Skript zur Vorlesung

Knowledge Discovery in Databases

im Sommersemester 2013

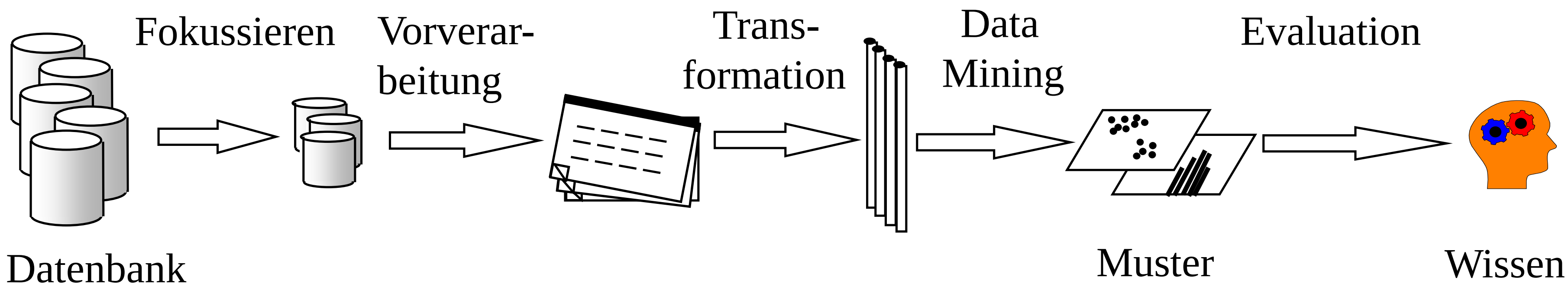
Kapitel 2: Preprocessing, Merkmalsräume

Vorlesung: Dr. Arthur Zimek
Übungen: Erich Schubert

http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)

Wiederholung: Das KDD-Prozessmodell

Prozessmodell nach Fayyad, Piatetsky-Shapiro & Smyth



Fokussieren:
- Beschaffung der Daten
- Verwaltung (File/DB)
- Selektion relevanter Daten

Vorverarbeitung:
- Integration von Daten aus unterschiedlichen Quellen
- Vervollständigung
- Konsistenzprüfung

Transformation
- Diskretisierung numerischer Merkmale
- Ableitung neuer Merkmale
- Selektion relevanter Merkm.

Data Mining
- Generierung der Muster bzw. Modelle

Evaluation
- Bewertung der Interessantheit durch den Benutzer
- Validierung: Statistische Prüfung der Modelle

Preprocessing

- Daten sind oft unsauber (verrauscht), unvollständig, inkonsistent:
  - Unsauber/verrauscht: Fehler, Outlier
    - Fehlerhafte Werte (z.B., Gehalt=-10000)
    - Unerwartete Werte (z.B. Gehalt=100000, wenn alle anderen Werte im Bereich 30000-50000 liegen)
  - Unvollständig (fehlende Werte)
    - Fehlende Attribute, die von Interesse für eine Aufgabe wären (z.B. keine Information über Beruf)
    - Fehlende Werte (z.B. Beruf=„“)
  - Inkonsistent
    - z.B. studentische Bewertungen können für verschiedene Universitäten unterschiedlich skalieren
- Unsaubere Daten ➔ schlechte Data Mining Ergebnisse

Hauptaufgaben im Preprocessing

- Data cleaning:
  - Fehlende Werte errechnen, verrauschte Daten glätten, Identifikation oder Entfernen von Outliern, Auflösung von Inkonsistenzen
- Data integration:
  - Integration mehrerer Datenbanken, Dateien (Entity identification, Value resolution)
- Data transformation:
  - z.B. Normalisierung
  - Generalisierung (z.B. durch Konzept-Hierarchie)
- Data reduction:
  - Aggregation
  - Feature-Reduktion
  - Duplikat-Eliminierung

Daten Objekte/Beispiele

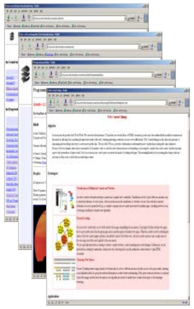
- Daten bestehen aus Objekten /Beispielen (objects, examples, instances)
  - z.B., Filmdatenbank: Filme, Schauspieler, Regisseure,…
  - z.B., Bibliotheksdatenbank: Bücher, Nutzer, Gebühren, Ausleihfristen ….
  - z.B., Universitätsdatenbank: Studenten, Professoren, Lehrveranstaltungen, Noten,…
- Objekte werden durch Merkmale (features/ attributes/variables) beschrieben
  - z.B. in einer Datenbank-Tabelle: Zeilen sind Objekte, Spalten sind Merkmale

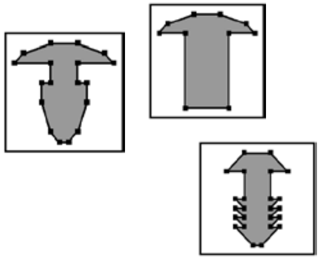
id	person	name	web	bio	location	following	followers
8	1	Justin Bieber	http://www.youtube.com/justinbieber	www.BieberFever.comRequest my NEW...	on the MY WORLD TOUR!!!	88045	5792472
9	2	Perez Hilton	http://www.PerezHilton.com	Perez Hilton is the creator and writer of o...	Hollywood, California	341	2566969
10	3	Paris Hilton	http://www.parishilton.com	Huge!	ÜT: 35.975487,-115.141709	842	2915057
11	4	Britney Spears	http://www.britneyspears.com	It's Britney Bitch!	Los Angeles, CA	417405	6168589
12	5	Kim Kardashian	http://kimkardashian.celebuzz.com/	business woman, exec producer, fashion...	on a plane...	96	5139761
13	6	Mariah Carey				0	3400111
14	7	Shakira	http://www.shakira.com	Welcome to Shakira's Official Twitter pag...	Bahamas	33	3318367
15	8	Justin Timberlake	http://www.justintimberlake.com	Official Justin Timberlake Twitter.	Memphis, TN	19	3339151
16	9	Gov. Schwarzenegger	http://gov.ca.gov	As California's 38th Governor I look forwa...	Sacramento, California	110763	1824274
17	10	Serena Williams	http://www.serenawilliams.com	Living, Loving, and working to help you...	Paris	84	1819778
18	11	Larry King	http://www.CNN.com/LarryKing	CNN's Larry King Live	LA	183	1721681
19	12	Panos Ipeirotis	http://behind-the-enemy-lines.blogspot.com/	Associate Professor at Stern School of B...	New York, NY	156	547

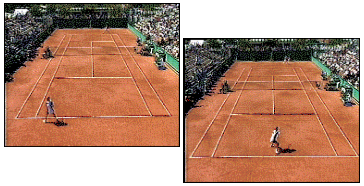
ID	title	URL	unknc	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentar	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFi	Thrille
1	Toy Story (1995)	http://us.imdb	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	
2	GoldenEye (1995)	http://us.imdb	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Four Rooms (1995)	http://us.imdb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Get Shorty (1995)	http://us.imdb	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	
5	Copycat (1995)	http://us.imdb	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	
6	Shanghai Triad (Yao a yao yao dao waipo qiao	http://us.imdb	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
7	Twelve Monkeys (1995)	http://us.imdb	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	
8	Babe (1995)	http://us.imdb	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	
9	Dead Man Walking (1995)	http://us.imdb	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
10	Richard III (1995)	http://us.imdb	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
11	Seven (Se7en) (1995)	http://us.imdb	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
12	Usual Suspects, The (1995)	http://us.imdb	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	

Merkmalsräume

- Motivation:
  - Zentrales Konzept beim Data Mining: Ähnlichkeit von Datenbankobjekten
    - Clustering: Zusammenfassen *ähnlicher* Objekte in Gruppen
    - Klassifikation: Zuordnung von Objekten zu einer Klasse *ähnlicher* Objekte
  - Definition einer geeigneten Distanzfunktion auf Datenbankobjekten nicht immer einfach (besonders in Nicht-Standard-Datenbanken)
    - Bilder
    - CAD-Objekte
    - Proteine
    - Textdokumente
    - Polygonzüge (GIS)
    - etc.



