DATABASE SYSTEMS GROUP

Ludwig-Maximilians-Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

LMU

Lecture notes

Knowledge Discovery in Databases

Summer Semester 2012

Lecture 2: Data Preprocessing / Feature spaces

Lecture: Dr. Eirini Ntoutsi
Exercises: Erich Schubert

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

Outline

- Data preprocessing
- Features
- Basic data descriptors
- Feature spaces and distance functions
- Feature transformation for text data
- Things you should know
- Homework/tutorial

Why data preprocessing?

- Real world data are noisy, incomplete and inconsistent:
  - Noisy: errors/ outliers
    - erroneous values : e.g. salary = -10K
    - unexpected values: e.g. salary=100K when the rest dataset lies in [30K-50K]
  - Incomplete: missing data
    - missing attributes of interest: e.g. no information on occupation
    - missing values: e.g., occupation=“ ”
  - Inconsistent: discrepancies in the data
    - e.g. student ratings between different universities might differ
- “Dirty” data → poor mining results
- Data preprocessing is necessary for improving the quality of the mining results !!!

Major tasks in data preprocessing

- Data cleaning:
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration:
  - Integration of multiple databases, data cubes, or files (Entity identification, Value resolution)
- Data transformation:
  - e.g. normalization in a given range
  - Generalization through e.g. some concept hierarchy
- Data reduction:
  - Aggregation
  - Dimensionality reduction
  - Duplicate elimination

Outline

- Data preprocessing
- Features
- Basic data descriptors
- Feature spaces and distance functions
- Feature transformation for text data
- Things you should know
- Homework/tutorial

Data objects/ Examples/Instances

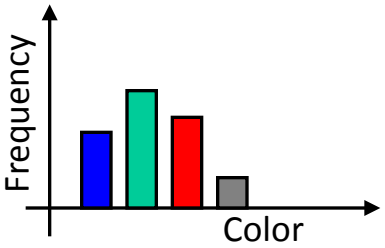
- Datasets consists of objects /examples / instances
  - e.g., in a movie database: movies, actors, director,…
  - e.g., in a library database: books, users, loans, publishers, ….
  - e.g., in a university database: students, professors, courses, grades,…
- Objects are described through features/ attributes/ variables
  - E.g. in a database table, the rows are the objects, the columns are the attributes/features/variables.

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

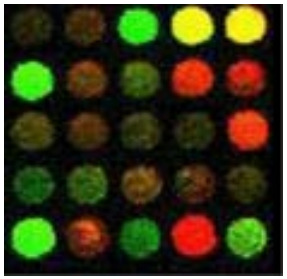
ID	title	URL	unknc	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentar	Drama	Fantasy	FilmNoir	Horror	Musical	Mystery	Romance	SciFi	Thrille
1	Toy Story (1995)	http://us.imdb	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	
2	GoldenEye (1995)	http://us.imdb	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Four Rooms (1995)	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	

Features

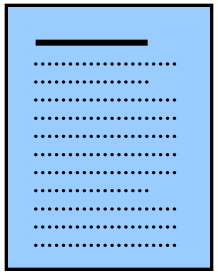
Image databases:
Color histograms



Gene databases:
gene expression level



Text databases:
Word frequencies



Data	25
Mining	15
Feature	12
Object	7
...	

The feature-approach allows uniform treatment of objects of different classes of applications.

Basic feature types

- **Binary/ Dichotomous variables:** the attribute can take 2 values 0 or 1
  - usually, 0 means absence, 1 means presence
  - Symmetric binary: both outcomes equally important:
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Categorical/ Nominal variables:** the attribute can take values within a set of *M* categories/ states.
  - e.g., eye color = {brown, green, blue}, occupation = {engineer, doctor, teacher}
  - There is *no ordering* in the categories/ states.
  - Only *equal*/ *different* relationships apply.
- **Ordinal variables:** similar to categorical variables, but the *M* states are ordered/ ranked in a meaningful way.
  - e.g., movie rating = {dislike, indifferent, like}, medal = {bronze, silver, gold}
  - There is an *ordering* between the values. But, we don't know the magnitude between successive values. The distance between these values has no meaning.

Basic feature types

Numeric variables: quantitative (integer or real)

- **Interval-scale variables**
  - continuous measurements of a roughly linear scale
    - e.g., weight, height, latitude, longitude, temperature C^o, citation counts
  - Measured on a scale of equal-sized units
    - It is assumed that the intervals keep the same importance throughout the scale.
  - Except for ordering, we can quantify and compare the magnitudes of differences between them.
    - e.g. 40°C > 30°C, 40°C is twice as high as 20°C, and an increase from 20°C to 40°C is twice as much as the increase from 30°C to 40°C.
- **Ratio-scale variables**
  - continuous positive measurements on a nonlinear scale, e.g. in an exponential scale.
    - e.g., the growth of bacterial population (say, with a growth function Ae^{Bt}.). In this model, equal time intervals multiply the population by the same ratio.
    - e.g., the decay of a radioactive element

Outline

- Data preprocessing
- Features
- Basic data descriptors
- Feature spaces and distance functions
- Feature transformation for text data
- Things you should know
- Homework/tutorial

Univariate descriptors: central tendency

Let $x_1,\ldots,x_n$ be a random sample of an attribute X. Measures of central tendency of X include:

- (Arithmetic) mean/ center/ average: ⟵ Algebraic measure

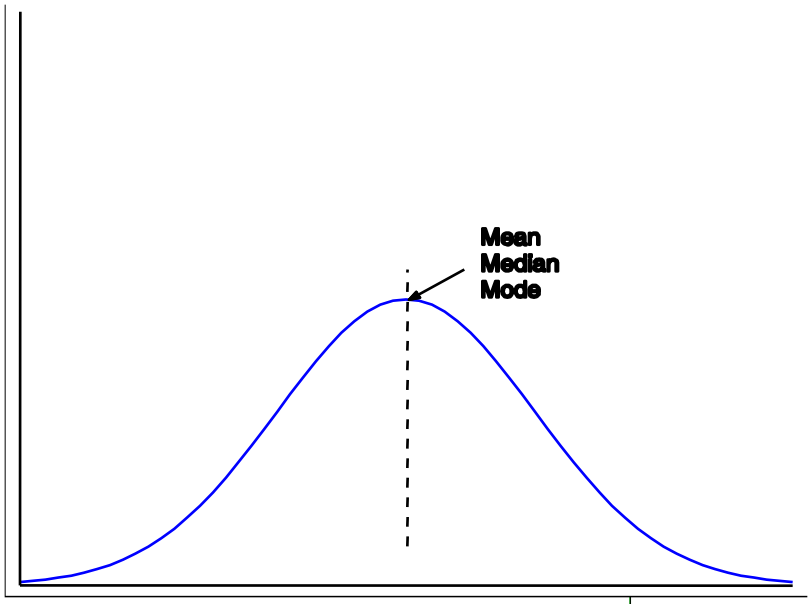
$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- Weighted average: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

