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

Skript zur Vorlesung

Knowledge Discovery in Databases

im Wintersemester 2003/2004

Kapitel 1: Einleitung

Vorlesung: Christian Böhm
Übungen: Karin Kailing und Peer Kröger

http://www.dbs.informatik.uni-muenchen.de/Lehre/KDD

Vorlesungs-Team

Peer Kröger
Oettingenstr. 67, Zimmer E 1.08
Tel. 089/2180-9327
Sprechstunde: Mi, 10^{00}-11^{00}

Karin Kailing
Oettingenstr. 67, Zimmer E 1.06
Tel. 089/2180-9325
Sprechstunde: Fr, 10^{00}-11^{00}

Christian Böhm
Oettingenstr. 67, Zimmer 1.58
Tel. 089/2180-9194
Sprechstunde: Do, 11^{00}-12^{00}

Motivation

- Riesige Datenmengen werden in Datenbanken gesammelt
- Analysen können nicht mehr manuell durchgeführt werden

Von den Daten zum Wissen

Daten	Methode	Wissen
		
		
		
		

Definition KDD

[Fayyad, Piatetsky-Shapiro & Smyth 1996]

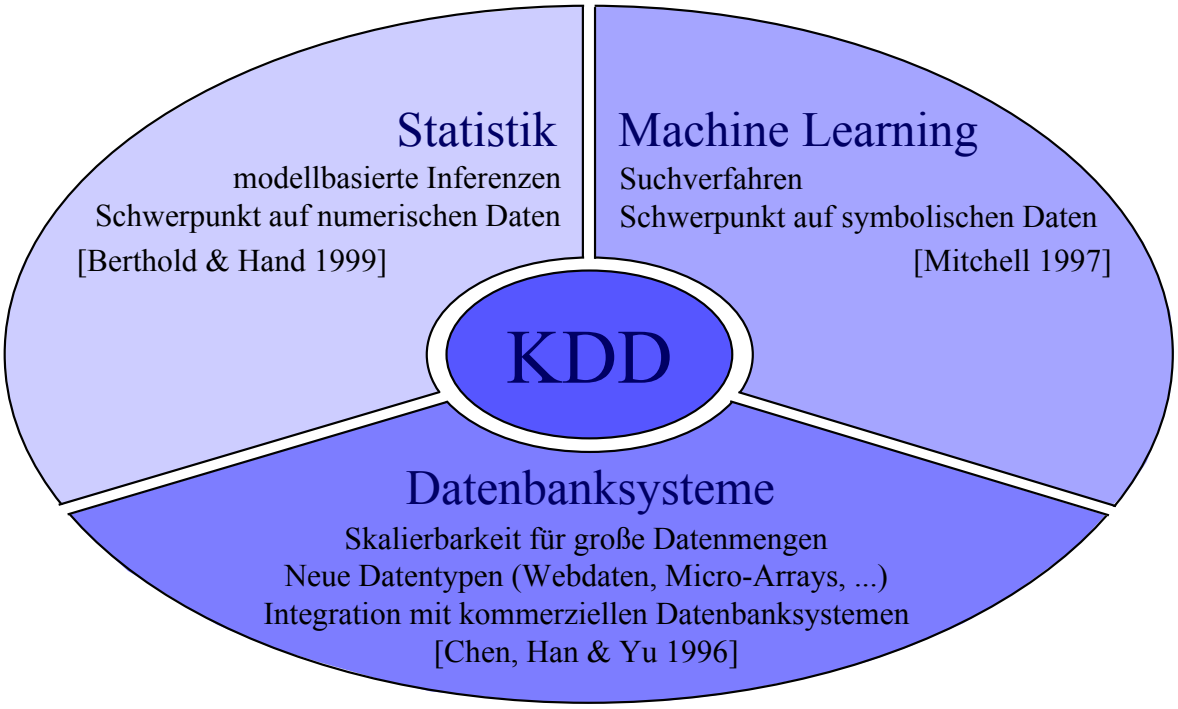
Knowledge Discovery in Databases (*KDD*) ist der Prozess der (semi-) automatischen Extraktion von Wissen aus Datenbanken, das

- *gültig*
- *bisher unbekannt*
- und *potentiell nützlich* ist.

