

MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2020

Prof. E.G. Schukat-Talamazzini

Stand: 11. Dezember 2020

Module und Studiengänge

Zur Vorlesung

Zur Modulprüfung

Zum Inhalt

Literatur zur Lehrveranstaltung

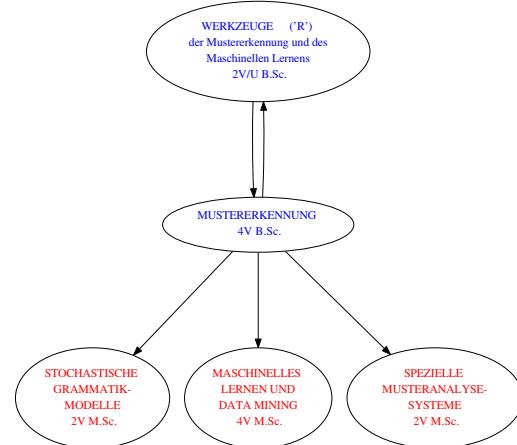
Professur für Musteranalyse

Lehrangebot Wintersemester & Sommersemester

Lehrbereich

Informatik
WP-Bereich
Intelligente Systeme
Vertiefung Künstliche Intelligenz und Mustererkennung
Statistische Musteranalyse

Inhaltliche Abhängigkeiten



Meine Lehrveranstaltungen für ...

Informatiker & Bioinformatiker & Informatikerinnen & Bioinformatikerinnen

	SOMMERSEMESTER	WINTERSEMESTER
ASQ und EF Inf.	Intelligente Systeme 4V	Literaturarbeit und Präsentation ASQ 2S
Bachelor	Mustererkennung 4V Werkzeuge ME/ML 2Ü	Strukturiertes Programmieren 4V+2Ü
Master	Stochastische Grammatikmodelle 2V Biometriesysteme (Seminar) 2S	Maschinelles Lernen und Datamining 4V Spezielle Musteranalyse-systeme 2V

Wie studiere ich MUSTERANALYSE ?

Studiengänge: Informatik · Bioinformatik · Angewandte Informatik

Bachelor

Mustererkennung^{6LP}



WeRkzeuge ME/ML^{3LP}



Vertiefungsangebote

auf Antrag beim Prüfungsamt:

Maschinelles Lernen^{6LP}



Master

Maschinelles Lernen^{6LP}



Stochast. Grammatik^{3LP}



Musteranalysesysteme^{3LP}



Nivellierungsmodule

auf Antrag beim Prüfungsamt:

Mustererkennung^{6LP}



WeRkzeuge ME/ML^{3LP}



Stochastische Grammatikmodelle

Vorlesung der Master-Studiengänge · 2V · 3 LP

Modul FMI-IN0146 Stochastische Grammatikmodelle	
Modulnummer/-code	FMI-IN0146
Modultitel (deutsch)	Stochastische Grammatikmodelle
Modultitel (englisch)	Stochastic Grammars
Modulverantwortlicher	Ernst Günter Schukat-Talamazzini
Voraussetzungen für Zulassung zum Modul	keine
Empfohlene bzw. erwartete Vorkenntnisse	keine
Art des Moduls (Pflicht-, Wahlpflicht- oder Wahlmodul)	Wahlpflichtmodul (KIME, INT) für den M.Sc. Informatik Wahlpflichtmodul für den M.Sc. Bioinformatik (Bereich bioinformatisch relevante Informatik) Wahlpflichtmodul für den M.Sc. Mathematik (Nebenfach Informatik) Wahlpflichtmodul (INF) für den M.Sc. Computational Science
Häufigkeit des Angebots (Zyklus)	jedes 2. Semester (ab Sommersemester)
Dauer des Moduls	1 Semester
Zusammensetzung des Moduls / Lehrformen (VL, Ü, S, Praktikum)	2V
Leistungspunkte (ECTS credits)	3LP
Arbeitsaufwand (work load)	90h
- Präsenzstunden	30h
- Selbststudium (einschl. Prüfungsvorbereitungen)	60h
Inhalte	Grammatische Modellierung von Zeichenfolgen natürlicher („Texte“) und künstlicher (z.B. Nukleotid- oder Aminosäuresequenzen) Sprachen. Vorlesungsthemen sind u.a.: <ul style="list-style-type: none">Schwach kontextfreie Grammatiken (IG, TAG, HG, CG)Information/Kompressionrobuste Häufigkeitsschätzung (Bayes, Good-Turing, Zipf)N-Gramme, Interpolation, Maximum-Entropiestochastische

M.Sc. Informatik

WP-Bereich Int.Syst.
Schwerpunkt KI/ME

M.Sc. Bioinform.

