DHBW CAS
Duale Hochschule
Baden-Württemberg
Center for Advanced Studies

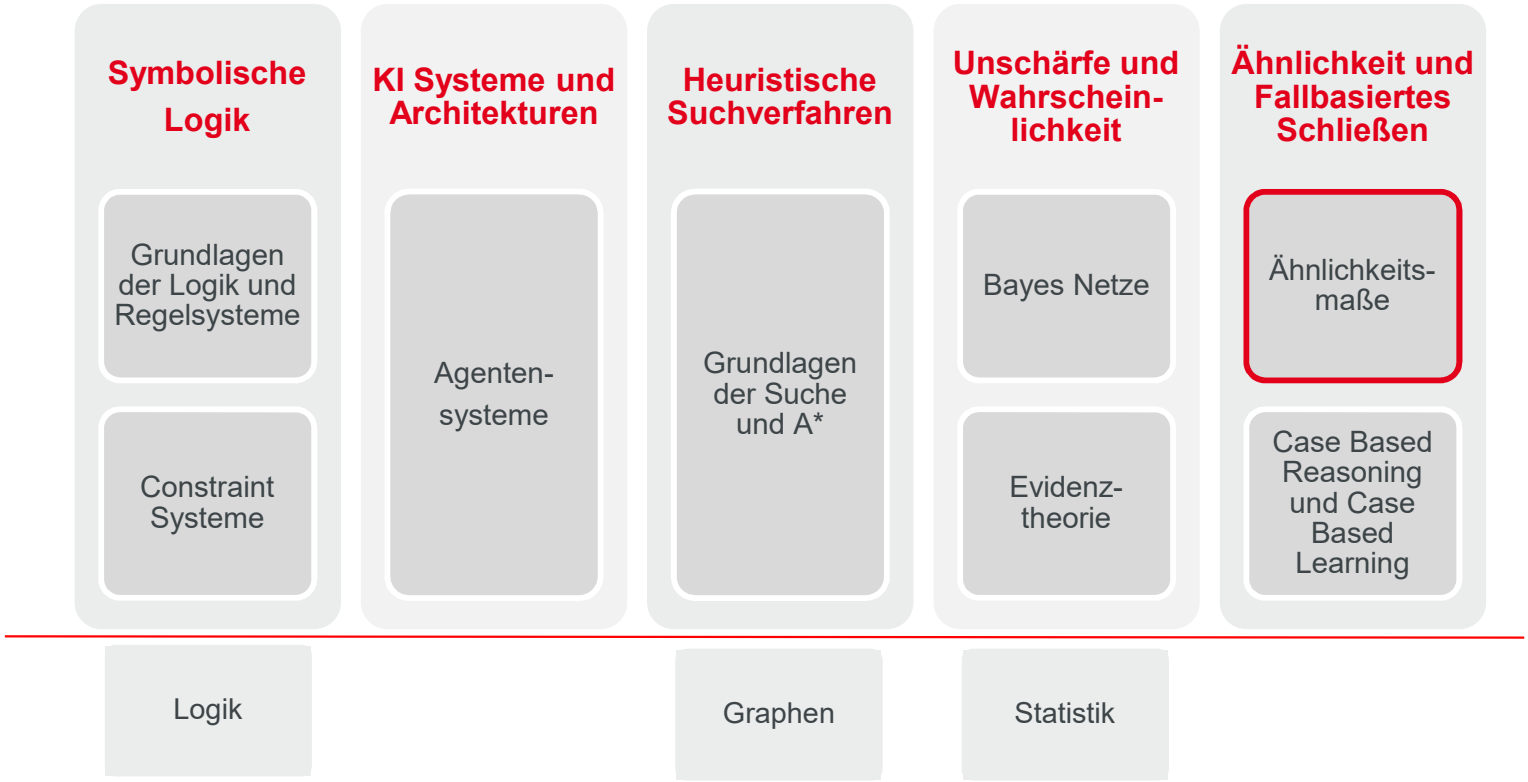
Methoden der Künstlichen Intelligenz & Computational Intelligence

Ähnlichkeit und Distanz – Konstruktion von Maßen

Prof. Dr. Dirk Reichardt

www.cas.dhbw.de

Modul – Teil 1 : Grundlagen Künstliche Intelligenz



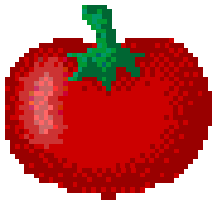
Formalisierung der Ähnlichkeit

Ähnlich?