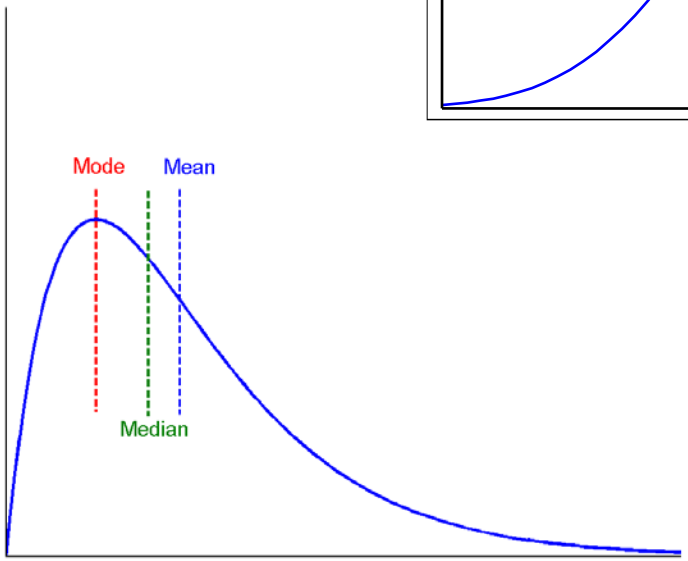
- Median: the central element in ascending ordering ⟵ Holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise

- Mode: Value that occurs most frequently in the data ⟵ Holistic measure
  - Unimodal, bimodal, trimodal

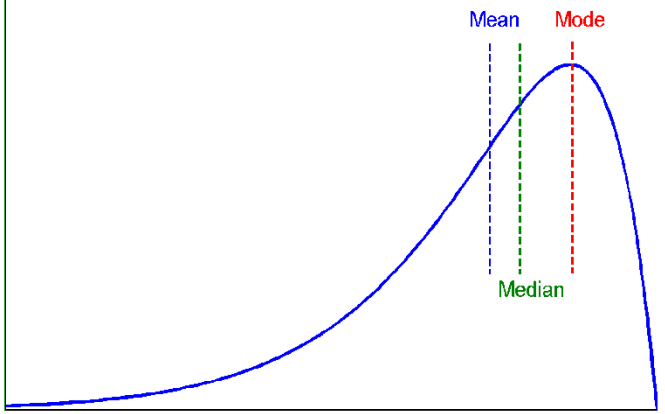
Symmetric vs Skewed data



Symmetric



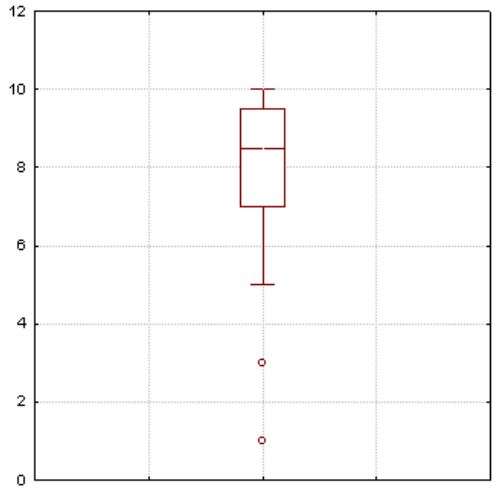
Positively skewed



Negatively skewed

Univariate descriptors: dispersion

Let $x_1,\ldots,x_n$ be a random sample of an attribute X. The degree to which X values tend to spread is called dispersion or variance of X :



Source: http://de.wikipedia.org/wiki/Boxplot

- Range: max value – min value
- Q_1 (25th percentile), Q_3 (75th percentile)
  - Median is the 50th percentile
- 5 number summary: min, Q_1, median, Q_3, max
  - Boxplots to visualize them
- Variance σ^2:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

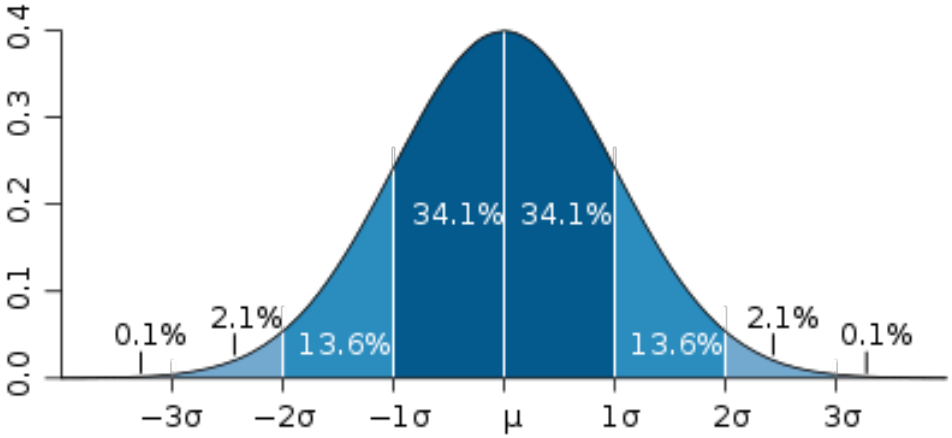
- Standard deviation σ:

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

Example: Normal distribution

The normal distribution curve

- ~68% of values drawn from a normal distribution are from μ-σ to μ+σ
- ~95% of the values lie from μ-2σ to μ+2σ
- ~99.7% of the values are from μ-3σ to μ+3σ



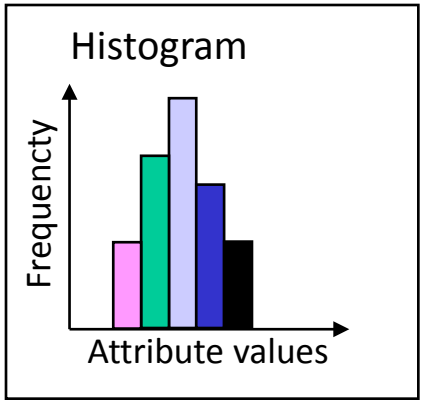
Source: http://en.wikipedia.org/wiki/Normal_distribution

Univariate descriptors: graphic displays

Let $x_1,\ldots,x_n$ be a random sample of an attribute X.