WP-Bereich

M.Sc. Comp.Science

WP-Bereich

Maschinelles Lernen & Data Mining

Vorlesung der Master-Studiengänge · 4V · 6 LP

Modultitel (deutsch)	Maschinelles Lernen und Datamining
Modultitel (englisch)	Machine Learning and Datamining
Modulnummer	FMI-IN0034
Art des Moduls (Pflicht-, Wahlpflicht- oder Wahlmodul)	02.12.09 Wahlpflichtmodul (KIME, INT) für den M.Sc. Informatik Wahlpflichtmodul (INT) für den B.Sc. Informatik (zusätzliches Lehrangebot) Wahlpflichtmodul für den M.Sc. Bioinformatik (Bereich Informatik) Wahlpflichtmodul (INF) für den M.Sc. Computational Science Wahlpflichtmodul für das Lehramt Informatik
Modul-Verantwortlicher	Ernst Günter Schukat-Talamazzini
Leistungspunkte (ECTS credits)	6
Arbeitsaufwand (work load) in:	180 Std. - Präsenzstunden - Selbststudium (einschl. Prüfungsvorbereitung)
Lehrform (SWS)	4V (mit Projektanteil)
Häufigkeit des Angebots (Modulturnus)	jährlich im Wintersemester
Dauer des Moduls	1 Semester
Voraussetzung für die Zulassung zum Modul	Keine
Empfohlene Vorkenntnisse für das Modul	FMI-IN0036 (Mustererkennung)
Voraussetzung für die Zulassung zur Modulprüfung	Keine
Voraussetzung für die Vergabe von Leistungspunkten (Prüfungsform)	Klausur (120min) oder mündliche Prüfung (30min) zur Vorlesung
Inhalte	Strukturaufdeckung, Klassifizierung oder Entwicklungsvorhersage aus großen Datenflüssen (Finanzprozesse, Handel und Transport, med./biol. Datensätze, Klimamesswerte, elektronische Dokumente, Fertigungsautomatisierung) – Vorlesungsthemen sind u.a.: Skalentypen; Visualisierung hochdimensionaler Daten (PCA, MDS, ICA); überwachte Lernverfahren (Versionenraum, Entscheidungsbaum, lineare/logistische Modelle); unüberwachte Lernverfahren (hierarchisch, (fuzzy) K-means, spektral); Graphische Modelle (Bayesnetze, Markovnetze, Induktion und Inferenz)
(Qualifikations-)Ziele	Tiefgreifende Fachkenntnisse des Gebiets Maschinelles Lernen Fähigkeit zur Analyse, Design und Realisierung von ML-Systemen Flächendeckende Übersicht aktueller Techniken des Datamining

Spezielle Musteranalysesysteme

Vorlesung der Master-Studiengänge · 2V · 3 LP

Modultitel (deutsch)	Spezielle Musteranalysesysteme
Modultitel (englisch)	Pattern Analysis Systems
Modulnummer	FMI-IN0054
Art des Moduls (Pflicht-, Wahlpflicht- oder Wahlmodul)	02.12.09 Wahlpflichtmodul (INT) für den M.Sc. Informatik Wahlpflichtmodul für den M.Sc. Bioinformatik (Bereich Informatik)
Modul-Verantwortlicher	Ernst Günter Schukat-Talamazzini
Leistungspunkte (ECTS credits)	3
Arbeitsaufwand (work load) in:	90 Std. - Präsenzstunden - Selbststudium (einschl. Prüfungsvorbereitung)
Lehrform (SWS)	2V (mit Projektanteil)
Häufigkeit des Angebots (Modulturnus)	jährlich im Sommersemester
Dauer des Moduls	1 Semester
Voraussetzung für die Zulassung zum Modul	keine
Empfohlene Vorkenntnisse für das Modul	- FMI-IN0036 (Mustererkennung) - Vorkenntnisse aus den Bereichen Künstliche Intelligenz und Digitale Bildverarbeitung
Voraussetzung für die Zulassung zur Modulprüfung	keine
Voraussetzung für die Vergabe von Leistungspunkten (Prüfungsform)	mündliche Prüfung (30min) zur Vorlesung oder Ausarbeitung/ Präsentation zu einer Projektaufgabe
Inhalte	- Komplexe Musteranalyseaufgaben mit longitudinalen Daten (Sprach- und Sprechererkennung, (Hand)schrifterkennung, DNA-Motive, Musikretrieval) - Geeignete Lernverfahren (z.B. Hidden Markov Modelle; siehe Webseite zum Kurs für Detailinformationen), unterstützende Werkzeuge, Vorverarbeitung und Etikettierung der Lerndaten und syntaktische Modellierungsverfahren am Beispiel einer oder mehrerer ausgewählter Aufgabenstellungen
(Qualifikations-)Ziele	- Vertiefte Kenntnis der Methoden syntaktischer Musteranalyse - Kompetenzen der Analyse, des Designs und der Realisierung von Musteranalysesystemen realistischer Größeordnung - Fertigkeiten der Nutzung ausgewählter Softwarewerkzeuge der syntaktischen Musteranalyse
Literatur	Scholar Technomic, Ernst G.: Automatische Spracherkennung, Wissens...

M.Sc. Informatik

WP-Bereich Int.Syst.

M.Sc. Bioinform.