Ja, beides sind **Autos**

... Ähnlichkeit bzgl. einer Kategorie, d.h. nicht unterscheidbar in der Abstraktion

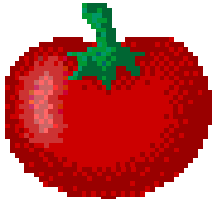
Ähnlich?



Ja, beide sind **rot**

... Ähnlichkeit bzgl. eines Merkmals (Farbe)

Ähnlich?

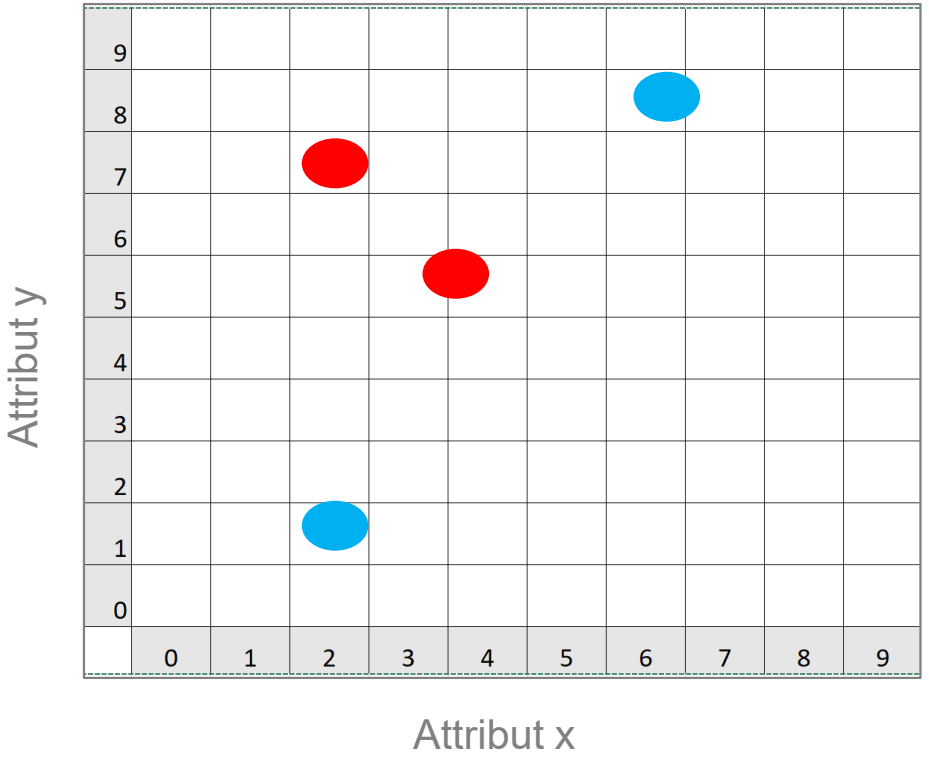


???

Ähnlichkeit und Distanz

Welche Objekte sind zueinander ähnlicher?

Die roten oder die blauen?



Wir unterscheiden zwei verschiedene Maße:

Ähnlichkeitsmaße : sim(x,y)

Distanzmaße: d (x,y)

Definition: Ähnlichkeitsmaße und Distanzmaße

Distanzmaß

Eine Abbildung d: M x M $\rightarrow \Re$ über einer Menge M heißt Distanzmaß, falls $\forall x,y,z \in M$ gilt:

(i) $d(x,x) = 0$ (Reflexivität)

(ii) $d(x,y) = d(y,x)$ (Symmetrie)

(iii) $d(x,y) = 0 \Leftrightarrow x = y$

Ergänzung:

Metrik Ist d ein Distanzmaß auf der Menge M und gilt zusätzlich

(iv) $d(x,y) + d(y,z) \geq d(x,z)$

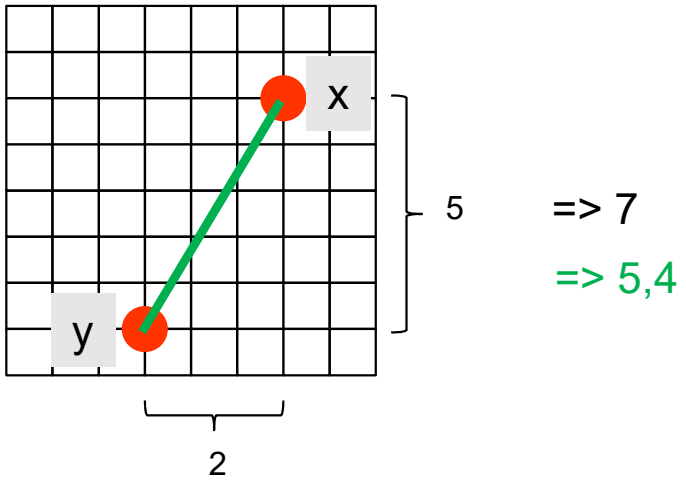
So ist d eine Metrik und (M,d) ein metrischer Raum.

Distanzmaße

Wertedistanz $d(x,y) = | x - y |$

Manhattan / City Block $d(x,y) = \sum_i |x_i - y_i|$

Euklidische Distanz $d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$



=> 7

=> 5,4

Minkowski Distanz

$$d_p(x,y) = \left[\sum_{k=1}^{n} |x_k - y_k|^p \right]^{\frac{1}{p}}$$

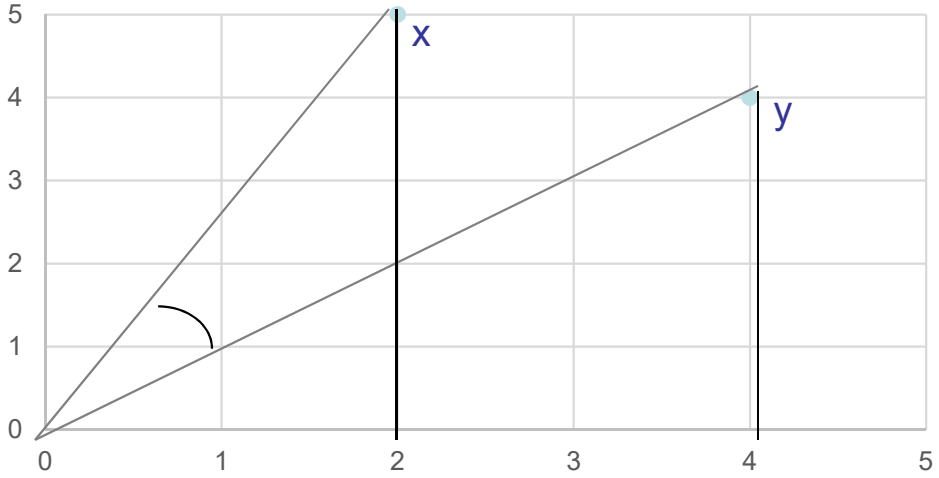
P = 1 — **MANHATTAN Distanz**

P = 2 — **EUKLIDISCHE Distanz**

Die Cosinus Distanz

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

Beispiel Cosinusähnlichkeit



x	2	5
y	4	4
xy	**8**	**20**
Cosinus Ähnlichkeit =		0,91914503
Cosinus Distanz =		0,08085497

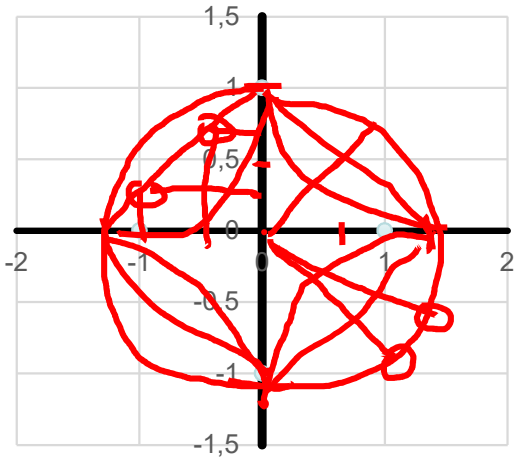
* Vorsicht: wenn gegenläufige Vektoren möglich sind, muss normiert werden auf [0,1]

Distanzmaße - Visualisierungsübung

Kurze Reflexionsübung:

Wo liegen die Elemente mit Einheitsdistanz 1
wenn die Minkowski Distanz mit folgenden
Parametern verwendet wird:

a) p = 1
b) p = 2
c) p = 1/2



Diskutieren Sie dies
mit der Nachbarin / dem Nachbarn.

