensionaler Vektor, der durch die Differenzberechnung aus der Formel verschwindet:

$$\begin{aligned} d_{L_m}(p_1+T, p_2+T) &= \left(\sum_{i=1}^{n} |(p_1[i]+T)-(p_2[i]+T)|^m\right)^{1/m} \\ &= \left(\sum_{i=1}^{n} |p_1[i]-p_2[i]|^m\right)^{1/m} \\ &= d_{L_m}(p_1, p_2) \end{aligned}$$

aber keine Skalierungs- oder Rotationsinvarianz

Einheitskreise



Holdersche Ungleichung

es gilt immer:

$(|a_1|^{m_1} + \ldots + |a_n|^{m_1})^{1/m_1} \leq (|a_1|^{m_2} + \ldots + |a_n|^{m_2})^{1/m_2}$ für $m_1 \geq m_2 \geq 1$

also: Einheitskreis mit niedrigem m-Wert liegt innerhalb Einheitskreises mit höherem m -Wert

Sonderfall euklidsche Distanzfunktion (m=2)

- entspricht Länge der Geraden durch beide Punkte
- Einheitskreis ist kreisförmig
- Rotationsinvarianz ist erfüllt, da R Orthonormalmatrix $(R^T * R = R * R^T = I)$

$$\begin{aligned} d_{L_2}(R*p_1, R*p_2) &= \sqrt{(R*p_1 - R*p_2)^T * (R*p_1 - R*p_2)} \\ &= \sqrt{(R*(p_1-p_2))^T * (R*(p_1-p_2))} \\ &= \sqrt{(p_1-p_2)^T * R^T * R * (p_1-p_2)} \\ &= \sqrt{(p_1-p_2)^T * (p_1-p_2)} \\ &= d_{L_2}(p_1, p_2) \end{aligned}$$

- Matrizenschreibweise: $d_{L_2}(p_1, p_2) = \sqrt{(p_1-p_2)^T * (p_1-p_2)}$

Berechnung von Reihenfolgen anhand Minkowski-Dist.-fkt. L_m

Achtung: unterschiedliche m-Werte erzeugen unterschiedliche Reihenfolgen!

Beispiel: $p_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ und $p_2 = \begin{pmatrix} 0,8 \\ 0,8 \end{pmatrix}$

Abstände dieser Punkte vom Koordinaten-ursprung :

$$d_{L_1}(O, p_1) = 1 \text{ und } d_{L_1}(O, p_2) = 1.6$$

$$d_{L_\infty}(O, p_1) = 1 \text{ und } d_{L_\infty}(O, p_2) = 0,8$$

Gewichtete Minkowski-Distanzfunktion L_m^w

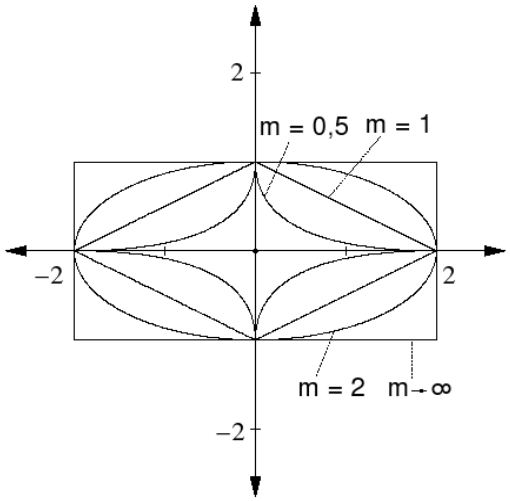
achsenparallele Stauchung und Streckung durch Gewichte $w_i \geq 0$:

$$d_{L_m}^w : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_0^+, d_{L_m}(p_1, p_2) \mapsto \left(\sum_{i=1}^{n} w_i * |p_1[i] - p_2[i]|^m \right)^{1/m}$$