WP-Bereich Informatik

Mustererkennung

Vorlesung der Bachelor/Master-Studiengänge · 4V · 6 LP

Modultitel (deutsch)	Mustererkennung
Modultitel (englisch)	Pattern Recognition
Modulnummer	FMI-IN0036
Art des Moduls (Pflicht-, Wahlpflicht-, oder Wahlmodul)	Wahlpflichtmodul (INT) für den B.Sc. Informatik Wahlpflichtmodul (INT) für den B.Sc. Angewandte Informatik Wahlpflichtmodul (Wahlpflichtbereich 2) für den B.Sc. Bioinformatik Wahlpflichtmodul (INT) für den M.Sc. Informatik (auf Antrag) Wahlpflichtmodul für den M.Sc. Bioinformatik (Bereich Informatik) Pflichtmodul für das Anwendungsfach Computational Neuroscience zum B.Sc. Angewandte Informatik Wahlpflichtmodul für das Lehramt Informatik
Modul-Verantwortlicher	Ernst Günther Schukat-Talamazzini
Leistungspunkte (ECTS credits)	6
Arbeitsaufwand (work load) in:	180 Std. - Präsenzstunden 60 Std. - Selbststudium (einschl. Prüfungsvorbereitung) 120 Std.
Lehrform (SWS)	3 V + 1 Ü
Häufigkeit des Angebots (Modulturnus)	jährlich im Sommersemester
Dauer des Moduls	1 Semester
Voraussetzung für die Zulassung zum Modul	
Empfohlene Vorkenntnisse für das Modul	- FMI-IN0070 (Grundlagen der Modellierung und Programmierung) oder FMI-IN0040 (Grundlagen der Modellierung und Programmierung (Grundteile)) oder FMI-IN0025 (Strukturierte Programme für Bioinformatiker) - FMI-IN0001 (Algorithmen und Datenstrukturen) - FMI-MA0007 (Einführung in die Wahrscheinlichkeitstheorie)
Voraussetzung für die Zulassung zur Modulprüfung	Bearbeitung der Übungsaufgaben Mindestens 50% der erzielbaren Punkte erreicht
Voraussetzung für die Vergabe von Leistungspunkten (Prüfungsform)	Klausur (120min) oder mündliche Prüfung (30min) zur Vorlesung Studienleistungsbegrenzte Erfolgsmetriken. Abgestufte (Prüfungs-)Anforderungen berücksichtigen das vom Bachelor- und Masterstudierenden jeweils erwartbare Leistungsniveau.
Inhalte	Einführung in die Methoden der Mustererkennung zur maschinellen Modellierung und Simulation komplexer Informationsverarbeitungsprozesse, insbesondere bei der Wahrnehmung und Auswertung visueller, akustischer oder taktiler Sinnesindrücke durch den Menschen auftreten. Diskretilisierung/Filtrierung/Normierung; Merkmalauswahl und Merkmalstransformation; statistische, diskriminative und nichtparametrische Klassifikatoren; unüberwachtes Lernen; Zeitreihen-

B.Sc. Informatik

WP-Bereich Int.Syst.

B.Sc. Bioinform.

WP-Bereich

B.Sc. Ang.Inform.

WP-Bereich Int.Syst.

Pflicht im Anw.fach CNS

M.Sc. Informatik

WP-Bereich Int.Syst.

M.Sc. Bioinform.

WP-Bereich Int.Syst.

LG Informatik

FS 6–9

Werkzeuge Mustererkennung & Maschinelles Lernen

Vorlesung der Bachelor/Master-Studiengänge · 2V/P · 3 LP

Modultitel (deutsch)	Werkzeuge der Mustererkennung und des Maschinellen Lernens
Modultitel (englisch)	Tools for Pattern Recognition and Machine Learning
Modulnummer	FMI-IN0086
Art des Moduls (Pflicht-, Wahlpflicht-, oder Wahlmodul)	Wahlpflichtmodul (INT) für den B.Sc. Informatik Wahlpflichtmodul (INT) für den B.Sc. Angewandte Informatik Wahlpflichtmodul (KIME, INT) für den M.Sc. Informatik (auf Antrag) Wahlpflichtmodul für den B.Sc. Bioinformatik (Bereich Informatik)
Modul-Verantwortlicher	Ernst Günter Schukat-Talamazzini
Leistungspunkte (ECTS credits)	3
Arbeitsaufwand (work load) in:	90 Std. - Präsenzstunden 30 Std. - Selbststudium (einschl. Prüfungsvorbereitung) 60 Std.
Lehrform (SWS)	2V (mit Übung)
Häufigkeit des Angebots (Modulturnus)	jedes Sommersemester
Dauer des Moduls	1 Semester
Voraussetzung für die Zulassung zum Modul	Keine
Empfohlene Vorkenntnisse für das Modul	FMI-IN0036 (Mustererkennung) sollte gleichzeitig belegt werden
Voraussetzung für die Zulassung zur Modulprüfung	50% der erreichbaren Punkte aus den Übungsaufgaben
Voraussetzung für die Vergabe von Leistungspunkten (Prüfungsform)	Mündliche Prüfung oder Klausur
Inhalte	Aufgabenstellungen aus den Bereichen Mustererkennung, Maschinelles Lernen, Datamining und ihre Bearbeitung mit geeigneten Softwarewerkzeugen: Klassifikation, Vorhersage, Clustering, Transformation, Visualisierung, Zeitreihen, Spektraldarstellung, Wahrscheinlichkeitsmodelle
(Qualifikations-)Ziele	- Fähigkeiten im praktischen Umgang mit Entwicklungswerkzeugen für maschinelles Lernen in Musteranalyse und Datamining - Grundlegende Kenntnisse über den Aufbau von Softwaresystemen und Programmierparadigmen für die maschinelle Datenanalyse - Kompetenzen in Datenanalyse, Versuchsplanung, Konfiguration von ML-Lösungen

Vorlesung

Nutzung der Folienpräsentation

Module und Studiengänge

Zur Vorlesung

Zur Modulprüfung

Zum Inhalt

Literatur zur Lehrveranstaltung

- Die Folien sollen vom Mitschreiben während der Vorlesung entlasten.
- Das Mitschreiben wird dadurch nicht überflüssig.
- Die Folien sind kein Lehrbuch.
- Die Folien sind daher im allgemeinen nur mit den Erläuterungen während der Vorlesung und entsprechenden eigenen Notizen verständlich.

Vorlesung

[Mathematische Sachverhalte](#)

- Wichtige mathematische Grundlagen werden in Steilkursen wiederholt.
- Die entsprechenden Fakten sind (oft) im letzten Abschnitt eines Vorlesungsteils dargestellt.
- Schwierige mathematische Zusammenhänge werden in der Anwendung verständlicher.
- Umfangreiche mathematische Formeln erscheinen viel harmloser, nachdem man/frau sie einmal programmtechnisch umgesetzt hat.