Distanzmaße

Etwas komplizierter:

Typisierte Vektoren

Attribut	Wohnung A	Wohnung B
Zimmerzahl	4	4
Stockwerk	3	7
Hausmeister	ja	ja
Garage	nein	nein
Alter	75	12
Aufzug	ja	ja
Balkon	ja	ja
Terrasse	nein	nein
Kehrwoche	ja	nein
Moebliert	nein	nein
Miete	575	720
Nebenkosten	205	230
Quadratmeter	50	65
Bad	Dusche	Dusche+Wanne

$$X = (x_1, x_2, \dots, x_n)$$

Integer Buchstabe Begriff

$$d(A, B) = ?$$

Analogie und Ähnlichkeit - Merkmalstypen und Skalierungen

Ein Merkmal heißt **qualitativ**, wenn es nur eine endliche Anzahl von Ausprägungen besitzt.

- binär (nur zwei Ausprägungen)
- mehrstufig (mehr als zwei Ausprägungen)

Ein Merkmal heißt **quantitativ**, wenn die Ausprägungen numerisch sind.
(Werte aus einem endlichen oder unendlichen Intervall [a,b].)

Skalen:

Nominalskala	Keine Relation zwischen den Ausprägungen, nur unterscheidbar.
Ordinalskala	Totale Ordnung der Ausprägungen
Kardinalskala	Nicht nur die Ordnung ist bekannt, auch der Abstand

Messung und Skalen

Nominalskala

Im empirischen Relativ besteht eine Äquivalenzrelation. Messwerte enthalten ausschließlich Information über Gleichheit von Ausprägungen.

Beispiele?

Beruf, Wohnort, Nationalität, Geschlecht, …

Messung und Skalen

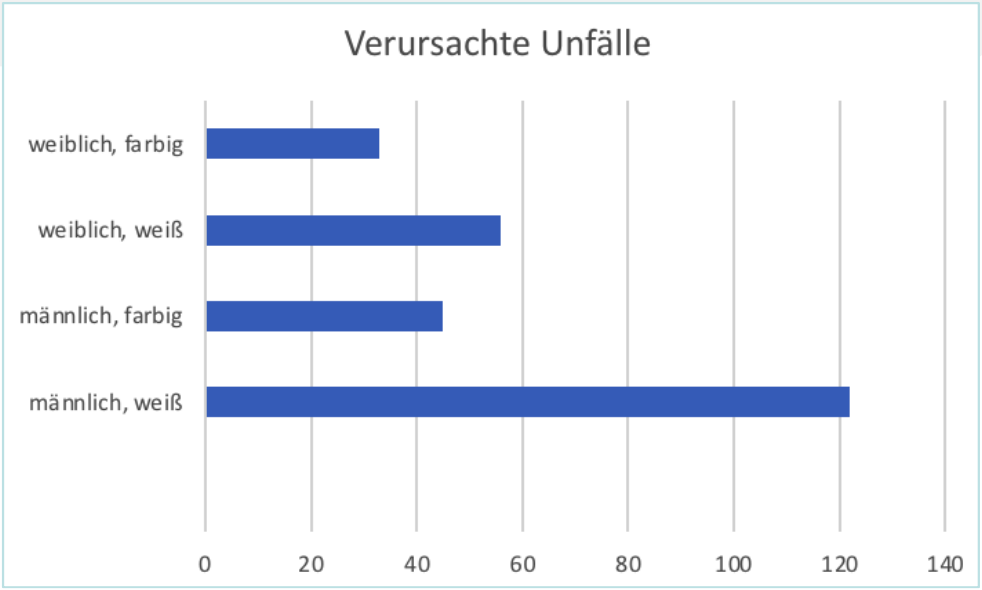
Ein Beispiel der Anwendung: *Unfallverursacher*

	farbig	weiß
männlich	1	0
weiblich	3	2

Mittelwert der Stichprobe: 1 ⟶

typische Unfallverursacher
sind männliche Farbige

Alles klar?



Messung und Skalen

Ordinalskala

Im empirischen Relativ besteht schwache Ordnungsrelation. Es ist also messbar ob ein Objekt eine stärkere/schwächere oder gleiche Merkmalsausprägung hat wie ein anderes.

Beispiele?

Charts, Militärische Ränge, Tabellenplatz, Beaufort Skala

Merken:
Mittelwert bei ordinal skalierten Werten nicht einsetzbar – alternativ: Median

Transformationen welche die Rangreihe nicht ändern sind zulässig.