For the visual inspection of the data, several types of graphs are useful, e.g.:

- Boxplots
  - 5 number summary
- Histograms:
  - Summarizes the distribution of X
  - X axis: attribute values, Y axis: frequencies
  - Absolute frequency: for each value a, # occurrences of a in the sample
  - Relative frequency: f(a) = h(a)/n
- Different types of histograms, e.g.:
  - Equal width:
    - It divides the range into *N* intervals of equal size
  - Equal frequency/ depth:
    - It divides the range into *N* intervals, each containing approximately same number of samples



Bivariate descriptors

- Given two attributes X, Y one can measure how strongly they are correlated
- For numerical data → correlation coefficient
- For categorical data → χ^2 (chi-square)

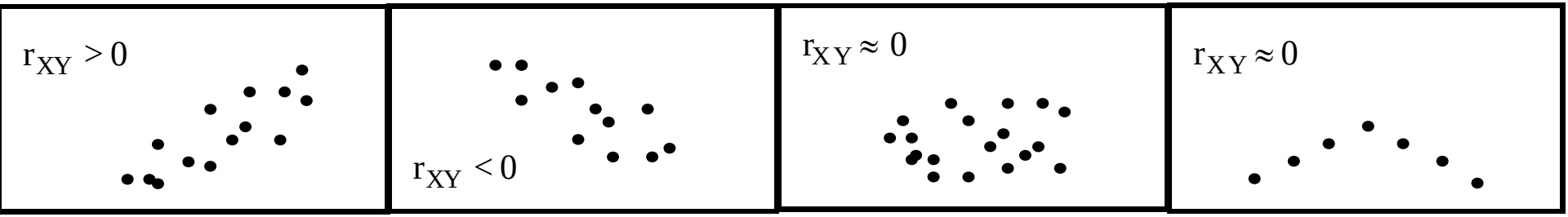
Bivariate descriptors for numerical features

Correlation coefficient (also called Pearson's product moment coefficient) :

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

– n: # tuples; x_i, y_i: the values in the i^{th} tuple for X, Y

- $-1 \leq r_{XY} \leq 1$
- the higher r_{XY} the stronger the correlation
  - $r_{XY} > 0$ positive correlation
  - $r_{XY} < 0$ negative correlation
  - $r_{XY} \sim 0$ no correlation/ independent



Visual inspection of correlation

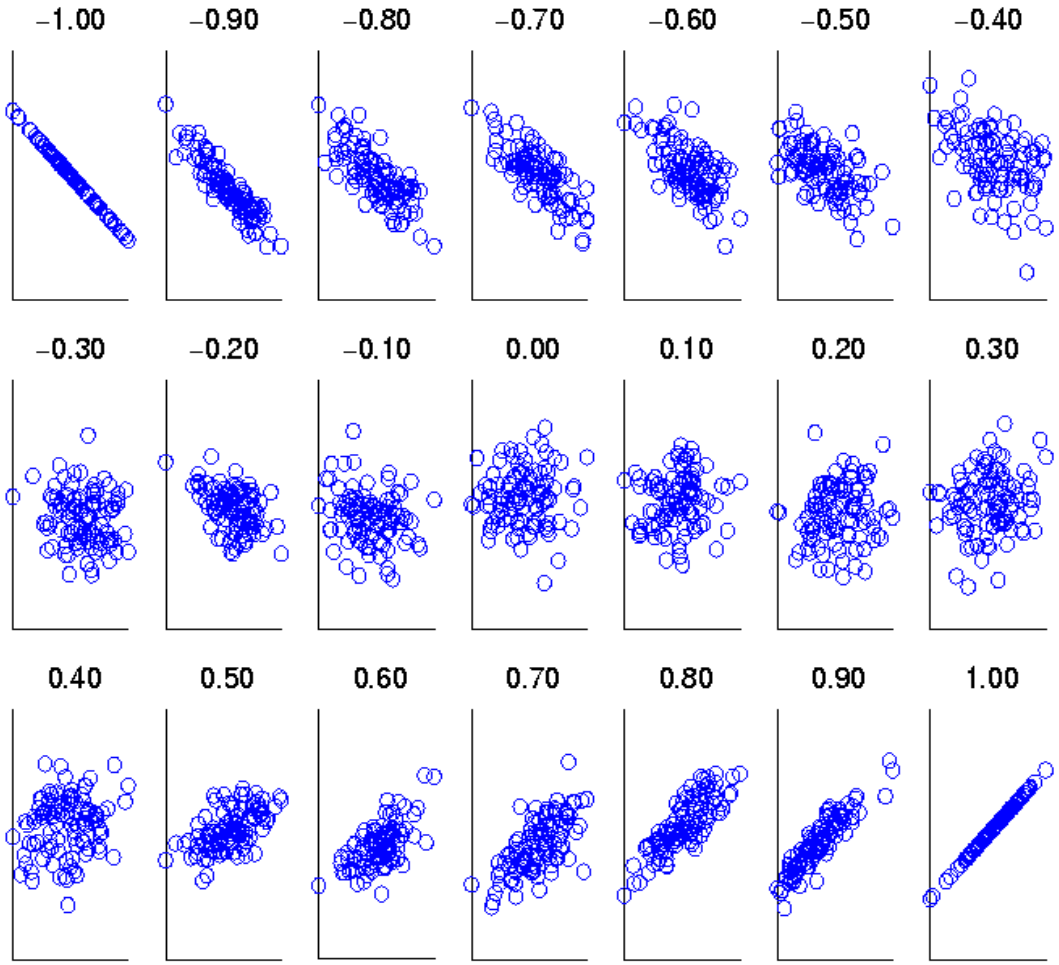


Figure 5.11. Scatter plots illustrating correlations from -1 to 1.

Bivariate descriptors for categorical features

Contingency table

- For categorical/ nominal features X={x_1, …, x_c }, Y={y_1, …, y_r}
- Represents the absolute frequency h_{ij} of each combination of values (x_i, y_J) and the marginal frequencies h_i, h_j of X, Y.

Attribute Y

Attribute X	Medium-term unemployment	Long-term unemployment	Total
No education	19	18	37
Teaching	43	20	63
Total	62	38	100

Chi-square χ^2 test

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

o_{ij}:observed frequency
e_{ij}: expected frequency

$$e_{ij} = \frac{h_i h_j}{n}$$

Chi-square example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Outline

- Data preprocessing
- Features
- Basic data descriptors
- Feature spaces and distance functions
- Feature transformation for text data
- Things you should know
- Homework/tutorial

Feature spaces and distance functions

Feature space

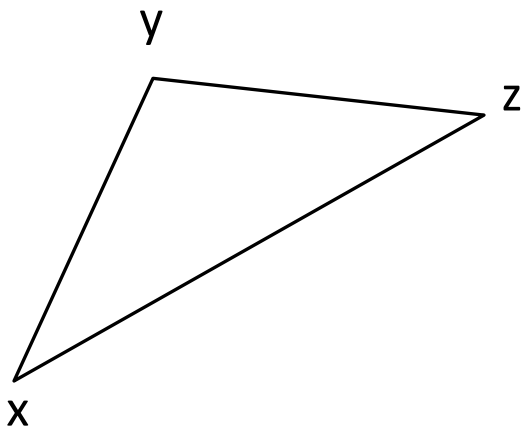
- Intuitively: a domain with a distance function
- Formally: feature space **F** = (*Dom*, *dist*)
- *Dom* is a set of attributes / features
- $dist : Dom \times Dom \rightarrow R^+_0$
- For all *x*, *y* in *Dom*, $x \neq y$, dist is required to satisfy the following properties:
  - $dist(x,y) > 0$ (strictness)
  - $dist(x,x) = 0$ (reflexivity)
  - $dist(x,y) = dist(y,x)$ (symmetry)