Vorlesung

[Elektronisches Folienskript](#)

Die PDF-Fassung des Folienskripts enthält einige Hyperlinks:

- **Verweise auf externe Webseiten**
Detaillierte Zusatzinformationen, Daten, Bilder
(funktioniert nicht während der Vorlesung ...)
- **Literaturangaben**
Verweis auf Quellenangaben am Ende des Dokuments
- **Programmcode**
'R'-Code zur Erstellung einer Grafik oder Tabelle
'dot'-Code zur Erzeugung eines (gerichteten) Graphen

[Module und Studiengänge](#)

[Zur Vorlesung](#)

[Zur Modulprüfung](#)

[Zum Inhalt](#)

[Literatur zur Lehrveranstaltung](#)

Prüfungsvorgang

Schriftliche Klausurarbeit (2–3 Stunden)

Prüfungstermine

[mehr Information](#)

Erstprüfung am Dienstag 2. März 2021 (**geändert!**)

Wiederholung am Mittwoch 31. März 2021

Prüfungsstoff

Vorlesungsinhalte in Schrift und Wort

(siehe Folienskript und Anleitungstexte im Moodle-Kurs)

„Generalprobe“ (mit Beispielfragen) in der letzten Vorlesungswoche

Anmeldung und Zulassung

zur Teilnahme **und** zur Prüfung

Deadline: **Montag, 11.01.2021**



Module und Studiengänge

Zur Vorlesung

Zur Modulprüfung

Zum Inhalt

Literatur zur Lehrveranstaltung

Lehrveranstaltungsform

Vorlesung (4V) · kein Übungsanteil

Zulassungsvoraussetzungen

keine, aber empfehlenswert: *Mustererkennung*

Themengebiet

Explorative Analyse großer Datenmengen mit maschinellen Lernverfahren

Zweck

Verstehen der wichtigsten Konzepte *hinter* den Schaltflächen einschlägiger Datamining-Systeme

Lernziele

Automatische Strukturaufdeckung in Datenbeständen
Behandlung numerischer und nichtnumerischer Attribute

Zum Vorlesungsinhalt

Die Grundrechnungsarten des Datamining

Datenpräparation

Akquisition · Auswahl · Filterung · Komplettierung

Visualisierung

hochdimensionaler oder nichtnumerischer Datensätze

Kategorisierung

von Objekten unterschiedlicher Attributskalen

Gruppierung

von Objekten unterschiedlicher Attributskalen

Vorhersage

verdeckter Attribute oder zukünftiger Objekte

Abhängigkeitsstruktur

Stärke und Richtung von Attributassoziationen

Module und Studiengänge

Zur Vorlesung

Zur Modulprüfung

Zum Inhalt

Literatur zur Lehrveranstaltung

Maschinelles Lernen

Empfohlene Bücher zur Vorlesung

-  **Tom M. Mitchell.**
Machine Learning.
McGraw-Hill Series in Computer Science. McGraw-Hill, New York, NY, 1997.
-  **Miroslav Kubat.**
An Introduction to Machine Learning.
Springer, 2015.
-  **Yuichiro Anzai.**
Pattern Recognition and Machine Learning.
Academic Press, San Diego, CA, 1992.
-  **Dana H. Ballard.**
An Introduction to Natural Computation.
Complex Adaptive Systems. MIT Press, Cambridge, MA, 1997.
-  **R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors.**
Machine Learning. An Artificial Intelligence Approach.
Symbolic Computation. Springer, Berlin, 1984.
-  **Paul Fischer.**
Algorithmisches Lernen.
Teubner, Wiesbaden, 2000.

Data Mining

Empfohlene Bücher zur Vorlesung

-  **Michael R. Berthold, Christian Borgelt, Frank Höppner, and Frank Klawonn.**
Guide to Intelligent Data Analysis.
Texts in Computer Science. Springer, 2010.
-  **J. Han and M. Kamber.**
Data Mining: Concepts and Techniques.
Morgan Kaufmann, 2000.
-  **Max Bramer.**
Principles of Data Mining.
Undergraduate Topics in Computer Science. Springer, 2013.
2. Auflage.
-  **Charu C. Aggarwal.**
Outlier Analysis.
Springer, 2017.
-  **Thomas A. Runkler.**
Data Analytics. Models and Algorithms for Intelligent Data Analysis.
Springer Vieweg, 2012.
-  **Michael Berthold and David J. Hand.**
Intelligent Data Analysis.
Springer, 2003.
2. Auflage.

Numerisch orientiertes Lernen

Empfohlene Bücher zur Vorlesung

-  **Christopher M. Bishop.**
Pattern Recognition and Machine Learning.
Springer, 2006.
Hardcover 60 EUR.
-  **Bertrand Clarke, Ernest Fokoue, and Hao Helen Zhang.**
Principles and Theory for Data Mining and Machine Learning.
Springer Series in Statistics. Springer, 2009.
-  **T. Hastie, R. Tibshirani, and J. Friedman.**
The Elements of Statistical Learning.
Springer, 2001.
-  **G. James, D. Witten, T. Hastie, and R. Tibshirani.**
An Introduction to Statistical Learning.
Number 103 in Springer Texts in Statistics. Springer, 2013.
-  **Alan Julian Izenman.**
Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning.
Springer Texts in Statistics. Springer, 2008.
-  **Hubert B. Keller.**
Maschinelle Intelligenz. Grundlagen, Lernverfahren, Bausteine intelligenter Systeme.

Statistische Modelle

Empfohlene Bücher zur Vorlesung

-  **Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang.**
Regression.
Springer, 2007.
-  **Lucas Drumond.**
Factorization Models for Multi-Relational Data.
Cuvillier Verlag, 2014.
-  **J. Kreiß and G. Neuhaus.**
Einführung in die Zeitreihenanalyse.
Springer, 2006.
-  **Peter Bühlmann and Sara van de Geer.**
Statistics for High-Dimensional Data.
Springer Series in Statistics. Springer, 2011.
-  **John C. Loehlin.**
Latent Variable Models.
Lawrence Erlbaum Assoc Inc, 2004.
-  **Giovanni Petris, Sonia Petrone, and Patrizia Campagnoli.**
Dynamic Linear Models with R.
Use R! Springer, 2009.