Frage: Wie geht man mit Mittelwertangaben um?

Beispiel: Beaufort 2,4 und 9 – Mittelwert = 5
entspricht nicht der durchschnittlichen Windgeschwindigkeit!

Messung und Skalen

Intervallskala

Die Größe des Unterschieds zwischen Merkmalsausprägungen muss empirisch ermittelt werden können.
Es ist jedoch keine Aussage über Verhältnisse zwischen Messwerten möglich.

Beispiel?

Celsius Temperaturskala:
willkürlicher Nullpunkt, somit 20 nicht doppelt so warm wie 10

Lineare Transformationen sind zulässig.

Beispiel?
Umrechnung in Fahrenheit: $F = 1.8 * C + 32$

Messung und Skalen

seltener benötigt

Verhältnisskala

Die Größe des Unterschieds zwischen Merkmalsausprägungen muss empirisch ermittelt werden können und es gibt einen echten Nullpunkt (Bedeutung).
Es ist jedoch keine Aussage über Verhältnisse zwischen Messwerten möglich.

Beispiele?

Gewicht, Länge etc.

Absolutskala

Keine Transformation zulässig. Enthält z.B. Maßeinheit.

Beispiele?

Mitglieder einer Gruppe, Fehltage etc.

Skalenübersicht

Nominal	Gleich/Ungleich	Modus	
Ordinal	Größer/Kleiner	+ Median	
Intervall	Gleichheit von Differenzen	+ arithmetisches Mittel	$\frac{x+y}{2}$
Verhältnis	Gleichheit von Verhältnissen	+ geometrisches Mittel	$\sqrt{x \times y}$
Absolut	Maßeinheit		

Definition: Ähnlichkeitsmaße und Distanzmaße

Ähnlichkeitsmaß

Eine Abbildung sim: M x M $\rightarrow$ [0,1] über einer Menge M heißt Ähnlichkeitsmaß, falls $\forall x,y,z \in M$ gilt:

(i) sim(x,x) = 1 (Reflexivität)

(ii) sim(x,y) = sim(y,x) (Symmetrie)

(iii) sim(x,y) = 1 $\Leftrightarrow$ x = y

Abschwächung:

Ähnlichkeitsmaße müssen nicht unbedingt symmetrisch und transitiv sein.

Man kann nicht von einer „absoluten“ Ähnlichkeit sprechen, die Ähnlichkeit ist immer abhängig vom Anwendungszweck.

Transformation

Kompatibilität

Für ein Distanzmaß d(x,y) kann eine Relation R(x,y,u,v) definiert werden durch:

(i) $R_d(x,y,u,v) \Leftrightarrow d(x,y) \leq d(u,v)$

Für ein Ähnlichkeitsmaß sim(x,y) kann eine entsprechende Definition gegeben werden:

(ii) $R_{sim}(x,y,u,v) \Leftrightarrow sim(x,y) \geq sim(u,v)$

sim(x,y) und d(x,y) heißen kompatibel, falls gilt:

(iii) $R_d(x,y,u,v) \Leftrightarrow R_{sim}(x,y,u,v)$

Oder:

Kann man eine bijektive, ordnungsinvertierende Abbildung mit f(0) = 1
und sim(x,y) = f(d(x,y)) angeben, so sind d und sim kompatibel.

Transformationsfunktionen

Beispiele:

$$F(x) = 1 - \frac{x}{x+1} \qquad F(x) = 1 - \left(\frac{x}{x+1}\right)^n$$

Ist ein maximaler Abstand max bekannt, dann auch:

$F(x) = 1 - x / max$

Weitere Transformationen:

$d(x,y) = \sqrt{(1- sim(x,y))}$

$d(x,y) = - \log(sim(x,y))$

Ähnlichkeit und Ähnlichkeitsmaße

Ähnlichkeitsmaße auf Boole'schen Vektoren

$x = (x_1,\ldots,x_n)$
$y = (y_1,\ldots,y_n)$

*Zwei Objekte, die mit den **Merkmalsvektoren** x und y beschrieben werden*

x \ y	1	0
1	a	b
0	c	d

Formale Festlegung von 4 Beschreibungsgrößen:

$a = \sum(x_i, y_i)$ $b = \sum(x_i, \neg y_i)$ $c = \sum(\neg x_i, y_i)$ $d = \sum(\neg x_i, \neg y_i)$

Anzahl der positiven Übereinstimmungen

Anzahl der negativen Übereinstimmungen

Wie muß ein Ähnlichkeitsmaß aussehen?