Feature spaces and distance functions

Metric space

- Formally: Metric space **M** = (*Dom*, *dist*) with the following property:
  - **M** is a feature space
  - $\forall x,y,z \in Dom : dist(x,z) \leq dist(x,y) + dist(y,z)$ (triangle inequality)

- Famous example: Euclidean vector space
  - Formally: Euclidean vector space **E** = (*Dom*, *dist*) :
    - (*Dom*, *dist*) is a metric space
    - *Dom* = $\mathbb{R}^d$



- Sprechweise:
  - Euclidean vector space = “Feature space”
  - Vectors (Objects in the Euclidean feature space) = “Feature vectors”
  - The *d* dimensions of the vector space = “Features”

Feature spaces and distance functions

Similarity/ distance between feature vectors (Euclidean vectors)

- Manhattan-Norm (L_1): $dist_2 = |p_1-q_1|+|p_2-q_2|+...$

The sum of the absolute differences of their coordinates

- Euclidean Norm (L_2): $dist_1 = ((p_1-q_1)^2+(p_2-q_2)^2+...)^{1/2}$

The length of the line segment connecting p and q

- Maximums-Norm ($L_¥$): $dist_¥ = \max\{|p_1-q_1|, |p_2-q_2|,...\}$

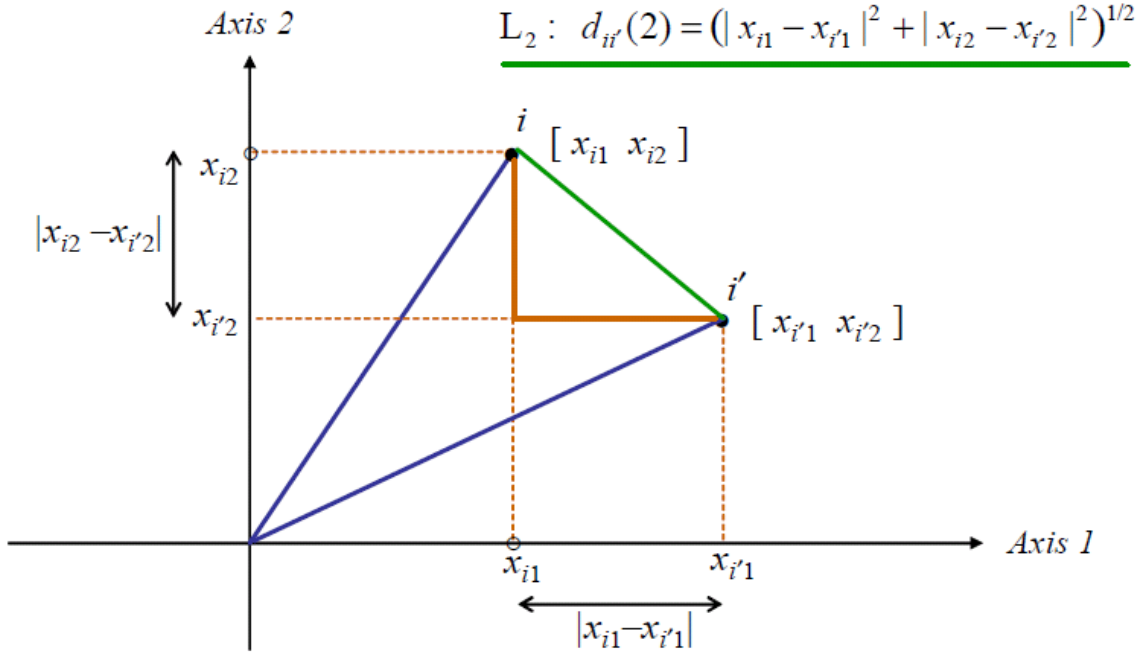
The distance between the least similar coordinates counts.

- Generalization of L_p-distance: $dist_p = (|p_1-q_1|^p + |p_2-q_2|^p + ...)^{1/p}$

Feature spaces and distance functions

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

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



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

Normalization

- Attributes with large ranges outweigh ones with small ranges
  - e.g. income [10K-100K]; age [10-100]
- To balance the "contribution" of each attribute in the resulting distance, the attributes are scaled to fall within a small, specified range
- min-max normalization: to [new_min$_A$, new_max$_A$]

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

  - e.g. normalize age=30 in [0-1], when min=10,max=100. new_age=(30-10)/(100-10)=2/9
- z-score normalization

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

e.g. normalize 70,000 iff μ=50,000, σ=15,000.
new_value = (70,000-50,000)/15,000=1.33

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

j is the smallest integer such that Max(|v'|) < 1

e.g., if v lies in [1, 999], if we set j=3 then v' will lie in [0.001, 0,999]

Dissimilarity between binary attributes

- A binary attribute has two states: 0 (absence), 1 (presence)
- A contingency table for binary data

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

- Simple matching coefficient

  (for symmetric binary variables)

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

- for asymmetric binary variables:

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

- Jaccard coefficient

  (for *asymmetric* binary variables)

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

Example

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	1	0	1	0	0	0
Mary	1	0	1	0	1	0
Jim	1	1	0	0	0	0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Dissimilarity measures between categorical/ nominal attributes

- A categorical/ nominal attribute has >2 states (generalization of a binary attribute)
  - e.g. color={red, blue, green}
- Method 1: Simple matching
  - m: # of matches, p: total # of variables

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

- Method 2: Map it to binary variables
  - create a new binary attribute for each of the *M* nominal states of the attribute

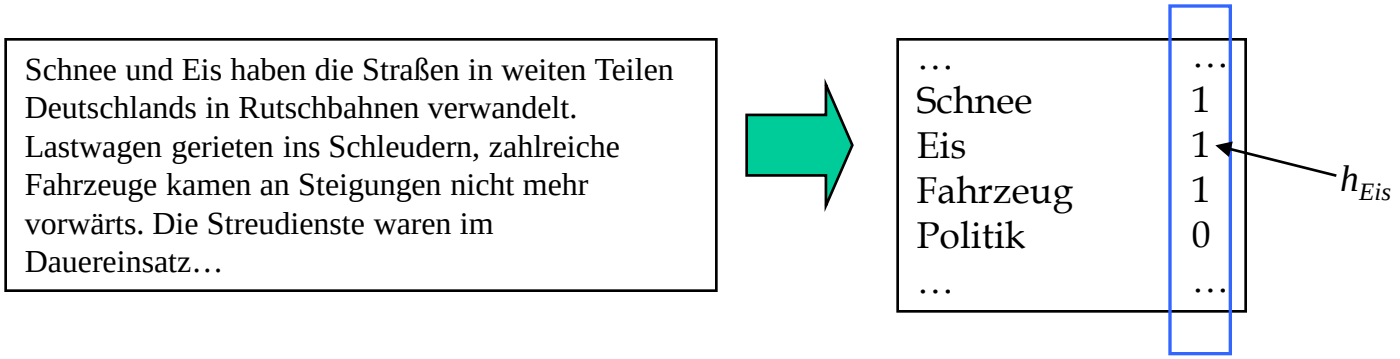
Outline

- Data preprocessing
- Features
- Basic data descriptors
- Feature spaces and distance functions
- Feature transformation for text data
- Things you should know
- Homework/tutorial