Graphische Modelle

Empfohlene Bücher zur Vorlesung

-  Daphne Koller and Nir Friedman.
Probabilistic Graphical Models. Principles and Techniques.
Adaptive Computation and Machine Learning. MIT Press, 2009.
-  Judea Pearl.
Probabilistic Reasoning in Intelligent Systems.
Morgan Kaufmann, 1997.
-  D.R. Cox and Nanny Wermuth.
Multivariate Dependencies. Models, Analysis and Interpretation.
Chapman & Hall, Boca Raton, 1996.
-  Joe Whittaker.
Graphical Models in Applied Multivariate Statistics.
John Wiley & Sons, Chichester, 1995.
-  Steffen L. Lauritzen.
Graphical Models.
Oxford Statistical Science Series. Clarendon Press, Oxford, 1996.

Graphische Modelle (DAG)

Empfohlene Bücher zur Vorlesung

-  Richard E. Neapolitan.
Learning Bayesian Networks.
Prentice Hall, 2003.
-  Richard E. Neapolitan.
Probabilistic Reasoning in Expert Systems.
John Wiley & Sons, 1990.
-  Russell G. Almond.
Graphical Belief Modeling.
Chapman & Hall, London, 1995.
-  F.B. Jensen.
Bayesian Networks and Decision Graphs.
Springer, 2001.
-  Judea Pearl.
Causality.
Cambridge University Press, 2000.

Assoziationsregeln, Warenkorbanalyse, Netzwerkdaten

Empfohlene Bücher zur Vorlesung

-  Charu C. Aggarwal.
Recommender Systems. The Textbook.
Springer, 2016.
-  Jean-Marc Adamo.
Data Mining for Association Rules and Sequential Patterns.
Springer, 2001.
-  Paul Alpar and Joachim Niedereichenholz, editors.
Data Mining im praktischen Einsatz.
Vieweg, Wiesbaden, 2000.
-  Alex A. Freitas.
Data Mining and Knowledge Discovery with Evolutionary Algorithms.
Springer, 2002.
-  Chengqi Zhang and Shichao Zhang.
Association Rule Mining. Models and Algorithms, volume 2307 of LNCS.
Springer, 2002.
-  Eric D. Kolaczyk and Gábor Csárdi.
Statistical Analysis of Network Data with R.
Use R! Springer, 2014.

Ausgewählte Verfahren ML/DM

Empfohlene Bücher zur Vorlesung

-  Lior Rokach and Oded Maimon.
Data Mining with Decision Trees, volume 81 of *Machine Perception and Artificial Intelligence*.
Springer, 2014.
-  D. Goldberg.
Genetic Algorithms: Search, Optimization and Machine Learning.
Addison-Wesley, Reading, MA, 1989.
-  A. Hyvärinen, J. Karhunen, and E. Oja.
Independent Component Analysis.
John Wiley & Sons, 2001.
-  A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari.
Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.
Wiley and Sons, 2009.
-  Richard S. Sutton and Andrew G. Barto.
Reinforcement Learning: An Introduction.
MIT Press, 1998.
-  Michel Neuhaus and Horst Bunke.
Bridging the Gap between Graph Edit Distance and Kernel Machines, volume 68 of *Series in Machine Perception and Artificial Intelligence*.

Spezielle Anwendungen ML/DM

Empfohlene Bücher zur Vorlesung

 Francesco Camastra and Alessandro Vinciarelli.
Machine Learning for Audio, Image and Video Analysis.
Springer, 2010.

 Pierre Baldi and Søren Brunak.
Bioinformatics. The Machine Learning Approach.
Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 1998.

 Thorsten Joachims.
Learning to Classify Text Using Support Vector Machines.
Kluwer Academic Publ., Boston, MA, 2002.

 Reginald Ferber.
Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web.
dpunkt.verlag, 2003.

 Paul S.P. Cowpertwait and Andrew V. Metcalfe.
Introductory Time Series with R.
Use R! Springer, 2009.

Softwaresysteme

Empfohlene Bücher zur Vorlesung

 Ian H. Witten and Eibe Frank.
Data Mining: Practical Machine Learning Tools and Techniques.
Morgan Kaufmann, 2005.

 David Edwards.
Introduction to Graphical Modelling.
Springer Texts in Statistics. Springer, New York, NY, 1995.

 Brian Everitt and Torsten Hothorn.
An Introduction to Applied Multivariate Analysis with R.
Use R. Springer, 2011.

 Graham Williams.
Data Mining with Rattle and R.
Use R. Springer, 2011.

 Søren Højsgaard, David Edwards, and Steffen Lauritzen.
Graphical Models with R.
Use R! Springer, 2012.

 Jeremy Kepner and Hayden Jananthan.
Mathematics of Big Data. Spreadsheets, Databases, Matrices, and Graphs.
MIT Press, 2018.