Bemerkungen:

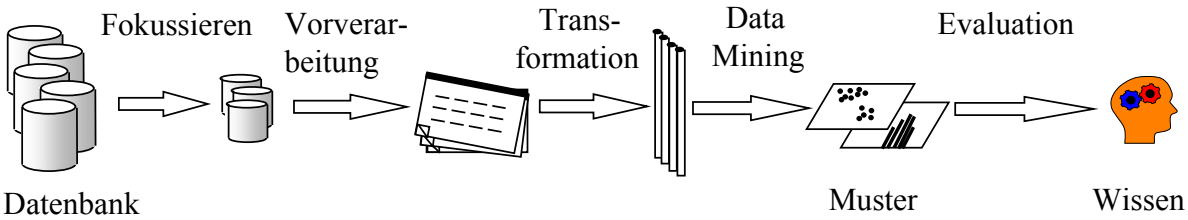
- (*semi-*) *automatisch*: im Unterschied zu manueller Analyse. Häufig ist trotzdem Interaktion mit dem Benutzer nötig.
- *gültig*: im statistischen Sinn.
- *bisher unbekannt*: bisher nicht explizit, kein „Allgemeinwissen".
- *potentiell nützlich*: für eine gegebene Anwendung.

Teilbereiche KDD



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.

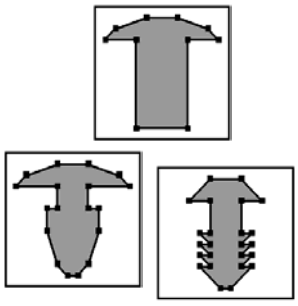
Data Mining
- Generierung der Muster bzw. Modelle

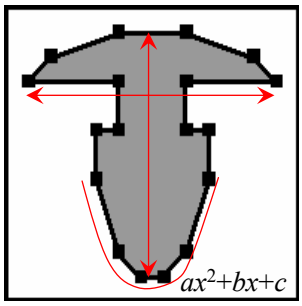
Evaluation
- Bewertung der Interessantheit durch den Benutzer
- Validierung: Statistische Prüfung der Modelle

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 Unterscheidung (Klassifikation, Ähnlichkeit) der Objekte relevant sind.

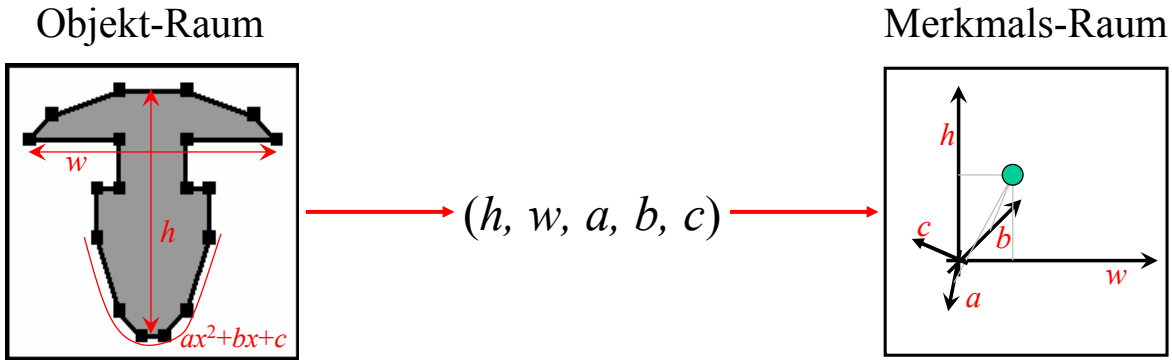
Beispiel: CAD-Zeichnungen:





Mögliche Merkmale:

Merkmals-Vektoren



- 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.)