Feature transformations for text data

- *Text represented as set of terms ("Bag-Of-Words" model)*
  - Term:
    - Single words ("cluster", "analysis"..)
      or
    - bigrams, trigrams, ...n-grams ("cluster analysis"..)
  - Transformation of a document D in a vector $r(D) = (h_1, ..., h_d)$

$h_i \geq 0$: the frequency of term t_i in D



Feature transformations for text data

- Challenges in Text Mining:
  1. Common words (“e.g.”, “the”, “and”, “for”, “me”)
  2. Words with the same root (“fish”, “fisher”, “fishing”,…)
  3. Very high-dimensional space (dimensionality $d > 10.000$)
  4. Not all terms are equally important
  5. Most term frequencies $h_i = 0$ (“sparse feature space“)
- More challenges due to language:
  - Different words have same meaning (synonyms)
    - “freedom” – “liberty”
  - Words have more than one meanings
    - e.g. “java”, “mouse”

Feature transformations for text data

- Problem 1: Common words (“e.g.”, “the”, “and”, “for”, “me”)
  - Solution: ignore these terms (Stopwords)

    There are stopwords list for all languages in WWW.

- Problem 2: Words with the same root (“fish”, “fisher”, “fishing”,…)
  - Solution: Stemming

    Map the words to their root

    - "fishing", "fished", "fish", and "fisher" to the root word, "fish".

    For English, the Porter stemmer is widely used.
    (Porters Stemming Algorithms: http://tartarus.org/~martin/PorterStemmer/index.html)

    Stemming solutions exist for other languages also.

    The root of the words is the output of stemming.

Feature transformations for text data

- Problem 3: Too many features/ terms
  - Solution: Select the most important features ("Feature Selection")
  - Example: average document frequency for a term
    - Very frequent items appear in almost all documents
    - Very rare terms appear in only a few documents

Ranking procedure:

1. Compute document frequency for all terms t_i :
2. Sort terms w.r.t. DF(t_i) and get rank(t_i)
3. Sort terms by score(t_i)= DF(t_i) · rank(t_i)
   e.g. score(t_{23}) = 0.82 · 1 = 0.82
   score(t_{17}) = 0.75 · 2 = 1.5
4. Select the k terms with the larger score(t_i)

$$DF(t_i) = \frac{|Dok\_t_i|}{|ALL\_Doks|}$$

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

Feature transformations for text data

- Problem 4: Not all terms are equally important
  - Idea: Very frequent terms are less informative than less frequent words. Define such a term weighting schema.
  - Solution: TF-IDF (Term Frequency · Inverse Document Frequency)

    Consider both the importance of the term in the document and in the whole DB.

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

The frequency of term t in d

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

Inverse frequency of term t in all DB

$$TF \times IDF = TF(t,d)\,IDF(t)$$

Feature vector with TF IDF : $r(d) = (TF(t_1,d) \cdot IDF(t_1), \ldots, TF(t_n,d) \cdot IDF(t_n))$

Feature transformations for text data

- Problem 5: for most of the terms $h_i = 0$
  - Euclidean distance is not a good idea: it is influenced by vectors lengths
  - Idea: use more appropriate distance measures

Jaccard Coefficient: Consider common terms

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

Cosine Coefficient: Consider terms values (e.g. TFIDF)

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

Outline

- Data preprocessing
- Features
- Basic data descriptors
- Feature spaces and distance functions
- Feature transformation for text data
- Things you should know
- Homework/tutorial

Things you should know

- Feature types: binary, categorical/nominal, ordinal, numeric
- Basic univariate descriptors
- Basic bivariate descriptors
- Feature spaces/ metric spaces
- Distance functions for numeric data
- Distance functions for binary data
- Distance functions for categorical data
- Text transformation

Homework/ Tutorial

- **Tutorial:** 1st tutorial on Thursday on:
  - basic data mining tasks
  - feature types

- **Homework:** Have a look at the tutorial in the website and try to solve the exercises.
  - Bring your questions ….

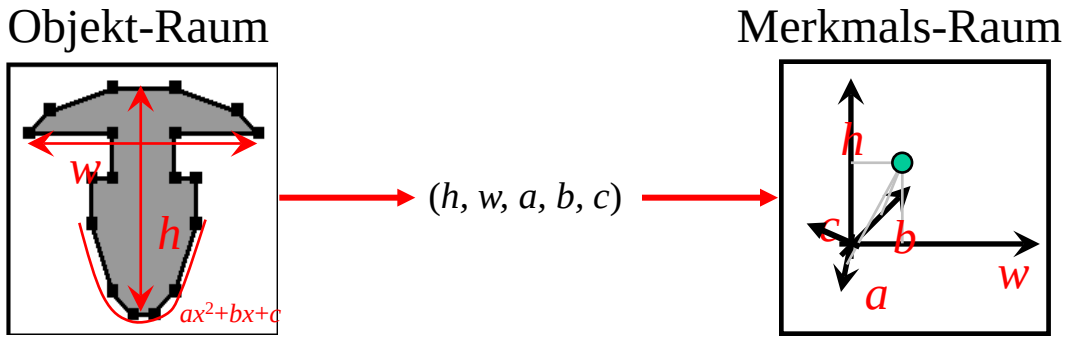
- **Suggested reading:**
  - Han J., KamberM., Pei J.*Data Mining: Concepts and Techniques3rd ed., Morgan Kaufmann, 2011* (Chapter 2) (Section 7.2)
  - Tutorial on Text Mining and Link Analysis for Web and Semantic Web
    - Video: http://videolectures.net/ess07_grobelnik_twdml/
    - Slides are also there (Download slides)
  - Natural language processing course: https://www.coursera.org/course/nlp

Feature-Transformationen für räumliche Objekte

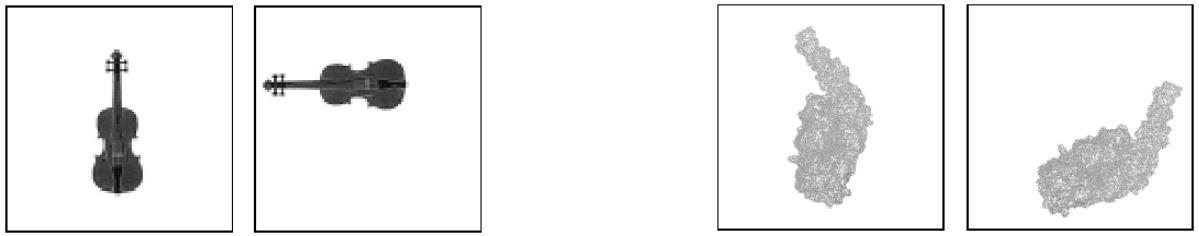
Feature Transformation

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:

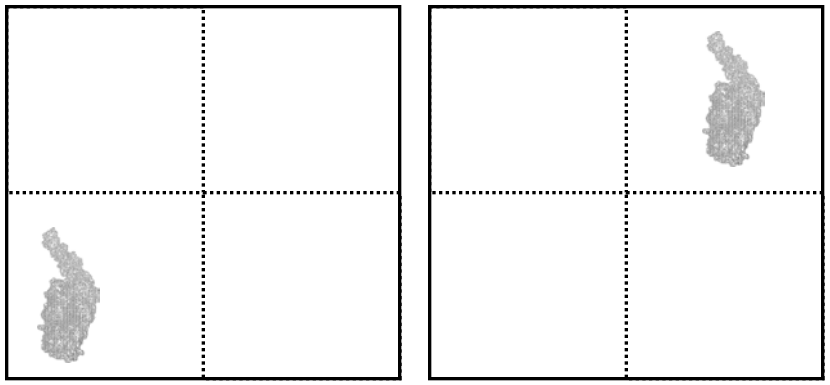


  - Erwünscht:
    - Kanonische Darstellung, d.h. ohne Lage- und Orientierungsinformation
    - Verallgemeinerung auf andere Objekteigenschaften

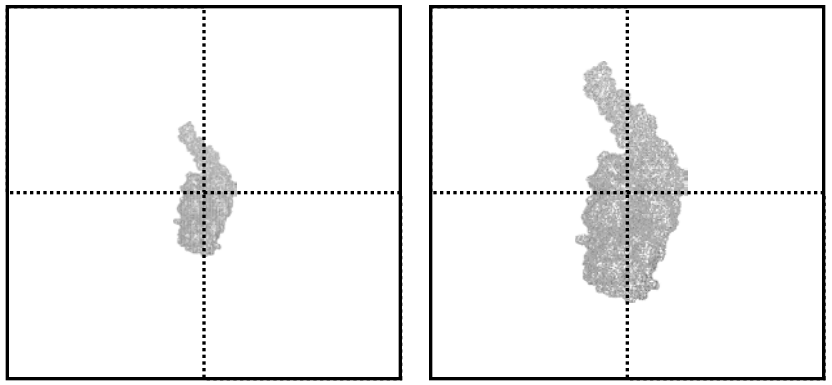
Feature-Transformationen für räumliche Objekte

Die wichtigsten Invarianzen

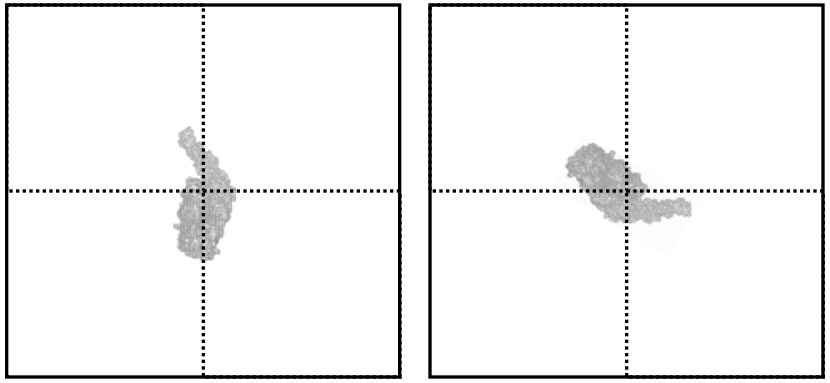
Translation



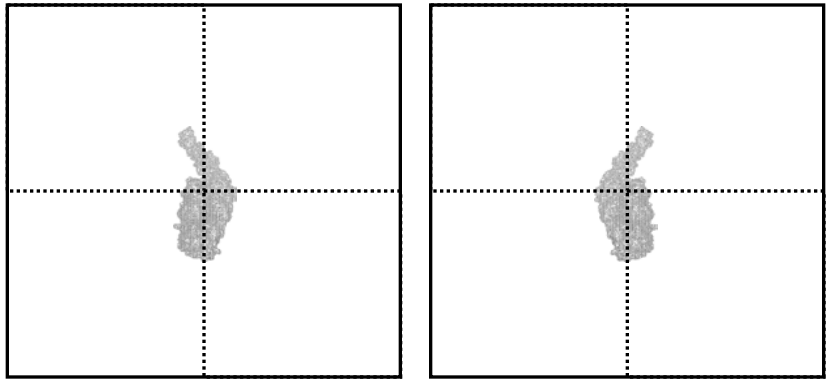
Skalierung



Rotation

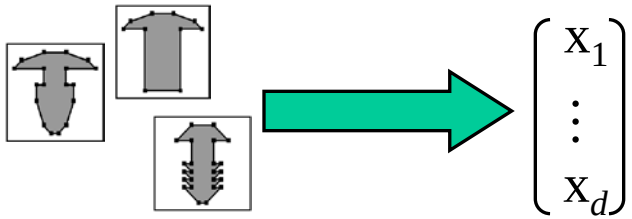


Spiegelung

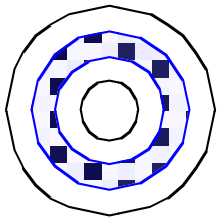


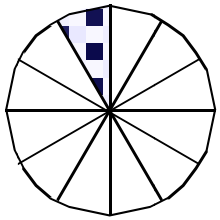
Feature-Transformationen für räumliche Objekte

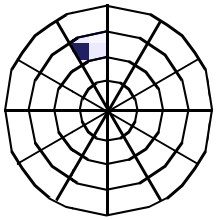
Volume Model [Ankerst, Kastenmüller, Kriegel, Seidl 99]



- 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

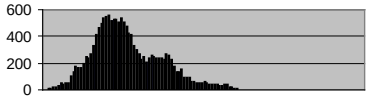

Schalenmodell


Sektorenmodell

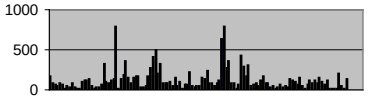

kombiniertes Modell

- Beispiel

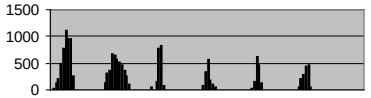

Seryl-tRNA Synthetase



Schalenmodell (120 Schalen)



Sektorenmodell (122 Sektoren)



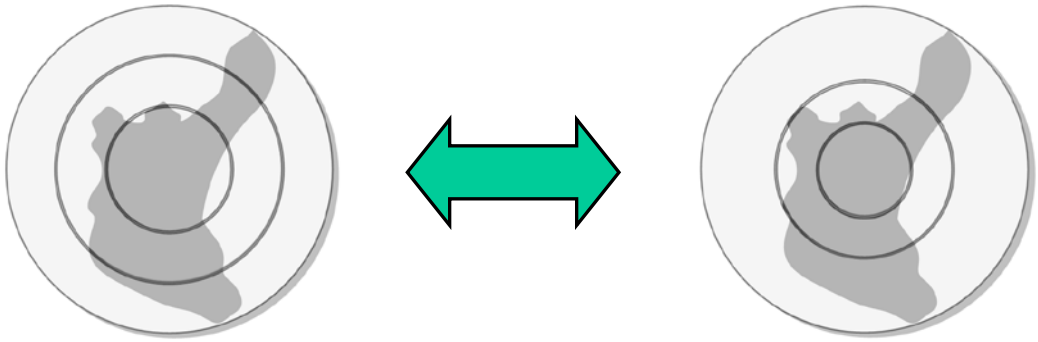
kombiniertes Modell (20 Schalen, 6 Sektoren)