Teil I

MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2020

Prof. E.G. Schukat-Talamazzini

Stand: 10. September 2020

Methoden und Aufgabenstellungen



ML \triangleq Statistik & Wahrscheinlichkeit

Interviewer: What's your biggest strength?
Me: I'm an expert in machine learning.
Interviewer: What's $9 + 10$?
Me: Its 3.
Interviewer: Not even close. It's 19.
Me: It's 16.
Interviewer: Wrong. Its still 19.
Me: It's 18.
Interviewer: No, it's 19.
Me: it's 19.
Interviewer: You're hired

ML \triangleq Iterative(!) Optimierung

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Lernen nach Herbert Simon

„Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task (or tasks drawn from the same population) more efficiently and more effectively the next time.“

(Automatic Performance Improvement)

Lernen nach Dana Scott

Prozeß des Aufbaus abrufbarer Repräsentationen von vergangenen Interaktionen mit der Umwelt

Lernen nach Ryszard Michalski

Konstruieren oder Verändern der Repräsentationen von Erfahrungen

Trifft Simons Definition unser intuitives Verständnis?

... zu weit?
Schärfen eines Messers schnellere CPU

... zu eng?
Zwangsarbeiter täuscht Leistung vor
Passant \rightsquigarrow Oper \rightsquigarrow Auskunft

Leistungsbegriff?!

Wozu maschinelles Lernen ?

Lernen ist der Schlüssel zur Intelligenz — bei Mensch und Maschine

Knowledge Acquisition Bottleneck

Experten sind oft unfähig, ihr Wissen zu formalisieren.

Wissenserwerb und -einpflege

... sind teuer, langsam und unsicher.

Problemstruktur ist zu komplex

Sprache, Schrift, Szenen, DNA, ...

Maschine findet überlegene Lösungen

Greifende/balancierende Roboter ...

SYNERGIE von Mensch & Maschine

- ↳ Lernfähigkeit des Menschen
- ↳ Kopierfähigkeit des Rechners
- ↳ Lerngeschwindigkeit des Rechners

Ziele des Lernens

Lösung	genauer
Aufgabenbereich	breiter
Arbeitsweise	ökonomischer
Wissensstruktur	einfacher



Alan Turing

den Computer **erziehen!**

Was wird gelernt ?

Kognitionspsychologie des menschlichen (früh/kindlichen) Lernens

Struktur
Erwerb
Nutzung

Begriffe

Aggregation (Extension von Begriffen)

- Gruppieren von Objekten in Kategorien
- Sinnvolle Begriffe → Vorhersage von Objektverhalten

Charakterisierung (Intension von Begriffen)

- Gemeinsame Eigenschaften aller Instanzen eines Begriffs
- Welche Merkmale? ↗ kultureller/sprachlicher Kontext

Klassifikation

- Zuordnen eines Objekts zu „seiner“ Kategorie
- Einordnen in eine Hierarchie von Unter- und Oberbegriffen

Induktives Lernen

Verallgemeinerndes Lernen aus (endlich vielen) Beispielen

$$\gamma_A \triangleq A(x) \wedge A(y) \wedge A(z)$$

$$\gamma_B \triangleq B(x) \wedge B(y) \wedge B(z)$$

$$\gamma_V \triangleq \forall x (A(x) \Rightarrow B(x))$$

Deduktion

allgemein → speziell

(formallogisch korrekte Schlußweise)

$$\gamma_V, \gamma_A \vdash \gamma_B$$

Induktion

speziell → allgemein

(formallogisch unbeweisbarer, oft lebensnotwendiger Schluß)

$$\gamma_A, \gamma_B \vdash \gamma_V$$

Abduktion

Folgerung → hinreichende Voraussetzung

(formallogisch unbeweisbarer, oft unhaltbarer Schluß)

$$\gamma_V, \gamma_B \vdash \gamma_A$$

Induktives Lernen

Philosophisches Reizthema eines Jahrtausends

Francis Bacon (1561–1626)

Relevanz positiver *und* negativer Lernbeispiele



John Stuart Mill (1806–1873)

Vier Methoden für den praktischen Induktionsschluß



Bertrand Russell (1872–1970)

Induktionsschluß ist Grundlage jeglicher Vorhersage, nicht beweisbar und essentiell probabilistischer Natur



Ludwig Wittgenstein (1889–1951) Tractatus Logico-Philosophicus

„Suche das einfachste Gesetz, das mit den Fakten harmoniert“



William von Ockham (1285–1347)

Occam's Razor: „Pluralitas non est ponenda sine necessitate“



Jorma Rissanen (*1932) 'minimum description length'-Prinzip

MDL → minimale Summe codierender & korrigierender Bits

Paradigmen maschinellen Lernens

Der „Lehrer“ befiehlt / demonstriert / präsentiert / fehlt

Lernen aus Instruktionen

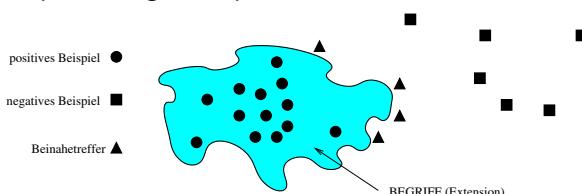
Natürlichsprachliche Systeme · Automatisches Programmieren

Lernen durch Analogiebildung

Wissentransfer auf neue, aber strukturell verwandte Aufgabenstellung

Lernen aus Beispielen (induktiv)

Beispiele, Gegenbeispiele und Beinahetreffer eines Begriffs



Lernen aus Beobachtung (explorativ)

Strukturieren von Objektmengen: $\left\{ \begin{matrix} \text{passiv} \\ \text{aktiv} \end{matrix} \right\} \triangleq \left\{ \begin{matrix} \text{Datenquelle = Prozeßbeobachtung} \\ \text{Interaktion Lernprogramm-Umwelt} \end{matrix} \right\}$

Konzeptuelles Lernen

Lernen eines Begriffs — wo kommen die benötigten Lernbeispiele (\pm) her ?