Weitere Beispiele für Merkmale

Bilddatenbanken:
Farbhistogramme

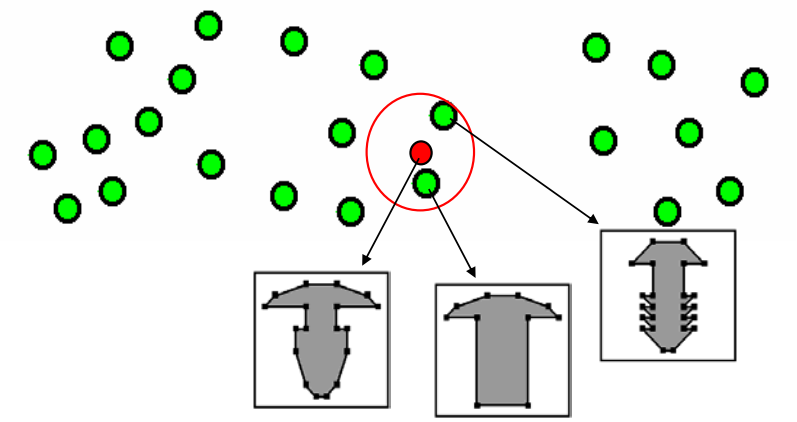
Gen-Datenbanken:
Expressionslevel

Text-Datenbanken:
Begriffs-Häufigkeiten

Der Feature-Ansatz ermöglicht einheitliche Behandlung von Objekten verschiedenster Anwendungsklassen

Ähnlichkeit von Objekten

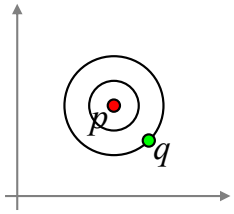
- Spezifiziere Anfrage-Objekt und…
  - … suche ähnliche Objekte – Range-Query (Radius ε)
  - … suche die *k* ähnlichsten Objekte – Nearest Neighbor



Ähnlichkeit von Objekten

Euklidische Norm (L_1):

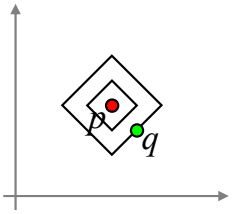
$\delta_1 = ((p_1-q_1)^2+(p_2-q_2)^2+...)^{1/2}$



Natürlichstes Distanzmaß

Manhattan-Norm (L_2):

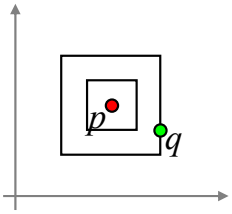
$\delta_2 = |p_1-q_1|+|p_2-q_2|+...$



Die Unähnlichkeiten der einzelnen Merkmale werden direkt addiert

Maximums-Norm (L_∞):

$\delta_\infty = \max\{|p_1-q_1|, |p_2-q_2|,...\}$



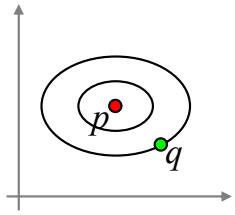
Die Unähnlichkeit des am wenigsten ähnlichen Merkmals zählt

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

Anpassbare Ähnlichkeitsmaße

Gewichtete Euklidische Norm:

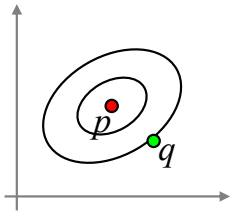
$\delta = (w_1(p_1-q_1)^2 + w_2(p_2-q_2)^2+...)^{1/2}$



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:

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



Bei den bisherigen Ähnlichkeitsmaßen wird jedes Merkmal nur mit sich selbst verglichen (p_1 nur mit q_1 usw.)
Besonders bei Farbhistogrammen müssen auch *verschiedene* Merkmale verglichen werden, z.B. p_{Hellblau} mit $q_{\text{Dunkelblau}}$

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

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)

Data Mining Aufgaben

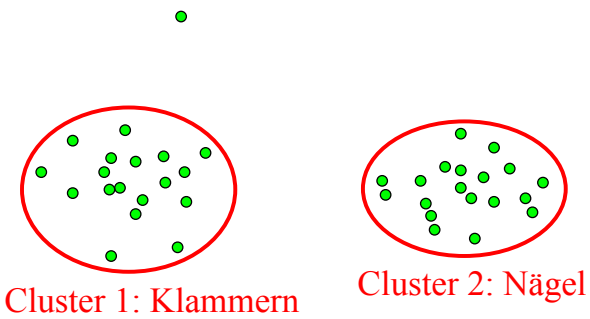
Wichtigste Data-Mining-Verfahren auf Merkmals-Vektoren:

Clustering Outlier Detection	Unsupervised Learning
Klassifikation Regression	Supervised Learning

Supervised: Ein Ergebnis-Merkmal soll gelernt/geschätzt werden
Unsupervised: Die Datenmenge soll lediglich in Gruppen unterteilt werden

Darüber hinaus gibt es zahlreiche Verfahren, die nicht auf Merkmalsvektoren, sondern z.B. auf Texten, Mengen, Graphen arbeiten.