Feature-Transformationen für räumliche Objekte

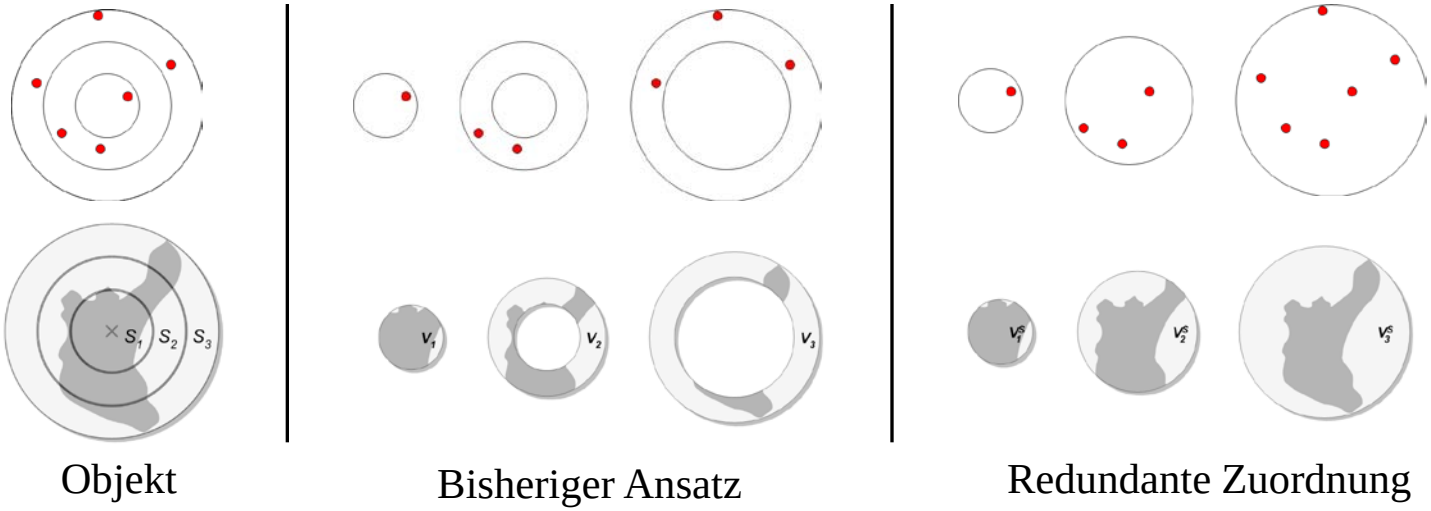
- Formale Definition der Histogramme
  - *Schalenmodell*: Definiere die Bins über den Abstand zum Mittelpunkt, d.h. Anzahl der Punkte auf der jeweiligen Schale.
  - *Sektorenmodell*: Anzahl der Punkte im jeweiligen Sektor.
  - *Kombiniertes Modell*: Synthese aus Schalen- und Sektorenmodell.
- Invarianzen
  - 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, dieses ist inhärent rotationsinvariant.

Feature-Transformationen für räumliche Objekte

- Verbesserung der Formhistogramme [Aßfalg, Kriegel, Kröger, Pötke 05]
  - Proportionale Aufteilung



  - Redundante Zuordnung zu den Bins

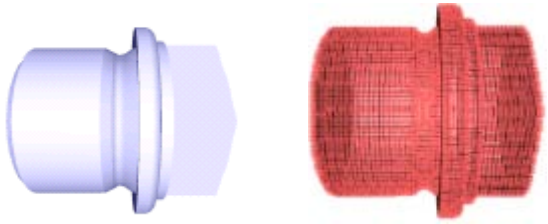


Objekt | Bisheriger Ansatz | Redundante Zuordnung

Feature-Transformationen für räumliche Objekte

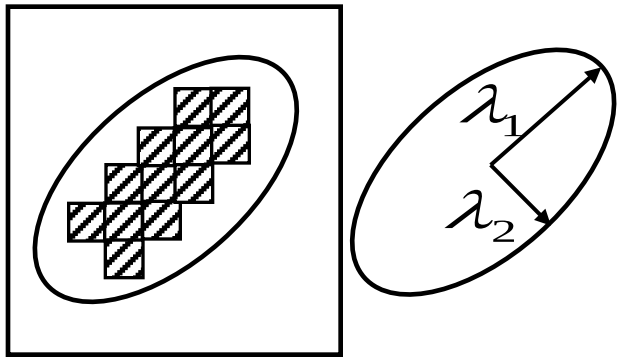
Eigenvalue Model [Kriegel, Kröger, Mashael, Pfeifle, Pötke , Seidl 03]

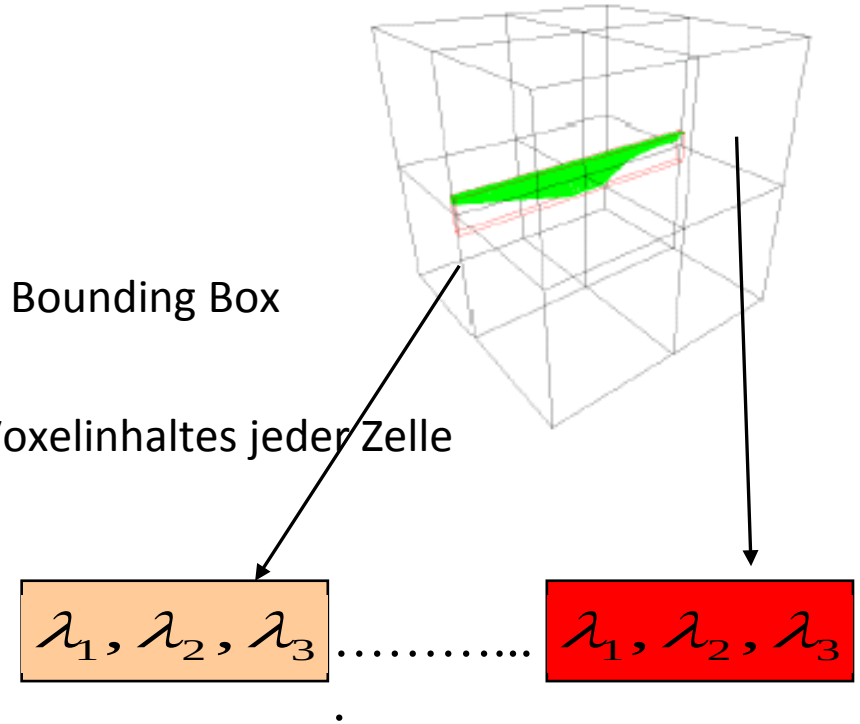
- Volumen-Diskretisierung durch Voxel (3dimensionale Pixel)