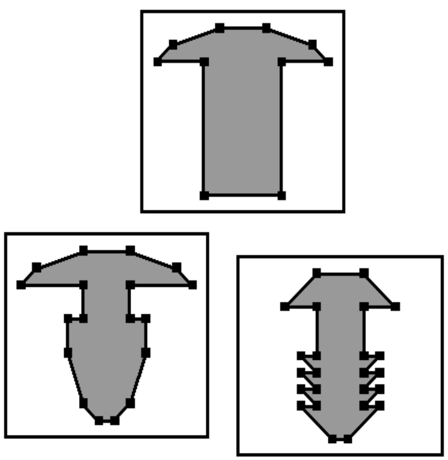


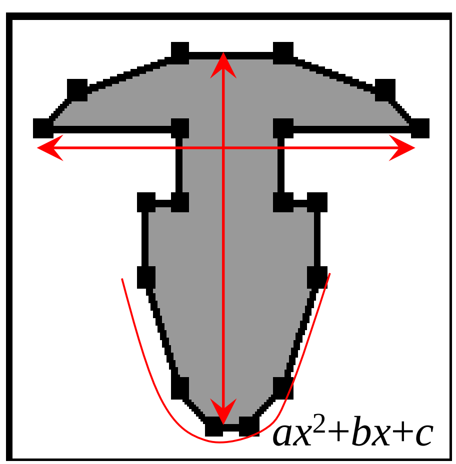


Merkmale („Features“ von Objekten)

- Oft sind die betrachteten Objekte komplex
- Eine Aufgabe des KDD-Experten ist dann, geeignete Merkmale (*Features*) zu definieren bzw. auszuwählen, die für die Unter-scheidung (Klassifikation, Ähnlichkeit) der Objekte relevant sind.

Beispiel: CAD-Zeichnungen:



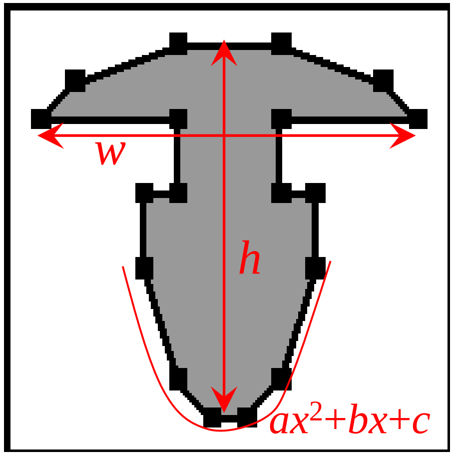


Mögliche Merkmale:

- Höhe *h*
- Breite *w*
- Kurvatur-Parameter (a,b,c)

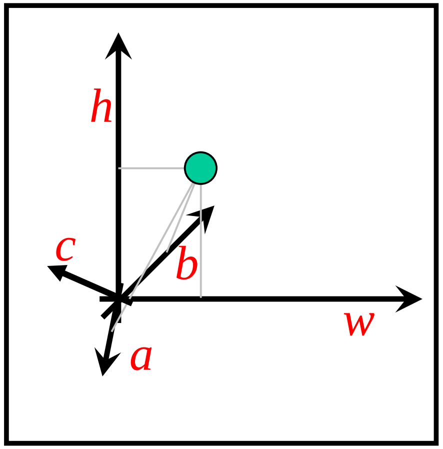
Beispiel: CAD-Zeichnungen (cont.)

Objekt-Raum



$$(h, w, a, b, c)$$

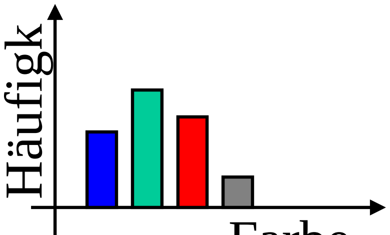
Merkmals-Raum



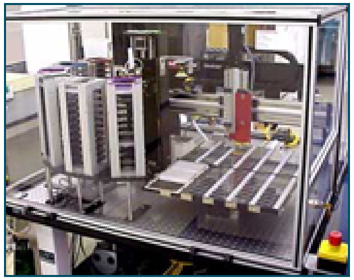
- Im Kontext von statistischen Betrachtungen werden die Merkmale häufig auch als *Variablen* bezeichnet
- Die ausgewählten Merkmale werden zu Merkmals-Vektoren (*Feature Vector*) zusammengefasst
- Der Merkmalsraum ist häufig hochdimensional (im Beispiel 5-dim.)

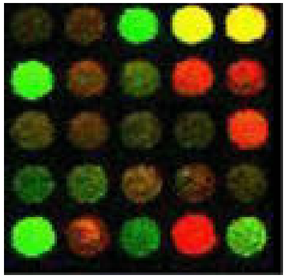
Merkmale

Bilddatenbanken:
Farbhistogramme

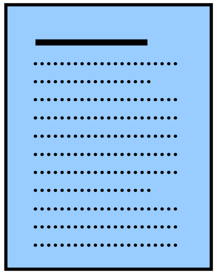


Gen-Datenbanken:
Expressionslevel





Text-Datenbanken:
Begriffs-Häufigkeiten



Data	25
Mining	15
Feature	12
Object	7
...	

Der Feature-Ansatz ermöglicht einheitliche Behandlung von Objekten verschiedenster Anwendungsklassen

Skalen-Niveaus von Merkmalen

Nominal (kategorisch)

Charakteristik:

Nur feststellbar, ob der Wert gleich oder verschieden ist. Keine Richtung (besser, schlechter) und kein Abstand. Merkmale mit nur zwei Werten nennt man *dichotom*

Beispiele:

Geschlecht (dichotom)
Augenfarbe
Gesund/krank (dichotom)

Ordinal

Charakteristik:

Es existiert eine Ordnungsrelation (besser/schlechter) zwischen den Kategorien, aber kein einheitlicher Abstand

Beispiele:

Schulnote (metrisch?)
Güteklasse
Altersklasse

Metrisch

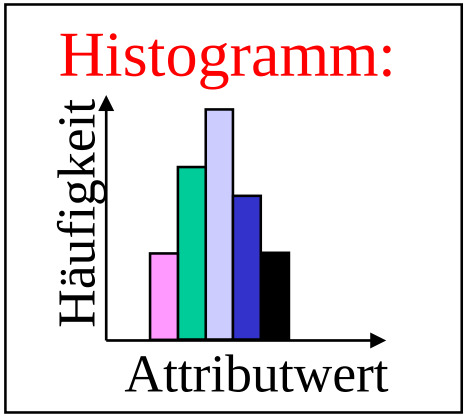
Charakteristik:

Sowohl Differenzen als auch Verhältnisse zwischen den Werten sind aussagekräftig. Die Werte können diskret oder stetig sein.

Beispiele:

Gewicht (stetig)
Verkaufszahl (diskret)
Alter (stetig oder diskret)

Univariate Deskription von Merkmalen



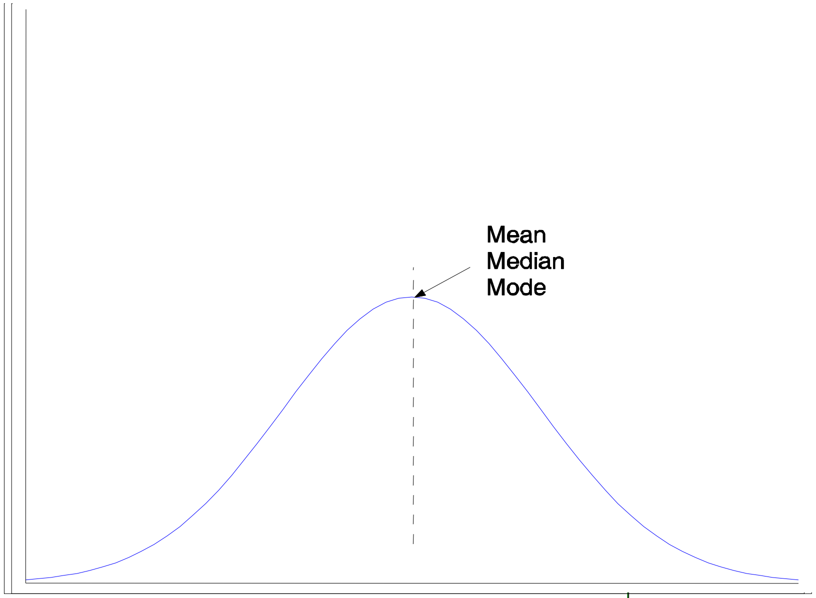
Sei $x_1,\ldots,x_n$ eine Stichprobe eines Merkmals X.