Forderung:

$$\sum_{i=1}^{n} w_i = 1$$

Einheitskreis



Invarianzen

- Translationsinvarianz
- keine Skalierungsinvarianz
- keine Rotationsinvarianz

Quadratische Distanzfunktion d_q

Matrizenschreibweise:

$$d_q(p_1, p_2) = (p_1 - p_2)^T * A * (p_1 - p_2)$$

A im n-dimensionalen Raum ist symmetrische, positiv definierte Matrix $\mathbb{R}^{n \times n}$

Matrix *A*

- Einheitsmatrix E: d_q identisch mit $d^2_{L_2}$
- Diagonalmatrix: d_q entspricht $d^2_{L_2^w}$ (Gewichte korrespondieren zu Diagonalelementen)
- ansonsten: nonuniforme Skalierung, Rotation, Spiegelung der Punkte

Symmetrische positiv definierte Matrix *A*

es gilt immer: $A = U * L * U^T$ (Eigenwertzerlegung):

- U ist orthonormale Matrix (Rotation anhand Eigenvektoren)
- L ist Diagonalmatrix (Skalierung anhand Eigenwerten)

Symmetrische positiv definierte Matrix *A (2)*

- Berechnung der Distanz mittels $d^2_{L_2}$ auf transformierten Punkten oft relativ schnell realisierbar

$$\begin{aligned} d_q(p_1, p_2) &= (p_1 - p_2)^T A\ (p_1 - p_2) \\ &= (p_1 - p_2)^T U L U^T (p_1 - p_2) \\ &= \left(L^{1/2} U^T (p_1 - p_2)\right)^T \left(L^{1/2} U^T (p_1 - p_2)\right) \\ &= \left(L^{1/2} U^T p_1 - L^{1/2} U^T p_2\right)^T \left(L^{1/2} U^T p_1 - L^{1/2} U^T p_2\right) \\ &= d^2_{L_2}(L^{1/2} U^T p_1, L^{1/2} U^T p_2) \end{aligned}$$

Invarianzen

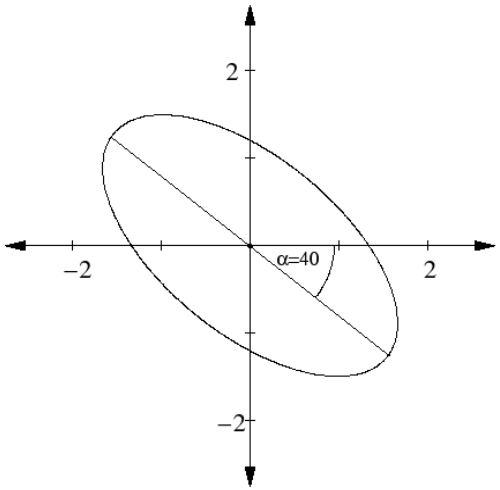
- Translationsinvarianz
- keine Skalierungsinvarianz
- keine Rotationsinvarianz

Beispielmatrix

$$A = \begin{pmatrix} 0,5599 & 0,3693 \\ 0,3693 & 0,6901 \end{pmatrix}$$

$$= \begin{pmatrix} \cos 40 & \sin 40 \\ -\sin 40 & \cos 40 \end{pmatrix} * \begin{pmatrix} 0,25 & 0 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} \cos 40 & -\sin 40 \\ \sin 40 & \cos 40 \end{pmatrix}$$

Einheitskreis des Beispiels



Mahalanobis-Distanzfunktion

- Einsatz der quadratischen Distanzfunktion d_q, wenn Distanzberechnung Kombination unterschiedlicher Dimensionen erfordert
- Grundlage kann Kovarianzmatrix C auf d Dimensionen sein
  → Mahalanobis-Distanzfunktion $d_M(p_1, p_2)$

$$d_M(p_1, p_2) = |det\ C|^{1/d}(p_1 - p_2)^T * C^{-1} * (p_1 - p_2)$$

Quadratische Pseudo-Distanzfunktion

- Aufgabe der Forderung nach Positiv-Definiertheit für *A*
- *Ziel: unsymmetrische Translationsinvarianz bzgl. Vektoren* t *des Vektorraums* T *:*

$$pd_q(p_1, p_2 + t) = pd_q(p_1, p_2)$$

- *Konstruktion der Matrix A aus geeigneter Orthogonalbases und Diagonalmatrix*

Quadratische Pseudo-Distanzfunktion (2)