- Würfelförmige Partitionierung der Bounding Box

- Bestimmung der Eigenwerte des Voxelinhaltes jeder Zelle

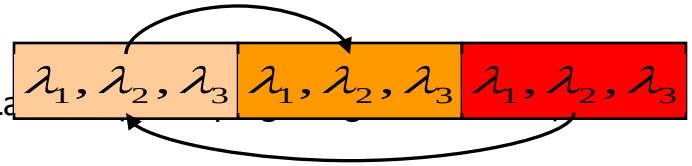




Feature-Transformationen für räumliche Objekte

Invarianzen

- Translationsinvarianz durch Lagenormierung:
  Verschiebung des Schwerpunkts eines Objektes in den Ursprung.
- Skalierungsinvarianz durch Voxelisierung der Bounding Box/Bounding Cube des Objekts mit immer gleicher Voxelauflösung
- Rotationsinvarianz
  - Hauptachsentransformation (völlig rotationsinvariant, aber bei manchen Objekten sensitiv gegenüber kleinen Änderungen)
  - CAD Objekte oft in „vernünftiger“ Lage durch Konstrukteur abgespeichert, dann besser 90-Grad-Rotationsinvarianz: Zur Laufzeit werden die 24 Würfelpositionen durch Permutation der Merkmalsvektor-Elemente simuliert, die Distanz zweier Objekte ist das Minimum über 24 Distanzen
- Reflektionsinvarianz
  - Betrachte 48 statt 24 Permutationen zur La

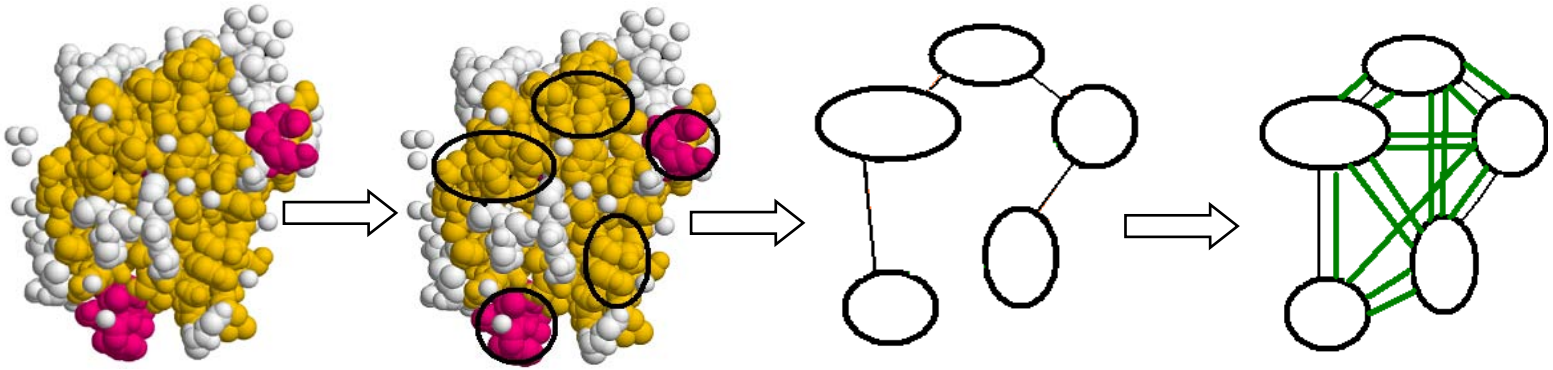


Feature-Transformationen für räumliche Objekte

Protein Datenbanken [Borgwardt, Ong, Schönauer, Vishwanathan, Smola, Kriegel 05]

Idee:

- Graphmodel für Protein 3D-Strukturen
- Knoten: Untereinheiten von Proteinen (secondary structure elements)
- Kanten: Nachbarschaft von Untereinheiten innerhalb der 3D-Struktur und entlang der Aminosäure Sequenz.

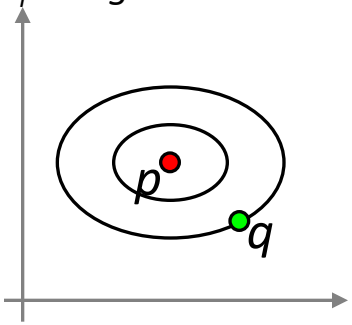


Feature spaces and distance functions

Weighted Euclidean Norm:

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

w_i: *weight in coordinate i*



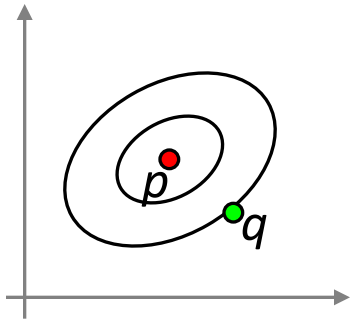
Häufig sind die Wertebereiche der Merkmale deutlich unterschiedlich.

Example: feature M_1 Î[0.01 .. 0.05]
feature M_2 Î[3.1 .. 22.2]

Since M_1 überhaupt berücksichtigt wird, should be weighted higher

Quadratic form:

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



In the previous case, the features are separately weighted.

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

Instead of distance measures, which measure the distance between two objects, we can use sometimes similarity measures

Descriptors for feature vectors

Descriptors for feature vectors

- Consider a set of feature vectors/ objects DB
- Centroid (vgl. Arithmetic mean):

$$\mu_{DB} = \frac{1}{DB} \cdot \sum_{o \in DB} o$$

  - Achtung: bei allgem. Metrischen Räumen muss Centroid nicht notwendigerweise existieren!!!
- Medoid m_{DB}:
  - The most centrally located object in DB
  - Bei allgem. Metrischen Räumen: The one with the smallest average distance to all the other objects in DB
- Variance:

$$Var_{DB} = \frac{1}{DB} \cdot \sum_{o \in DB} dist(o, \mu_{DB})$$

- Standard deviation: square root of variance

Feature spaces and distance functions

Hauptachsenanalyse eine Menge *DB* von *Euklidischen Vektoren*

- Covariance matrix: $$\Sigma_{DB} = \frac{1}{|DB|}\sum_{o \in D}(o - \mu_{DB})(o - \mu_{DB})^{T}$$
- The matrix is decomposed in:
  - an orthonormal matrix V = [e_1,...,e_d] (Eigen vectors)
  - and a diagonal matrix x Λ = diag (λ_1,..., λ_d) (Eigen values)
  - so dass gilt: Σ_{DB} = V Λ V^T
- Interpretation:
  - Eigenvektoren:
    Hauptausrichtung der Datenpunkte in *DB*
  - Eigenwerte:
    Varianz der Datenpunkte in *DB* entlang der entspr. Eigenvektoren