- Absolute Häufigkeit: Für jeden Wert a ist $h(a)$ die Anzahl des Auftretens in der Stichprobe
- Relative Häufigkeit: $p(a) = h(a)/n$

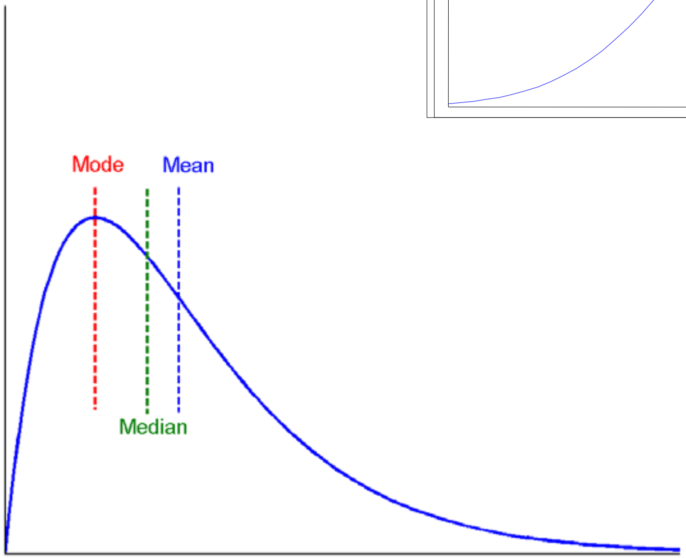
Die folgenden Maße sind nur für metrische Merkmale sinnvoll:

- Arithmetisches Mittel: $\mu = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$
- Median: *Das mittlere Element bei aufst. Sortierung*
- Modus (mode): *Ausprägung mit größter Häufigkeit*
- Varianz: $VAR(X) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$
- Standardabweichung: $\sigma = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \mu)^2}$

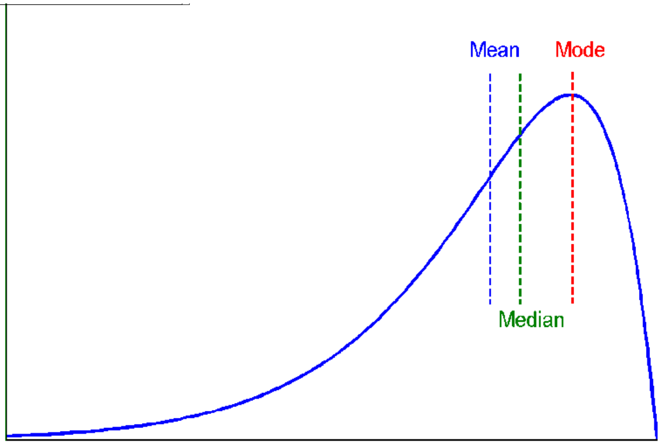
Symmetric vs. Skewed



Symmetric



Positively skewed



Negatively skewed

Kontingenztabelle

- für kategorische Merkmale *X* und *Y*
- repräsentiert für zwei Merkmale *X* und *Y* die absolute Häufigkeit h_{ik} jeder Kombination (x_i, y_k) und alle Randhäufigkeiten $h_{.k}$ und $h_{i.}$ von *X* und *Y*

	MittelfristigeArbeitslosigkeit	Langfristige Arbeitslosigkeit	
Keine Ausbildung	19	18	37
Lehre	43	20	63
	62	38	100

- Wie sollten die relativen Häufigkeiten verteilt sein, wenn die beiden Merkmale keinerlei Abhängigkeit besitzen? $p_i = \frac{h_{i.}}{n}, p_{ij} = p_i p_j$
- χ^2-*Koeffizient*
  Differenz zwischen dem bei Unabhängigkeit erwarteten und dem tatsächlich beobachteten Wert von h_{ij} (Maß für die Stärke der Abhängigkeit)

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

o_{ij}: beobachtete Häufigkeit
e_{ij}: erwartete Häufigkeit

$$e_{ij} = n \cdot p_i \cdot p_j = \frac{h_i h_j}{n}$$

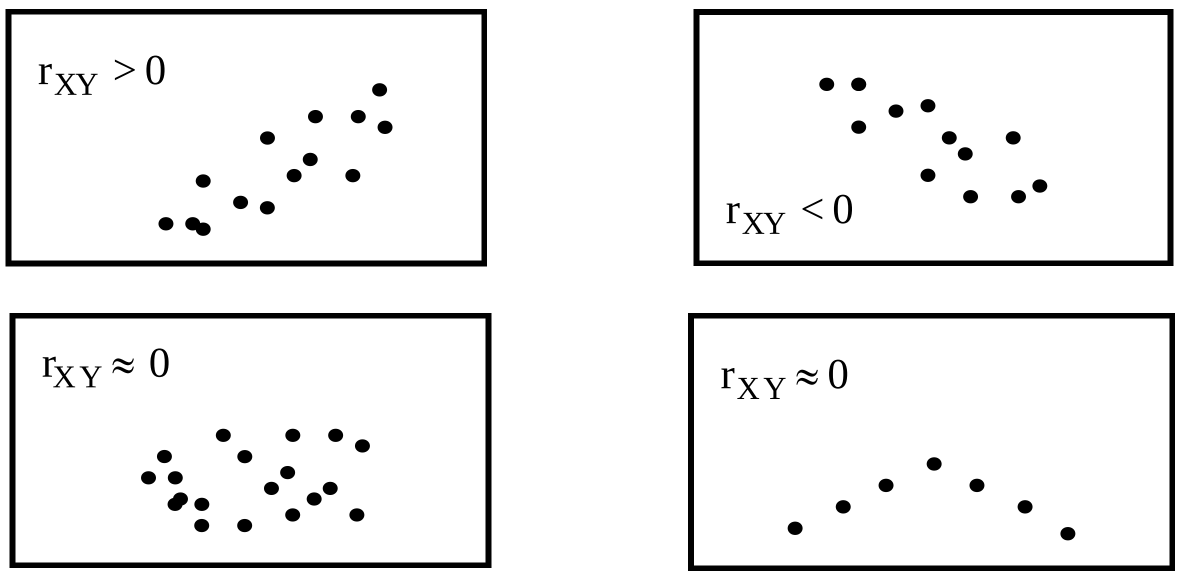
Multivariate Deskription von Merkmalen

Korrelationskoeffizient

- für numerische Merkmale *X* und *Y*
- wie stark sind die Abweichungen vom jeweiligen Mittelwert korreliert?