Clustering



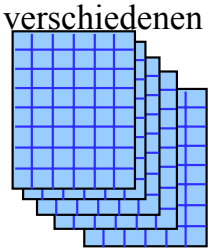
Clustering heisst: Zerlegung einer Menge von Objekten (bzw. Feature-Vektoren) so in Teilmengen (Cluster), dass

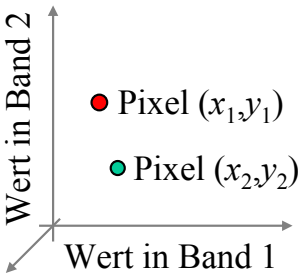
- die Ähnlichkeit der Objekte innerhalb eines Clusters maximiert
- die Ähnlichkeit der Objekte verschiedener Cluster minimiert wird

Idee: Die verschiedenen Cluster repräsentieren meist unterschiedliche Klassen von Objekten; bei unbek. Anzahl und Bedeutung der Klassen

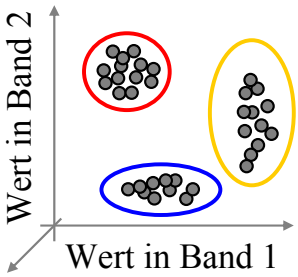
Anwendung: Thematische Karten

Aufnahme der Erdoberfläche
in 5 verschiedenen Spektren



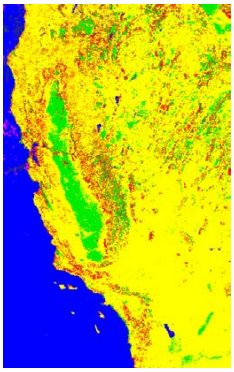


Cluster-Analyse

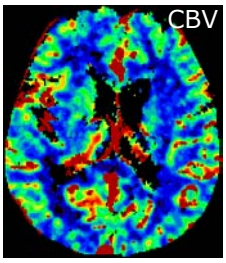


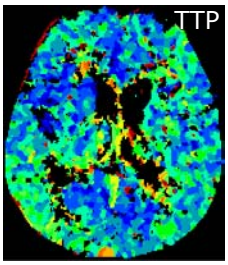
Rücktransformation
in *xy*-Koordinaten

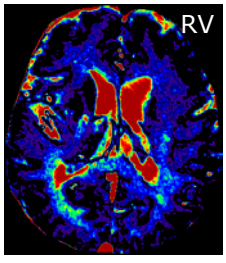
Farbcodierung nach
Cluster-Zugehörigkeit



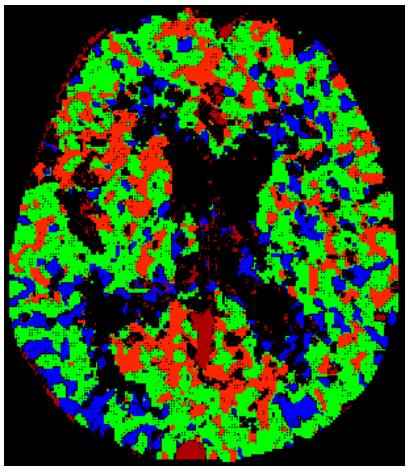
Anwendung: Gewebeklassifikation







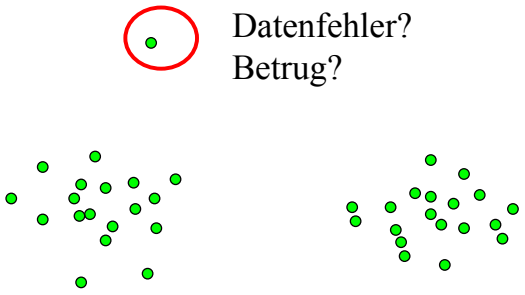
- Schwarz: Ventrikel + Hintergrund
- Blau: Gewebe 1
- Grün: Gewebe 2
- Rot: Gewebe 3
- Dunkelrot: Große Gefäße



	Blau	Grün	Rot
TTP (s)	**20.5**	**18.5**	**16.5**
CBV (ml/100g)	**3.0**	**3.1**	**3.6**
CBF (ml/100g/min)	**18**	**21**	**28**
RV	**30**	**23**	**21**

Ergebnis: Klassifikation cerebralen Gewebes anhand funktioneller Parameter mittels dynamic CT möglich.