1) sim(x,y) ***wächst monoton*** mit a und d
2) sim(x,y) ist ***symmetrisch*** in b und c
3) sim(x,y) ***fällt monoton*** mit b und c

Ähnlichkeit und Ähnlichkeitsmaße

Distanzmaß: d(x,y) = b + c „***Hammingabstand***“

Daraus abgeleitet: der „***Simple Matching Coefficient***“ (SMC)

sim(x,y) = 1 - (b+c) / max = 1 - (b+c)/(a+b+c+d) = (a+d)/(a+b+c+d)

Optimistische und *pessimistische* Auslegung des SMC

$$sim(x,y) = \frac{\alpha\,(a+d)}{\alpha(a+d) + (1-\alpha)(c+b)}$$

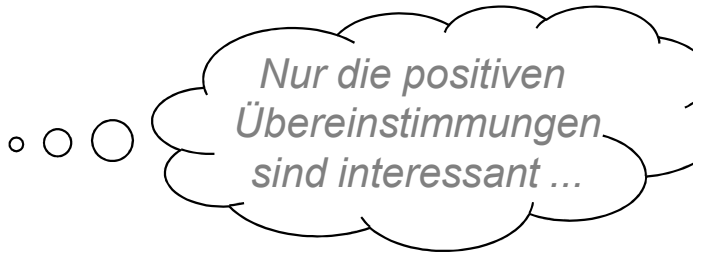
Für α = 1/2 ist es der SMC

Ist $\alpha > 1/2$ so ist die Auslegung *optimistisch*, bei $\alpha < 1/2$ *pessimistisch*.

Asymmetrische Maße

sim(x,y) = a / (a+b+c) ***Jaccard Koeffizient***

sim(x,y) = a / (a+b+c+d) ***Russel Rao***



Ähnlichkeit und Ähnlichkeitsmaße

"Informierte" Ähnlichkeitsmaße

Die bisher vorgestellten Ähnlichkeitsmaße behandeln alle Attribute gleich.

Ein Mietkarteibeispiel:

Attribut A :	*Hausmeister*	{ja, nein}	2 Ausprägungen
Attribut B :	*Zimmerzahl*	{1,1-2,2,2-3,3,3-4,4,4-5,5,6}	10 Ausprägungen

Welche 2 Wohnungen sind sich ähnlicher?

? W_1 : (Hausmeister=ja) $\wedge$ (Zimmerzahl=3)
W_2 : (Hausmeister=ja) $\wedge$ (Zimmerzahl=6)

? W_3 : (Hausmeister=nein) $\wedge$ (Zimmerzahl=6)
W_4 : (Hausmeister=ja) $\wedge$ (Zimmerzahl=6)

Ein Ähnlichkeitsmaß, das die *Alternativenzahl* berücksichtigt

$$sim(x,y) = 1/m * \sum m_i * f(x_i,y_i)$$

mit $m = \sum m_i$ und f(x,y) = 1 wenn x=y, sonst 0

$$sim(W_1,W_2) = (2 \cdot f(w_{1,1},w_{2,1}) + 10 \cdot f(w_{1,2},w_{2,2}))/12 = 1/6$$

$$sim(W_3,W_4) = (2 \cdot f(w_{3,1},w_{4,1}) + 10 \cdot f(w_{3,2},w_{4,2}))/12 = 5/6$$

Ähnlichkeit und Ähnlichkeitsmaße

"Informierte" Ähnlichkeitsmaße

Eine weitere Information bei Vektoren binärer Attribute:

Attributvektor A: { A1, A2, A3, A4, A5 }

P_i **10% 95% 2% 50% 80%** → *... aller Daten haben den Wert 1!*

Welche Objekte sind sich ähnlicher?