$$r_{XY} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

- Beispiele



Multivariate Deskription von Merkmalen

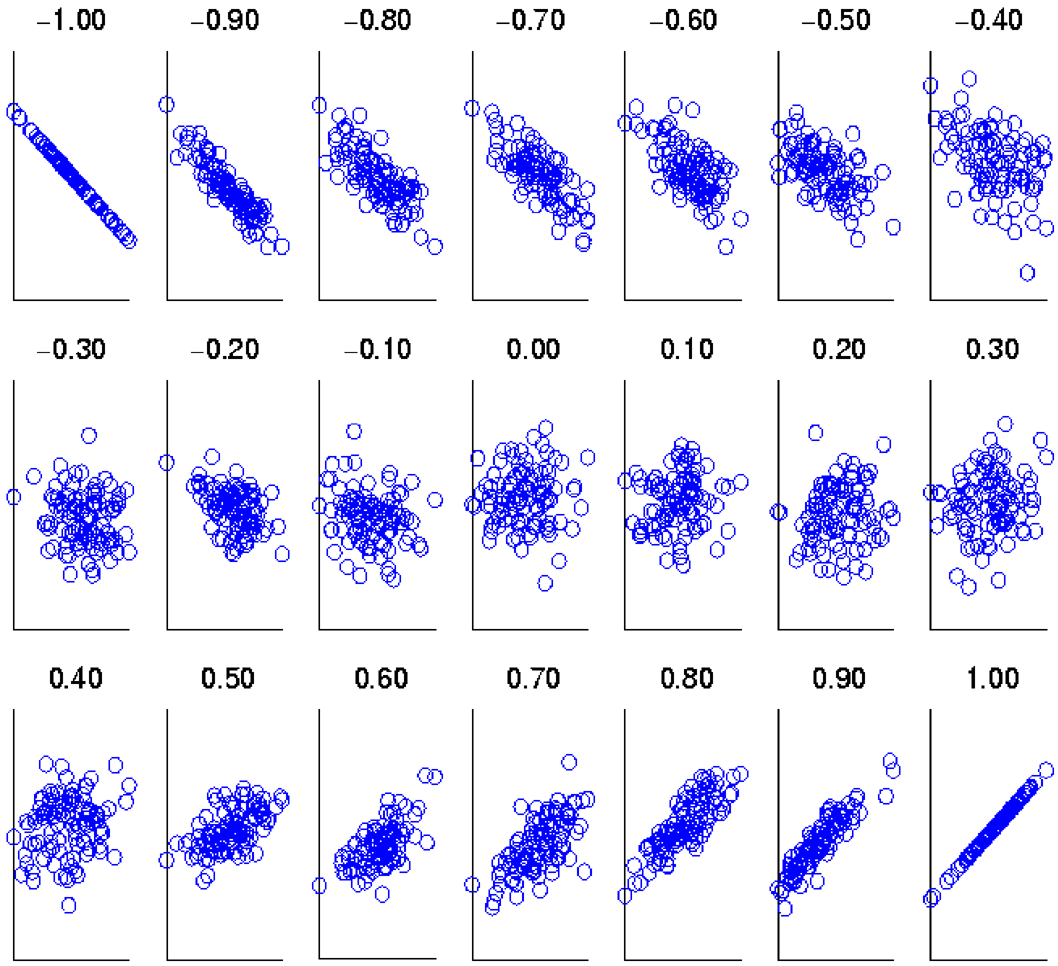


Figure 5.11. Scatter plots illustrating correlations from -1 to 1.

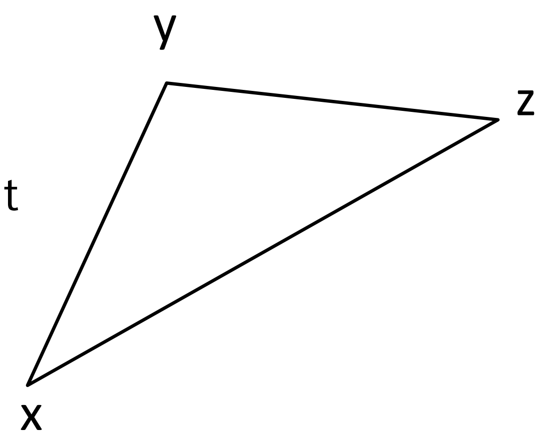
Räume und Distanzfunktionen

Merkmalsraum (Featureraum)

- Intuitiv: ein Wertebereich/Domain mit Distanzfunktion
- Formal: Featureraum **F** = (*Dom*, *dist*)
- *Dom* ist eine (geordnete) Menge von Merkmalen (Features)
- $dist : Dom \times Dom \rightarrow \mathbb{R}_0^+$ ist eine totale (Distanz)-Funktion mit den folgenden Eigenschaften
  - $\forall p,q \in Dom,\ p \neq q : dist(p,q) > 0$ Striktheit
  - $\forall o \in Dom : dist(o,o) = 0$ Reflexivität
  - $\forall p,q \in Dom : dist(p,q) = dist(q,p)$ Symmetrie

Räume und Distanzfunktionen

- Metrischer Raum
  - Formal: Metrischer Raum **M** = (*Dom*, *dist*) mit den folgenden Eigenschaften
    - **M** ist ein Featureraum
    - $\forall o,p,q \in Dom : dist(o,p) \leq dist(o,q) + dist(q,p)$ Dreiecksungleichung



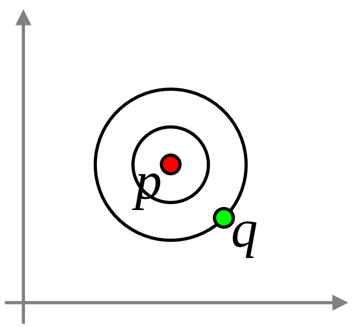
- Wichtigstes Beispiel: Euklidischer Vektorraum
  - Formal: Euklidischer Vektorraum **E** = (*Dom*, *dist*) mit
    - (*Dom*, *dist*) ist ein metrischer Raum
    - $Dom = \mathbb{R}^d$
    - $dist = (x, y) \mapsto \|x - y\|_2$
- Sprechweise:
  - Euklidischer Vektorraum = „Featureraum“
  - Vektoren (Objekte im Euklidischen Featureraum) = „Featurevektoren“
  - Die *d* Dimensionen des Vektorraums = „Features“

Räume und Distanzfunktionen

- Ähnlichkeit von Feature Vektoren (Euklidische Vektoren)

Euklidische Norm (L_2):

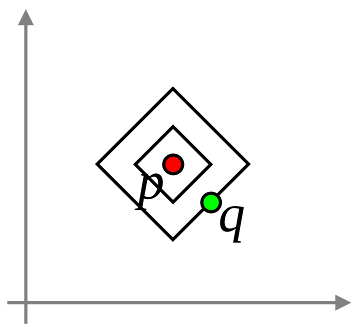
$dist_2 = ((p_1-q_1)^2+(p_2-q_2)^2+\ldots)^{1/2}$



Natürlichstes Distanzmaß

Manhattan-Norm (L_1):

$dist_1 = |p_1-q_1|+|p_2-q_2|+\ldots$

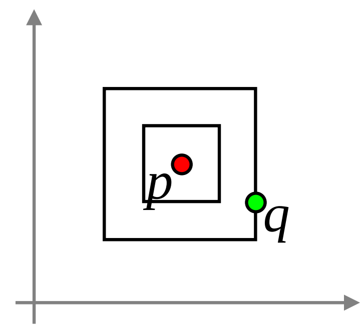


Die Unähnlichkeiten der einzelnen Merkmale werden direkt addiert

Maximums-Norm (L_∞):

$dist_\infty = \max\{|p_1-q_1|, |p_2-q_2|,\ldots\}$

auch: L_{max}, supremum dist., Chebyshev dist.



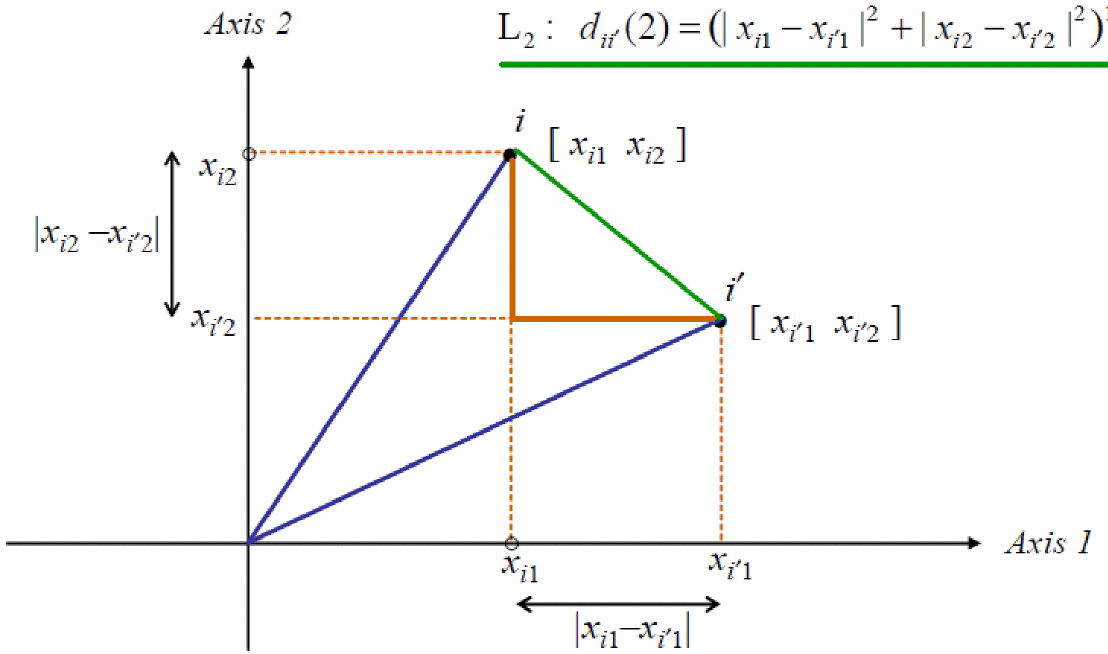
Die Unähnlichkeit des am wenigsten ähnlichen Merkmals zählt