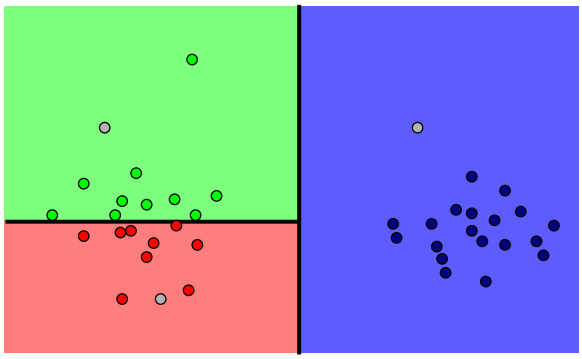
Outlier Detection



Outlier Detection bedeutet:
Ermittlung von untypischen Daten

Anwendungen:
- Entdeckung von Missbrauch etwa bei
  - Kreditkarten
  - Telekommunikation
- Datenfehler

Klassifikation



Aufgabe:
Lerne aus den bereits klassifizierten *Trainingsdaten* die *Regeln*, um neue Objekte nur aufgrund der Merkmale zu klassifizieren

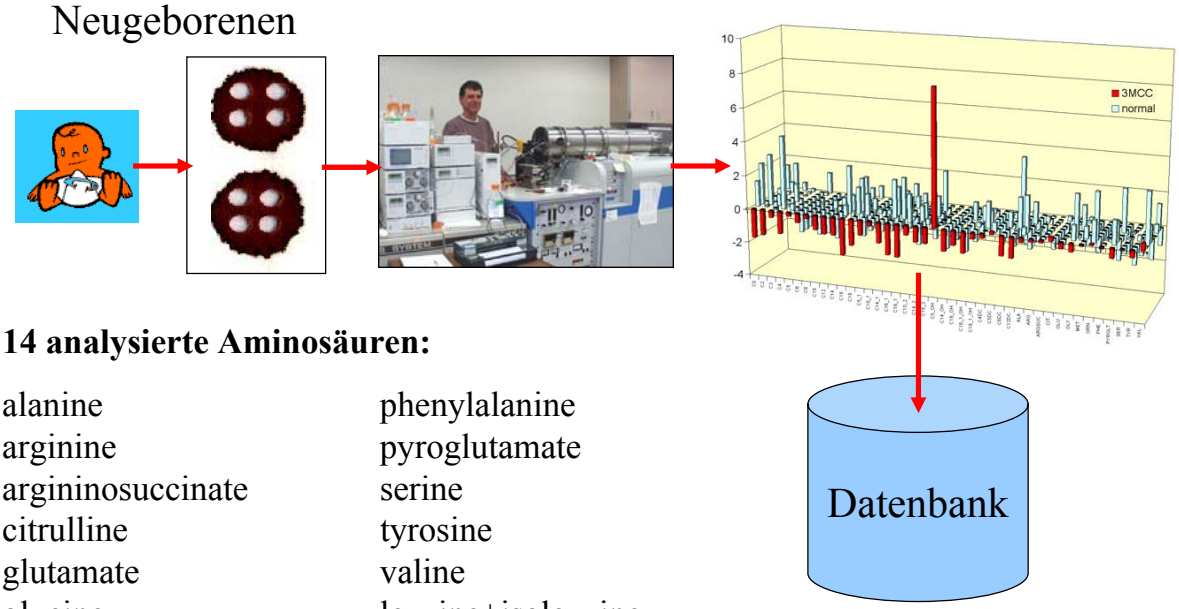
Das Ergebnismerkmal (Klassenvariable) ist nominal (*kategorisch*)