- den U -Vektoren entsprechende Diagonal-werte auf Null setzen
- seien s_i mit $i = 1, \ldots m$ die durch l_i auf Null gesetzten U -Spaltenvektoren, dann gilt:

$$T = \left\{ t \in \mathbb{R}^n | t = \sum_{i=1}^{m} \lambda_i * s_i : \lambda_i \in \mathbb{R} \right\}$$

Nachweis der Translationsinvarianz

$$\begin{aligned}
& pd_q(p_1, p_2 + t) \\
= \; & (p_1 - p_2 - t)^T A (p_1 - p_2 - t) \\
= \; & (p_1 - p_2 - t)^T U L U^T (p_1 - p_2 - t) \\
= \; & (p_1 - p_2 - t)^T U L^{1/2} L^{1/2} U^T (p_1 - p_2 - t) \\
= \; & \left(L^{1/2} U^T (p_1 - p_2 - t)\right)^T \left(L^{1/2} U^T (p_1 - p_2 - t)\right) \\
= \; & \left(L^{1/2} U^T (p_1 - p_2) - L^{1/2} U^T t\right)^T \left(L^{1/2} U^T p_1 - p_2) - L^{1/2} U^T t\right) \\
= \; & \left(L^{1/2} U^T (p_1 - p_2)\right)^T \left(L^{1/2} U^T p_1 - p_2)\right) \\
= \; & pd_q(p_1, p_2)
\end{aligned}$$

Beispiel Quadratische Pseudo-Distanzfunktion

Konstruktion Translationsinvarianz im Winkel von 40 Grad:

$$U = \begin{pmatrix} \cos 40 & -\sin 40 \\ \sin 40 & \cos 40 \end{pmatrix}$$

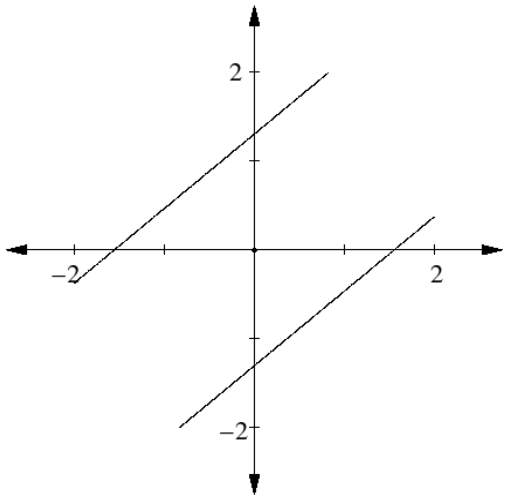
$$L = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Beispiel Quadratische Pseudo-Distanzfunktion (2)

Die Kombination dieser Matrizen ergibt die gewünschte Matrix *A:*

$$U * L * U^T = \begin{pmatrix} 0,4132 & -0.4924 \\ -0,4924 & 0,5868 \end{pmatrix}$$

Einheitskreis des Beispiels



Dynamical-Partial-Semi-Pseudo-Distanzfunktion

folgende Beobachtungen Chang/Wu03 bzgl. Unähnlichkeit im hochdimensionalen Raum:

- ähnliche Objekte liegen meist nur in wenigen Dimensionen nebeneinander
- Ähnlichkeit kann häufig nicht an bestimmten Dimensionen festgemacht werden