Verallgemeinerung L_p-Abstandsmaß: $dist_p = (|p_1-q_1|^p + |p_2-q_2|^p + \ldots)^{1/p}$

Veranschaulichung: L_1 vs. L_2

$$L_1: \quad d_{ii'}(1) = |x_{i1} - x_{i'1}| + |x_{i2} - x_{i'2}|$$

$$L_2: \quad d_{ii'}(2) = (|x_{i1} - x_{i'1}|^2 + |x_{i2} - x_{i'2}|^2)^{1/2}$$

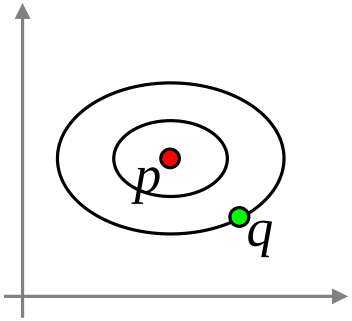


Quelle: http://www.econ.upf.edu/~michael/stanford/maeb5.pdf

Räume und Distanzfunktionen

Gewichtete Euklidische Norm:

$dist = (w_1(p_1\text{-}q_1)^2 + w_2(p_2\text{-}q_2)^2 + \ldots)^{1/2}$



Häufig sind die Wertebereiche der Merkmale deutlich unterschiedlich.

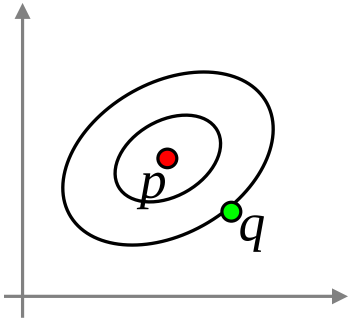
Beispiel: Merkmal $M_1 \in [0.01 .. 0.05]$

Merkmal $M_2 \in [3.1 .. 22.2]$

Damit M_1 überhaupt berücksichtigt wird, muss es höher gewichtet werden

Quadratische Form:

$dist = ((p - q)\,\mathbf{M}\,(p - q)^{\mathbf{T}})^{1/2}$



Bei den bisherigen Distanzmaßen werden die Merkmale nur getrennt gewichtet.

Besonders bei Farbhistogrammen müssen auch *verschiedene* Merkmale gemeinsam gewichtet werden.

Normalisierung

- Attribute mit großem Wertebereich gehen stärker in Distanzen ein als solche mit kleinem Wertebereich
  - z.B. Einkommen [10K-100K]; Alter [10-100]
- Skalierung von Attributen in einen einheitlichen Wertebereich, um die Beiträge aller Attribute zur Distanz gleich zu gewichten
- min-max Normalisierung zu [new_min_A , new_max_A]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - z.B. normalisiere Alter=30 in [0,1], mit min=10,max=100. new_age=(30-10)/(100-10)=2/9
- z-score Normalisierung

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

z.B. normalisiere 70000 mit μ=50000, σ=15000.
new_value = (70000-50000)/15000=1.33

Distanzen und Ähnlichkeiten

Statt mit Distanzmaßen, die die Unähnlichkeit zweier Objekte messen, arbeitet man manchmal auch mit Ähnlichkeitsmaßen:

$$\mathrm{sim}(x,y) = 0 \approx \text{unendliche Distanz}$$

häufig maximale Ähnlichkeit 1: $\mathrm{sim}(x,y) = 1 \approx \mathrm{dist}(x,y) = 0$

Abbildungen von Ähnlichkeiten auf Distanzen:

$$\mathrm{dist}(x,y) = 1 - \mathrm{sim}(x,y)$$

$$\mathrm{dist}(x,y) = -\ln(\mathrm{sim}(x,y))$$

(Un-)Ähnlichkeiten für binäre Variablen

Binär-Variable hat zwei Zustände: 0 (absence), 1 (presence)

Contingency-table für Binärdaten:

Object *i* \ Object *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

Einfacher matching coefficient

(für symmetrische Binär-Variablen)

$$d(i, j) = \frac{r+s}{q+r+s+t}$$

für asymmetrische Binär-Variablen:

$$d(i, j) = \frac{r+s}{q+r+s}$$

Jaccard coefficient

(für *asymmetrische* Binär-Variablen)

$$sim_{Jaccard}(i, j) = \frac{q}{q+r+s}$$

(Un-)Ähnlichkeiten für kategorische/nominale Attribute

- Kategorische Variable hat mehr als zwei Zustände
  - z.B. Farbe {rot, blau, grün}
- Methode 1: simple matching
  - m: # matches, p: # variables

$$d(i,j) = \frac{p-m}{p}$$

  - entspricht (skalierter) Hamming-Distanz

$$dist(x,y) = \sum_{i=1}^{d} \delta(x_i, y_i) \ mit\ \delta(x_i, y_i) = \begin{cases} 0, & falls\ x_i = y_i \\ 1, & sonst \end{cases}$$

- Methode 2: Abbildung auf binäre Variablen
  - Erzeuge eine neue binäre Variable für jeden der nominalen Zustände

    rot = (ja, nein), blau = (ja, nein), grün = (ja, nein)

ID	title	URL	unknc	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentar	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFi	Thrille
1	Toy Story (1995)	http://us.imdb	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	
2	GoldenEye (1995)	http://us.imdb	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Four Rooms (1995)	http://us.imdb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Get Shorty (1995)	http://us.imdb	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	
5	Copycat (1995)	http://us.imdb	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	
6	Shanghai Triad (Yao a yao yao dao waipo qiao	http://us.imdb	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
7	Twelve Monkeys (1995)	http://us.imdb	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	
8	Babe (1995)	http://us.imdb	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	
9	Dead Man Walking (1995)	http://us.imdb	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
10	Richard III (1995)	http://us.imdb	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
11	Seven (Se7en) (1995)	http://us.imdb	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
12	Usual Suspects, The (1995)	http://us.imdb	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	

Räume und Distanzfunktionen

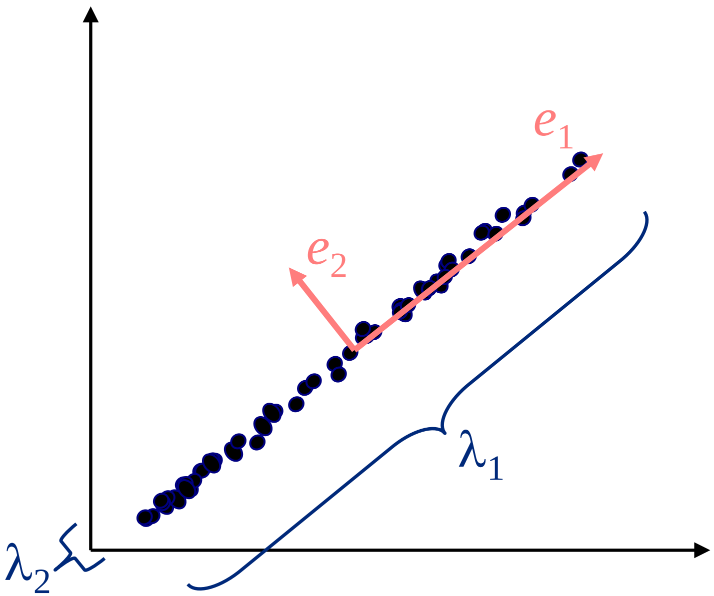
Deskription von Featurevektoren

- Gegeben: Menge *DB* von Featurevektoren
- Zentroid (Centroid, vgl. arithmetisches Mittel): $\mu_{DB} = \frac{1}{|DB|} \cdot \sum_{o \in DB} o$
  - Achtung: bei allgemeinen metrischen Räumen muss der Centroid nicht notwendigerweise existieren!!!
- Medoid m_{DB}:
  - Der Featurevektor, der am nächsten zum Centroiden gelegen ist (die kleinste Distanz zum Zentroiden hat)
  - Bei allg. metrischen Räumen: Objekt mit dem kleinsten durchschn. Abstand zu allen anderen Objekten aus *DB*
- Varianz (der Distanzen): $Var_{DB} = \frac{1}{|DB|} \cdot \sum_{o \in DB} dist(o, \mu_{DB})$
- Standardabweichung analog