Anwendung: Neugeborenen-Screening

Blutprobe des Neugeborenen

Massenspektrometrie

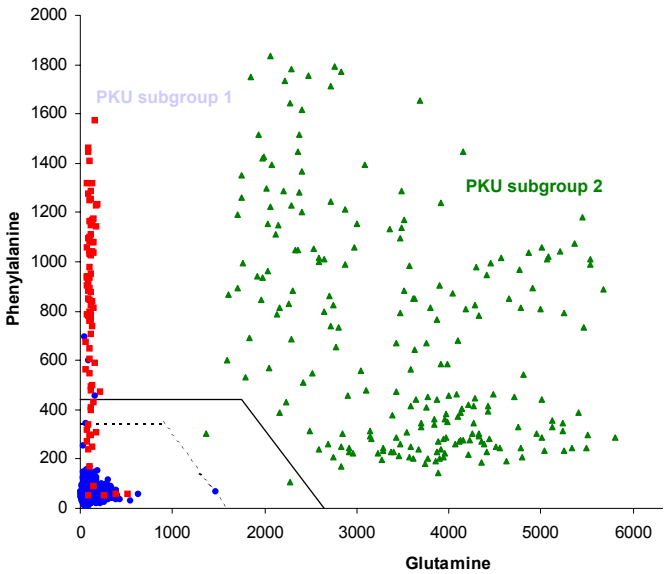
Metabolitenspektrum



14 analysierte Aminosäuren:

alanine
arginine
argininosuccinate
citrulline
glutamate
glycine
methionine
phenylalanine
pyroglutamate
serine
tyrosine
valine
leuzine+isoleuzine
ornitine

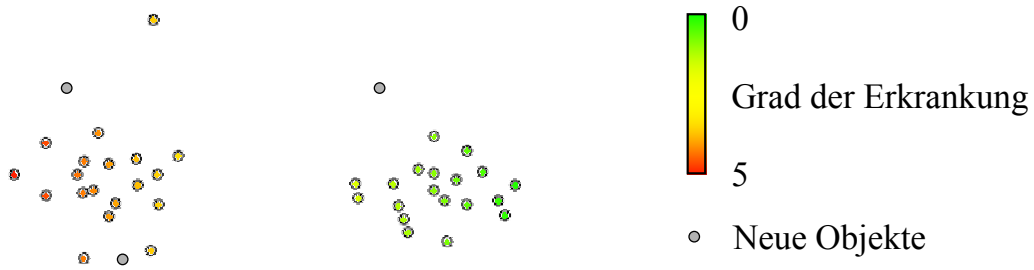
Anwendung: Neugeborenen-Screening



Ergebnis:

- Neuer diagnostischer Test
- Glutamin als bisher unbekannter Marker

Regression



Aufgabe:
Ähnlich zur Klassifikation, aber das Ergebnis-Merkmal, das gelernt bzw. geschätzt werden soll, ist *metrisch*

Überblick über die Vorlesung

1 Einleitung
2 Grundlagen
- Datenbanksysteme
- Statistik

3 Clustering und Outlier
- Hierarchische Verfahren
- Partitionierende Verf.
- Subspace Clustering
- Correlation Clustering

4 Klassifikation
- Bayes-Klassifikation
- Nächste-Nachbarn
- Support Vector Machines

5 Assoziationsregeln
- A Priori Algorithmus
- Taxonomien

6 Besondere Anwendungen
- Temporales Data Mining
- Verteiltes/paralleles DM
- Text-Mining
- Web-Mining
- Molekularbiologie

7 Andere Paradigmen
- Neuronale Netze
- Genetische Algorithmen
- Induktive Logik

Literatur

Lehrbuch zur Vorlesung (deutsch):

Ester M., Sander J.

Knowledge Discovery in Databases: Techniken und Anwendungen

ISBN: 3540673288, Springer Verlag, September 2000, € 39,95

Weitere Bücher (englisch):

Berthold M., Hand D. J. (eds.)

Intelligent Data Analysis: An Introduction

ISBN: 3540430601, Springer Verlag, Heidelberg, 1999, € 63,24

Han J., Kamber M.

Data Mining: Concepts and Techniques

ISBN: 1558604898, Morgan Kaufmann Publishers, August 2000, € 54,30

Mitchell T. M.

Machine Learning

ISBN: 0071154671, McGraw-Hill, 1997, € 61,30

Witten I. H., Frank E.

Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations

ISBN: 1558605525, Morgan Kaufmann Publishers, 2000, € 50,93