Assistiertes Lernen

Handverlesene Auswahl von \oplus/\ominus -Beispielen

⇒ Optimaler Lernerfolg durch kompetenten Reiseführer

Lernen mit Orakel

Lernprogramm wählt interessante neue Beispiele

Orakelbefragung liefert \oplus/\ominus -Information

⇒ Entdeckungsreise zu den Grenzfällen

Überwachtes Lernen

Beispiele wie vom natürlichen Erzeugungsprozeß produziert

Lehrer vergibt (die korrekten) \oplus/\ominus -Etiketten

⇒ Zufälliges Abrastern des Objektraums

Verstärkungslernen ('reinforcement learning')

Lernbeispiele liegen unetikettiert vor

Lehrer erteilt summarische Leistungsnote („Lob und Tadel“)

⇒ Strategie zwischen Exploration & Exploitation

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Beispiele induktiver Lernaufgaben

Aufgabenbereich · Leistungskriterium · Erfahrungsquelle

QUBIC (4 × 4 × 4 Tic Tac Toe)

AB — alle QUBIC-Partien gegen Bobby Fisher

LK — Prozentsatz aller gewonnenen Partien

EQ — die Möglichkeit, 3 Wochen gegen Fisher zu trainieren

Postanschriftenleser

AB — Erkenne Zielorte handgeschriebener Anschriften

LK — Prozentsatz korrekt sortierter Briefsendungen

EQ — 10^5 handadressierte Briefe mit bekanntem Zielort

Steuerung eines (auto-)mobilen Roboters

AB — selbständiges Manövriren im öffentlichen Fernverkehr

LK — Geschwindigkeit / $(1 + \text{Karambolagen})^{1.000.000}$

EQ — 20 Minuten Bewegtbilder mit Steuerkommandos

Natürlichsprachlicher Datenbankzugang

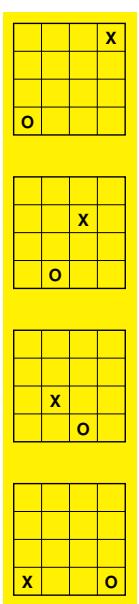
AB — autom. Beantwortung natürlichsprachlicher Datenbankanfragen

LK — Prozentsatz korrekter Antworten

EQ — Texte natürlichsprachlicher Benutzeranfragen nebst SQL-Kodierung

Beispiel QUBIC

Dreidimensionales Tic tac toe · Kubus mit $4^3 = 64$ Feldern



Zielfunktion $\text{eval}^* : \mathcal{B} \mapsto [-100, +100]$

$$\text{eval}^*(\mathbf{b}) = \begin{cases} +100 & \text{wenn 4 X in einer Reihe} \\ -100 & \text{wenn 4 O in einer Reihe} \\ 0 & \text{wenn Remisstellung erreicht} \\ \mathcal{E}[\cdot] & \text{Erwartungswert der Endstellung bei optimaler Strategie} \end{cases}$$

Lösungsmodell (lineare Näherung für eval^*)

$$\text{eval}(\mathbf{b}) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_{10} x_{10} =: \mathbf{w}^\top \mathbf{x}$$

mit den Prädiktorvariablen $x_i = x_i(\mathbf{b})$:

- $x_1(x_2) = \# \text{ offener Reihen mit einem X (O)}$
- $x_3(x_4) = \# \text{ offener Reihen mit zwei X (O)}$
- $x_5(x_6) = \# \text{ offener Reihen mit drei X (O)}$
- $x_7(x_8) = \# \text{ Schnittpunkte von X-Reihen (O-Reihen)}$
- $x_9(x_{10}) = \# \text{ Schnittpunkte s.o.; } \geq 2 \text{ X (O) je Reihe}$

Das Münchhausen-Prinzip

Was tun, wenn das Lösungsverfahren die Lösung selbst als Eingabe benötigt?

Problem

Woher bekommen wir die benötigten Werte

$$\text{eval}^*(\mathbf{b}_t) = ?$$

Lösung

Vorwärtssuche mit der Näherungsfunktion $\text{eval}(\cdot)$

$$\begin{aligned} \text{eval}^*(\mathbf{b}) &= \max\{\text{eval}^*(\mathbf{b}') \mid \mathbf{b}' \text{ Nachfolger von } \mathbf{b}\} \\ &\approx \max\{\text{eval}_w(\mathbf{b}') \mid \mathbf{b}' \text{ Nachfolger von } \mathbf{b}\} \end{aligned}$$

- Je besser die Näherung $\text{eval}(\cdot)$, desto genauer ist obige Approximation
- Wird dieses „bootstrapping“-Verfahren konvergieren?
- Welche Nachfolger von \mathbf{b} sollten betrachtet werden?
- Kann $\text{eval}^*(\cdot)$ überhaupt durch lineare Funktion angenähert werden?

Lernen der Stellungsbewertungsfunktion

Die Kenntnis von $\text{eval}^*(\cdot)$ ermöglicht eine optimale Zugauswahl

Benötigte Lernstichprobe

Partiestellungen $\mathbf{b}_1, \dots, \mathbf{b}_T$ mit bekannten Werten $y_t = \text{eval}^*(\mathbf{b}_t)$

Minimierung des Modellfehlers

Parameteroptimierung nach LSE-Prinzip („least squared error“)

$$\varepsilon = \sum_{t=1}^T (\underbrace{\text{eval}^*(\mathbf{b}_t)}_{\varepsilon_t} - \text{eval}(\mathbf{b}_t))^2$$

Iterative Lösung durch Gradientenabstieg

- 1 Initialisiere die Gewichte $w_0, w_1, w_2, \dots, w_{10}$
- 2 Führe je Lernbeispiel \mathbf{b}_t einen Verbesserungsschritt durch:

$$\mathbf{w}' = \mathbf{w} + \frac{2\beta \cdot (\text{eval}^*(\mathbf{x}_t) - \mathbf{w}^\top \mathbf{x}_t)}{\|\mathbf{x}_t\|^2}$$

Dabei bezeichnet β die **Lernrate** des Verfahrens.