Räume und Distanzfunktionen

Hauptachsenanalyse eine Menge *DB* von *Euklidischen Vektoren*

- Kovarianz-Matrix: $$\Sigma_{DB} = \frac{1}{|DB|} \sum_{o \in D} (o - \mu_{DB})(o - \mu_{DB})^{\mathrm{T}}$$
- Die Matrix wird zerlegt in
  - eine Orthonormalmatrix $V = [e_1, \ldots, e_d]$ (Eigenvektoren)
  - und eine Diagonalmatrix $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_d)$ (Eigenwerte)
  - so dass gilt: $\Sigma_{DB} = V \Lambda V^{\mathrm{T}}$
- Interpretation:
  - Eigenvektoren:
    Hauptausrichtung der Datenpunkte in *DB*
  - Eigenwerte:
    Varianz der Datenpunkte in *DB* entlang der entspr. Eigenvektoren

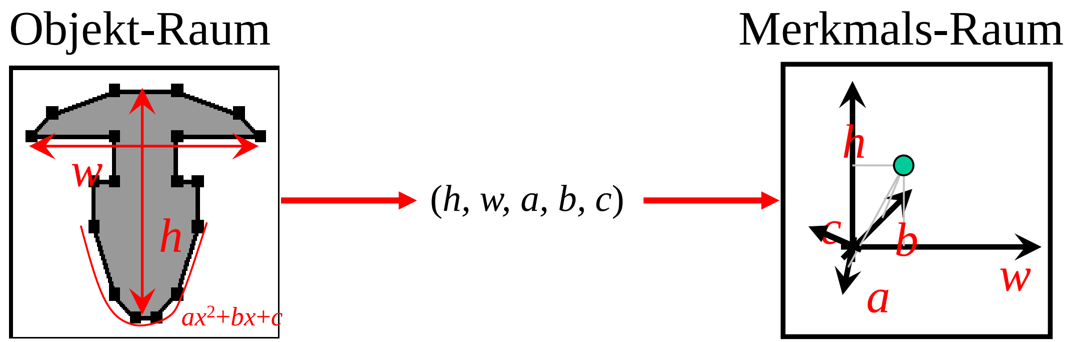


Feature-Transformationen für räumliche Objekte

Feature Transformation

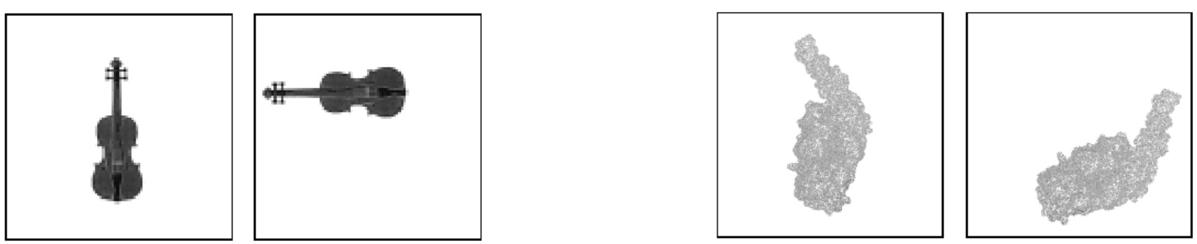
für räumliche Objekte

(CAD-Daten, Proteine, ...)



– Invarianzen

- Gleichheit (oder Ähnlichkeit) von Formen unabhängig von Lage und Orientierung im Raum
- Beispiele gleicher Formen im 2D und im 3D:

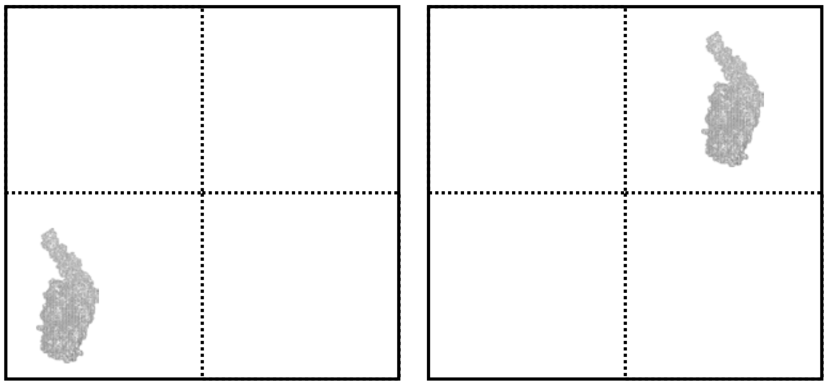


- Erwünscht:
  – Kanonische Darstellung, d.h. ohne Lage- und Orientierungsinformation
  – Verallgemeinerung auf andere Objekteigenschaften

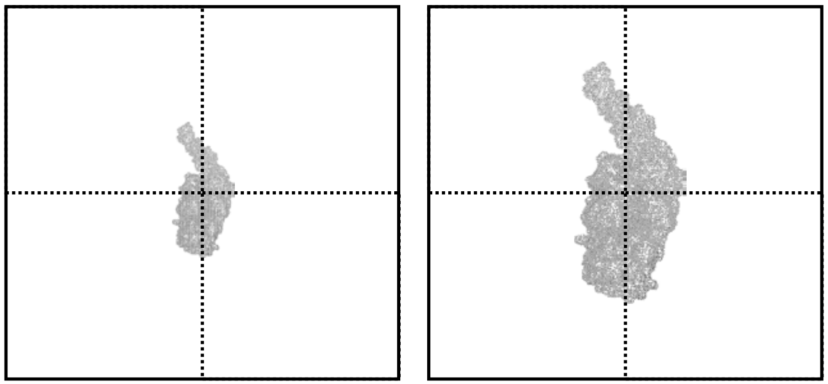
Feature-Transformationen für räumliche Objekte

Die wichtigsten Invarianzen

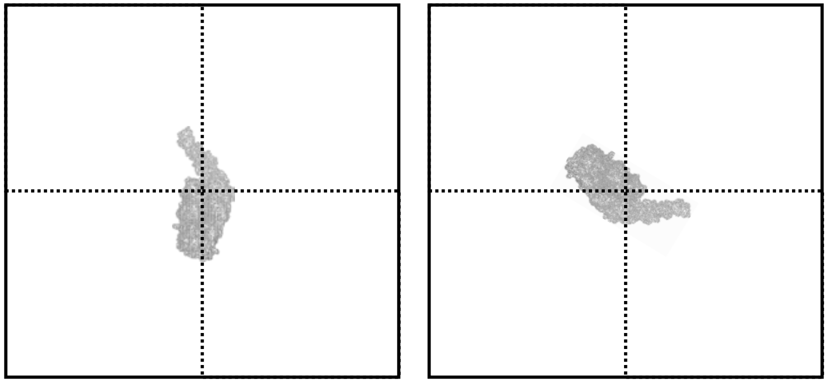
Translation



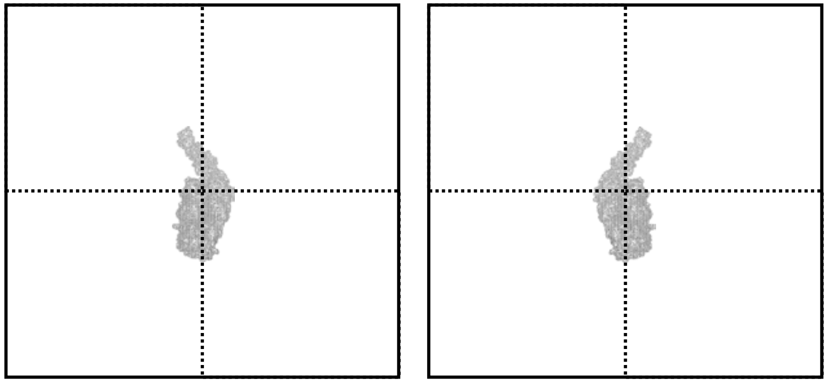
Skalierung



Rotation

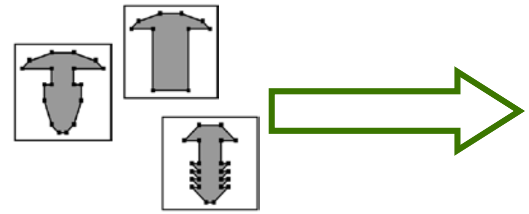


Spiegelung



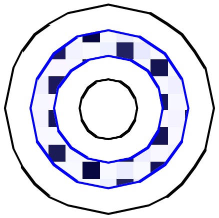
Feature-Transformationen für räumliche Objekte