? W_1 : 1, 0, 0, 0, 0
W_2 : 1, 1, 1, 1, 1

? W_3 : 0, 1, 0, 1, 0
W_4 : 1, 1, 1, 0, 1

Ein *mögliches* Ähnlichkeitsmaß, das die *Merkmalshäufigkeit* berücksichtigt

$$sim(x,y) = \frac{(a' + d')}{(a' + b + c + d')}$$

mit $a' = \sum_i (x_i \cdot y_i) \cdot (1-P_i)$
$d' = \sum_i (\neg x_i \cdot \neg y_i) \cdot P_i$

$sim(W_1,W_2) = 0.9/(0.9+4) = 0.18$

$sim(W_3,W_4) = 0.05/(0.05+3+1) = 0.01$

Ähnlichkeit und Ähnlichkeitsmaße

"Informierte" Ähnlichkeitsmaße - Wichtigkeit von Attributen

Nicht alle Angaben sind für ein Konzept gleich wichtig!

Beispiel: Die Ähnlichkeit zweier Personen wird eher am *Gesicht* als am *Körperbau* gemessen.

Welche Objekte sind sich ähnlicher? **W = 0.5 0.1 0.1 0.15 0.15**

? X_1 : 1, 0, 0, 0, 0
X_2 : 1, 1, 1, 1, 1

? X_3 : 0, 1, 0, 1, 0
X_4 : 1, 1, 1, 0, 1

Ein *mögliches* Ähnlichkeitsmaß, das die *Merkmalswichtigkeit* berücksichtigt

$$sim(x,y) = w_1 \cdot (x_1 = y_1) + w_2 \cdot (x_2 = y_2) + ... + w_n \cdot (x_n = y_n) \qquad \text{mit } 1 = \sum_i w_i$$

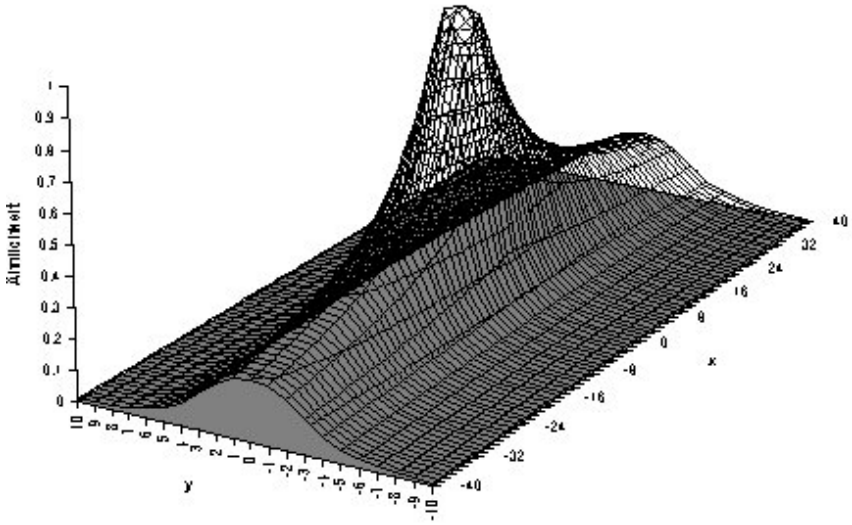
$sim(X_1,X_2) = 0.5$

$sim(X_3,X_4) = 0.1$

Ähnlichkeit - Anwendung im Fahrzeug

„Ähnlichkeit“ verwendet als Kriterium
zum Matching zwischen zwei Messungen
von Objekterkennern.

$$sim(o_1,o_2) = 1 - (d(o_1,o_2)/(1+d(o_1,o_2)))^k$$



K =4, $w_{Distanz}$ = 1, w_{Offset} = 2,
$e_{Distanz}$ = 2, e_{Offset} = 3

Gewichtungsfunktion:

$$f(w,e,d) = w * d^e$$

w = Gewichtung eines Attributs
d = Betrag der Differenz von zwei Attributwerten
e = Toleranzfaktor

r(sensor,attribut,hindernis) im Wertebereich [0,1] - eine Zuverlässigkeitsfunktion

v(A) = Wert des Attributs A

$$d(o_1,o_2) = \frac{\sum_i f(w_{Ai},e_{Ai},|v(_{Ai},o_1) - v(A_i,o_2)|)*r(s(o_1),A_i,o_1)*r(s(o_2),a_i,o_2)}{(\sum_i w_{Ai} * r(s(o_1),A_i,o_1)*r(s(o_2),A_i,o_2)}$$

Was ist die Minkowski Distanz?

Was sind informierte Ähnlichkeitsmaße?

Welche Transformationsfunktionen gibt es?

Was sind die Eigenschaften von Skalen?