Dynamical-Partial-Semi-Pseudo-Distanzfunktion

Problem mit Minkowski-Distanzfunktion: alle Dimensionen werden berücksichtigt

- Berücksichtigung einer dynamischen Untermenge der Dimensionen

Dynamic-Partial-Semi-Pseudo-Distanzfunktion (2)

- p_1 und p_2 seien zwei Punkte im n-dimensionalen Raum und der Abstand in Dimension i
- nur die m kleinsten Abstände werden berücksichtigt: $\delta_i = |p_1[i] - p_2[i]|$

$\Delta_m = \{\texttt{die kleinsten m } \delta\texttt{-Werte aus } (\delta_1, \delta_2, \ldots, \delta_n)\}$

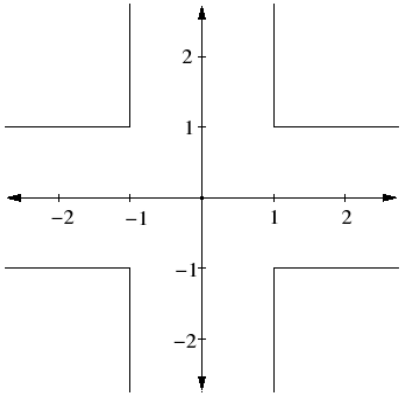
$$d_{dp}^{m,r} = \left(\sum_{\delta_i \in \Delta_m} \delta_i^r \right)^{\frac{1}{r}}$$

Eigenschaften

- Selbstidentität und Symmetrie sind erfüllt
- Verletzung der Positivität und Dreiecksungleichung

Einheitskreis

zweidimensionaler Raum und $m = 1$



Chi-Quadrat-Semi-Pseudo-Distanzfunktion

- Abstand zwischen Histogrammen mit absoluter Häufigkeit
- ursprünglich in Statistik entwickelt Untersuchung von Abhängigkeit zwischen Zufallsvariablen
- basiert auf Nullhypothese: Häufigkeitsverteilungen sind gleich also Differenz zwischen erwarteter und tatsächlicher Häufigkeit sind 0

Chi-Quadrat-Semi-Pseudo-Distanzfunktion (2)

$$spd_{\chi^2}(p_1,p_2)=\sum_{j=1}^{n}\frac{(p_1[j]-\hat{p}_1[j])^2}{\hat{p}_1[j]}+\sum_{j=1}^{n}\frac{(p_2[j]-\hat{p}_2[j])^2}{\hat{p}_2[j]} \text{ für } p_1,p_2\in\mathbb{N}_0^n$$

erwartete Häufigkeiten:

$$\hat{p}_i[j]=\frac{(p_1[j]+p_2[j])*\sum_{a=1}^{n}p_i[a]}{\sum_{a=1}^{n}(p_1[a]+p_2[a])}$$

Beispiel

- Test, ob Grippedoppelimpfung Grippe verhindern kann
- Befragung verschiedener Personen über Auftreten von Grippe und Impfungen
- erwartete Werte sind in Klammern notiert

	keine Impfung	eine Impfung	Doppelimpfung	$\sum$
Grippe	24 (14,398)	9 (5,014)	13 (26,588)	46
keine Grippe	289 (298,602)	100 (103,986)	565 (551,412)	954
$\sum$	313	109	578	1000

Berechnung der erwarteten Häufigkeiten

- wenn kein Zusammenhang zwischen Impfung und Gruppe, dann Wert jeder Zelle abschätzbar

Berechnung der erwarteten Häufigkeiten

- Beispiel Grippe/keine Impfung
  - Wahrsch. für Grippe ist 46/1000
  - Wahrsch. für keine Impfung ist 313/1000
  - Wahrsch. für Grippe/keine Impfung ist 46/1000*313/1000
  - erwartete Häufigkeit: 46/1000*313/1000*1000=46*313/1000=14,398

Eigenschaften

- Selbstidentität und Symmetrie sind erfüllt
- Rotationsinvarianz
- keine Positivität
- keine Dreiecksungleichung

Einheitskreis