Volume Model [Ankerst, Kastenmüller, Kriegel, Seidl 99]

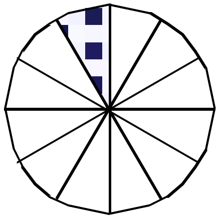


$$\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

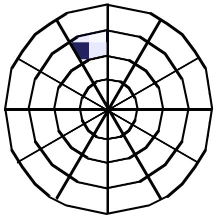
- Applikationen: CAD, Protein 3D-Strukturen
- Idee: *Formhistogramme* für 3D Objekte
  - Partitioniere den 3D-Raum in Zellen (Histogramm-Bins).
  - Bestimme den Anteil an Punkten des Objektes pro Zelle (normiertes Histogramm).
  - Durch die Normierung werden die Histogramme unabhängig von der Punktedichte.
- Partitionierungen



Schalenmodell



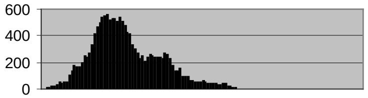
Sektorenmodell



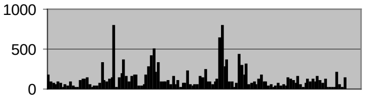
kombiniertes Modell

- Beispiel

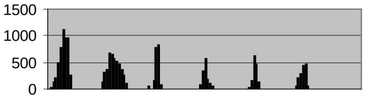
Seryl-tRNA Synthetase



Schalenmodell (120 Schalen)



Sektorenmodell (122 Sektoren)



kombiniertes Modell (20 Schalen, 6 Sektoren)

Feature-Transformationen für räumliche Objekte

- Formale Definition der Histogramme
  - *Schalenmodell*: Definiere die Bins über den Abstand zum Mittelpunkt, d.h. Anzahl der Punkte auf der jeweiligen Schale.
  - *Sektorenmodell*: Anzahl der Punkte im jeweiligen Sektor
  - *Kombiniertes Modell*: Synthese aus Schalen- und Sektorenmodell
- Invarianzen
  - Translationsinvarianz durch Lagenormierung:
    Verschiebung des Schwerpunkts eines Objektes in den Ursprung
  - Rotationsinvarianz durch Hauptachsentransformation:
    - Drehung der Objekte, so dass die Hauptachsen auf den Koordinatenachsen liegen
    - unnötig beim Schalenmodell (ist inhärent rotationsinvariant)

Feature-Transformationen für Texte

- *Text als Mengen/Vektoren von Termen: („Bag-Of-Words“)*
  - Term:
    - einzelnes Wort („Schnee“, „Eis“..)
      oder
    - zusammenhängendes Satzfragment („nicht mehr vorwärts“..)

  - Transformation eines Dokuments D in Vektor $r(D) = (h_1, \ldots, h_d)$
    $h_i \geq 0$: die Häufigkeit des Terms t_i in D

Schnee und Eis haben die Straßen in weiten Teilen Deutschlands in Rutschbahnen verwandelt. Lastwagen gerieten ins Schleudern, zahlreiche Fahrzeuge kamen an Steigungen nicht mehr vorwärts. Die Streudienste waren im Dauereinsatz…

⇒

…	…
Schnee	1
Eis	1
Fahrzeug	1
Politik	0
…	…

h_{Eis}

Feature-Transformationen für Texte

- Probleme im Textmining
  1. Viele Wörter nutzlos (z.B. er, sie, es, und, als, der, dies, das...)
  2. Wörter haben gleichen Wortstamm („gehen“ „ging“)
  3. Sehr hochdimensionale Featureräume (häufig $d > 10.000$)
  4. Nicht alle Terme sind gleich wertvoll
  5. Die meisten Termhäufigkeiten $h_i = 0$ („sparse feature space“)
- weitere Probleme aus der Linguistik:
  - unterschiedliche Wörter haben gleiche Bedeutung
    „laufen“ ⇔ „rennen“
  - Wörter haben mehrere Bedeutungen
    „Maus“: Computermaus, Nagetier…

Feature-Transformationen für Texte

- Problem 1: Viele Wörter nutzlos (z.B. er, sie, es, und, als, der, dies, das...)
  - Lösung: Streichen solcher Terme (Stopwords)

    Für alle Sprachen werden Stopwordlisten im WWW publiziert.

- Problem 2: Wörter haben gleichen Wortstamm („gehen“ „ging“)
  - Lösung: Stemming

    Worte auf Wortstamm rückführen (z.B. lief, läuft, lauft => laufen)

    Im Englischen algorithmisches Stemming möglich.
    (Porters Stemming Algorithms: http://tartarus.org/~martin/PorterStemmer/index.html)

    In anderen Sprachen werden Dictionaries benötigt, die die Wortstämme zu den Vokabeln enthalten.

Feature-Transformationen für Texte

- Problem 3: Sehr viele Terme müssen betrachtet werden.
  - Lösung: Auswahl der wichtigsten Features („Feature Selection“)
  - Beispiel: Mittlere Dokumentenhäufigkeit
    - Sehr häufige Terme kommen scheinbar in allen Dokumenten vor
      => Vorkommen unterscheidet kaum Dokumente
    - Sehr seltene Terme kommen nur in Bruchteil der Dokumente vor
      => Nichtvorkommen unterscheidet kaum Dokumente

Vorgehen:

1. Berechne Dokumenthäufigkeit für alle Terme t_i : $DF(t_i) = \frac{|Dok\_t_i|}{|ALL\_Doks|}$
2. Sortiere Terme nach DF(t_i) und vergebe Rang rank(t_i)
3. Sortiere Terme nach score(t_i)= DF(t_i) · rank(t_i)
   z.B. score(t_{23}) = 0.82 · 1 = 0.82
   score(t_{17}) = 0.75 · 2 = 1.5
4. Wähle die k Terme mit dem größten Wert für score(t_i)

Rank	Term	DF
1.	t_{23}	0.82
2.	t_{17}	0.65
3.	t_{14}	0.52
4.	…	…

Feature-Transformationen für Texte

- Problem 4: Nicht alle Terme sind gleich wertvoll.
  - Idee:
    1. Gewichte seltene Terme höher als häufige.
    2. Gewichte häufig in einem Dokument auftretende Terme höher als solche die nur einmal vorkommen.
  - Lösung: TF-IDF (Term Frequency · Inverse Document Frequency)

    Berücksichtige sowohl die relative Anzahl der Vorkommen im Dokument als auch die Seltenheit des Terms.

$$TF(t,d) = \frac{n(t,d)}{\sum_{w \in d} n(w,d)}$$

relative Häufigkeit von t in d

$$IDF(t) = \frac{|DB|}{\left|\{d \mid d \in DB \wedge t \in d\}\right|}$$

inverse Häufigkeit von t bzgl. aller Dokumente

Featurevektor mit TF IDF : $r(d) = (TF(t_1, d) \cdot IDF(t_1), \ldots, TF(t_n, d) \cdot IDF(t_n))$

Feature-Transformationen für Texte

- Problem 5: die meisten Termhäufigkeiten $h_i = 0$

  => Euklidische Abstände sehr ähnlich

  – Lösung: Verwendung anderer Abstandsmaße

    Idee: Verwende Terme, die beide Dokumente (D_1, D_2) gemeinsam haben.

    Jaccard Coefficient: Dokumente als Termmengen

$$d_{Jaccard}(D_1, D_2) = 1 - \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

Cosinus Coefficient: Abstand für Wortvektoren (evtl. TF IDF)

$$d_{\text{cosinus}}(D_1, D_2) = 1 - \frac{\langle D_1, D_2 \rangle}{\|D_1\| \cdot \|D_2\|} = 1 - \frac{\sum_{i=0}^{n} \left(d_{1,i} \cdot d_{2,i}\right)}{\sqrt{\sum_{i=0}^{n} d_{1,i}^{\;2}} \cdot \sqrt{\sum_{i=0}^{n} d_{2,i}^{\;2}}}$$

Feature-Transformationen für Bilder

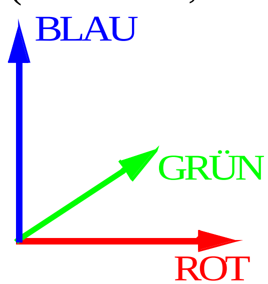
- Hauptkategorien von Features für Bilder
  - Farbverteilung (Farbhistogramme)
  - Textur (Oberflächen-Beschaffenheit von Bildsegmenten, z.B. Holzmaserung, Kieselsteine, Karomuster)
  - Formen (Konturen)

- Farbhistogramme:
  - Repräsentation der Farbverteilung in einem Bild (auf Pixelbasis)
  - Definition der Farbhistogramme
    - Farbraum festlegen (z.B. RGB, HSV, HLS, ...)
    - Menge von Repräsentanten im Farbraum auswählen (sample points), z.B. Gitter im Farbraum mit $4 \times 4 \times 4 = 64$ Farben oder $8 \times 8 \times 8 = 512$ Farben

Farbhistogramme

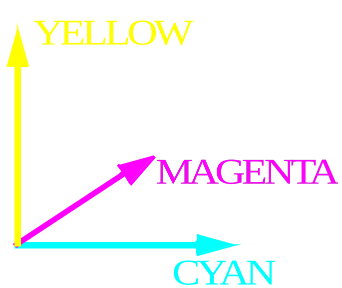
- Farbräume: Technische Modelle (RGB, CMY) und anschauliche Modelle (HSV, HLS)

RGB-Modell
(Bildschirm, additiv)

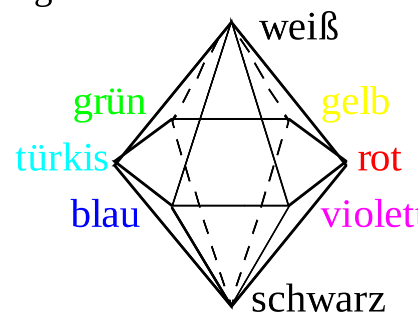


HSV-Modell: Hue, Saturation, Value
(Farbton, Sättigung, Helligkeit)

CMY-Modell
(Drucker, subtraktiv)

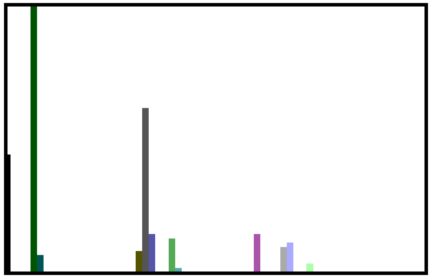


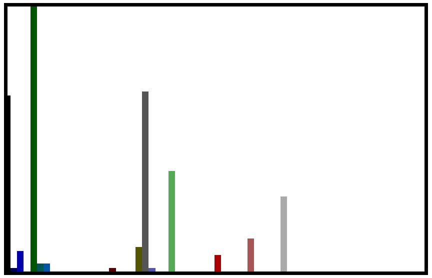
HLS-Modell: Hue, Luminance, Saturation
(Farbton, Leuchtkraft, Sättigung



Farbhistogramme

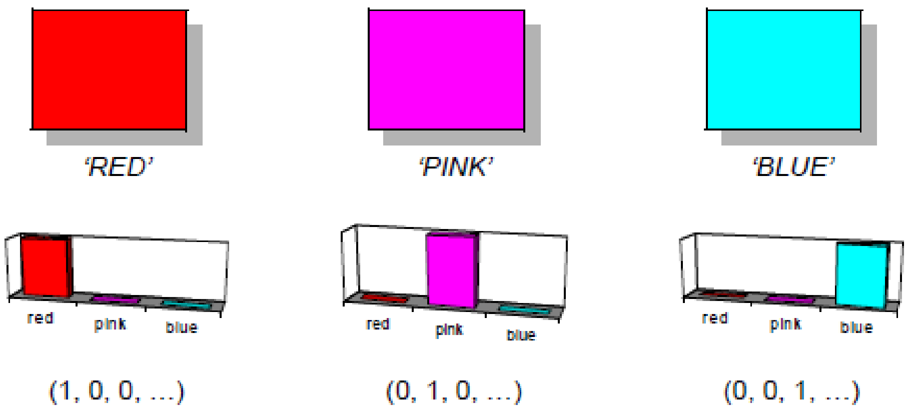
- Berechnung der Farbhistogramme
  - Für jedes Pixel, erhöhe den Zähler des nächstgelegenen Repräsentanten um eins
  - Evtl. Normierung, um Histogramm von der Bildgröße unabhängig zu machen
  - Beispiel (64 Repräsentanten):



Distanzen für Farbhistogramme

- Beispiel: euklidische Distanz für Farbhistogramme h_P und h_Q der Bilder P und Q: $dist(P,Q) = \sqrt{(h_P - h_Q) \cdot (h_P - h_Q)^T}$



$$dist('RED','PINK') = \sqrt{2}$$

$$dist('RED','BLUE') = \sqrt{2}$$

$$dist('PINK','BLUE') = \sqrt{2}$$

- Alle Paare von Bildern haben denselben Abstandswert $\sqrt{2}$
- Distanz berücksichtigt nicht, dass rot (subjektiv) ähnlicher zu lila ist als zu blau.

Distanzen für Farbhistogramme

- Quadratische Form mit Ähnlichkeitsmatrix:

$$dist_A(P,Q) = \sqrt{(h_P - h_Q) \cdot A \cdot (h_P - h_Q)^T}$$
$$= \sqrt{\sum_i \sum_j a_{ij} \cdot (h_{P_i} - h_{Q_i}) \cdot (h_{P_j} - h_{Q_j})}$$

$$A = \begin{bmatrix} 1 & a_{21} & \cdots & \\ a_{12} & 1 & a_{ij} & \vdots \\ \vdots & & 1 & \\ & \cdots & & 1 \end{bmatrix}$$

- Einträge a_{ij} ($= a_{ji}$?) beschreiben die Ähnlichkeit der Dimensionen i und j in den Vektoren (Bins i und j in den Histogrammen)

$$A' = \begin{bmatrix} 1 & 0{,}9 & 0 \\ 0{,}9 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

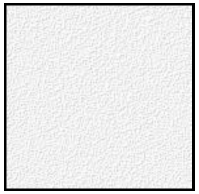
$$dist_{A'}('RED', 'PINK') = \sqrt{0{,}2}$$
$$dist_{A'}('RED', 'BLUE') = \sqrt{2}$$
$$dist_{A'}('PINK', 'BLUE') = \sqrt{2}$$

- Ähnlichkeitsmatrizen werden aus Ergebnissen der Perzeptionsforschung abgeleitet

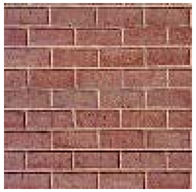
Texturen für Bilder

- Gerichtetheit, Orientiertheit (Directionality)
  - Vorhandensein von Vorzugsrichtungen
    (Verteilung der Gradientenrichtungen)


- Kontrast
  - Lebendigkeit (Unruhe) eines Musters
  - Berechnung aus Varianz im Grauwerthistogramm




- Granularität (Coarseness)
  - Größenordnung der Textur
  - Berechnung durch über das Bild verschobene Fenster unterschiedlicher Größe




Toolbox für Feature-Extraktion von Bildern:
http://code.google.com/p/jfeaturelib/

Was haben Sie gelernt?

- Objekte und Merkmale
- Arten von Merkmalen: binär, kategorisch/nominal, ordinal, numerisch
- grundlegende univariate Deskriptoren
- grundlegende bivariate Deskriptoren
- Feature-Räume / metrische Räume
- Distanzfunktionen für numerische Daten
- Distanzfunktionen für Binär-Daten
- Distanzfunktionen für kategorische Daten
- einige Feature-Transformationen und Distanz-Maße für
  - räumliche Objekte
  - Text (Dokumente)
  - Bilder