Beispiel: Konzeptuelles Lernen

Unter welchen Witterungsbedingungen empfiehlt sich ein Segelturn?

GEGEBEN

- Objekte/Instanzen $\hat{=}$ mögliche Kalendertage
- Attribute/Prädikate $\hat{=}$ {sky, air, humidity, ...}
- Zielfunktion $\hat{=}$ gosailing : $\mathcal{X} \mapsto \{T, F\}$

Lerndaten

Objekte mit allen Attributwerten & der Begriffzugehörigkeit:

#	sky	air	humidity	wind	water	forecast	gosailing
1	sunny	warm	normal	strong	warm	same	T
2	sunny	warm	high	strong	warm	same	T
3	rainy	cold	high	strong	warm	change	F
4	sunny	warm	high	strong	cold	change	T

Beispiel: Konzeptuelles Lernen

Induktion als Versuch der Datenbeschreibung mit unzureichenden Mitteln

GESUCHT

Passende Hypothese $h \in \mathcal{H}$ aus geeignetem Repräsentationenraum.

- Hypothesenraum $\mathcal{H} \triangleq$ Konjunktionen von Attribut-Wert-Paaren
(z.B. $sky = sunny \wedge water = cool$)
- Lerndaten \triangleq positive und negative Beispiele
- Optimale Vorhersage der Urteile $gosailing(\cdot)$ durch h

Postulat des induktiven Lernens

Wenn Hypothese h approximiert Zielfunktion auf (großer)
Lernstichprobe

Dann Hypothese h approximiert Zielfunktion auf bislang
unbeobachteten Beispielen

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Repräsentationsformalismen

für **Datenobjekte** · zugrundeliegende **Begriffe** · gelernte **Hypothesen**

Parametersätze Diskriminanten, Neuronetze, Verteilungsfamilien

Formale Sprachen reguläre Ausdrücke, endliche Automaten, CFG

Produktionsregeln IF-THEN-Regeln, Assoziationen

Logik Aussagen-/prädikatenlogische Formeln, Klauselmengen

Graphen Semantische Netze, Drahtmodelle, Bayes/Markovnetze

Relationen Totale-, partielle- und Intervallordnungen

Frames Attribut-Wert-Paare, Dämonen, Defaults

Prozeduralformen Programme, Operatoren

Hierarchien Taxonomien, Partitionen, Entscheidungsbäume

Intensionale Repräsentationen

Endliche(!) formalsprachliche Beschreibung unendlicher(!) Gesamtheiten

Logische Formeln

$\text{elefant}(x) \Leftrightarrow \text{grau}(x) \wedge \text{groß}(x) \wedge \text{hat}(x, \text{Rüssel})$
 $\wedge \text{ist}(x, \text{nachtragend}) \wedge \neg \text{frißt}(x, \text{Rollmops})$

Programme, Algorithmen

```
proc prim (nat n) bool:
    for i from 2 to sqrt(n) do
        if mod(n,i) = 0 then return false fi
        od
    return true
```

Grammatiken

$$\begin{array}{lcl} S & \rightarrow & NP\ VP \\ NP & \rightarrow & N \mid Det\ N \\ VP & \rightarrow & V \mid VP\ NP \\ N & \rightarrow & John \mid Mary \\ V & \rightarrow & loves \end{array}$$

Räumliche Strukturen

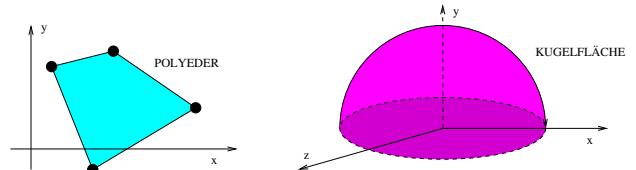
Kontinuum geometrischer Punkte als Lösung einer parametrisierten Gleichung

Polyeder

Drahtmodelle im \mathbb{R}^n :

$$(x_{(1)}, \dots, x_{(m)}) , \quad x_{(i)} \in \mathbb{R}^n$$

z.B. ein Viereck $((x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4))$, $x_i, y_i \in \mathbb{R}$, in der Ebene



Punkte auf einer Hyperfläche

z.B. auf einer \mathbb{R}^3 -Sphäre mit Radius r :

$$\mathbf{x} = (r \cos \theta, r \sin \theta, r \cos \omega) , \quad \theta, \omega \in [0, 2\pi]$$

Bäume

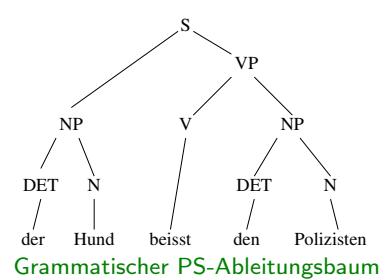
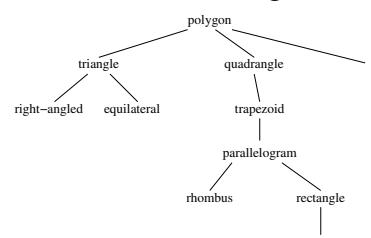
Zyklenfreie zusammenhängende ungerichtete Graphen bzw. ...

Definition

Der gerichtete Graph $\mathcal{G} = (U, L)$ heißt **Baum**, falls gilt:

1. \mathcal{G} ist einfach zusammenhängend.
2. Ex. genau ein **Wurzelknoten** $u_0 \in U$ ohne Vorgängerknoten.
3. Alle $u \in U \setminus \{u_0\}$ besitzen genau einen Vorgängerknoten.

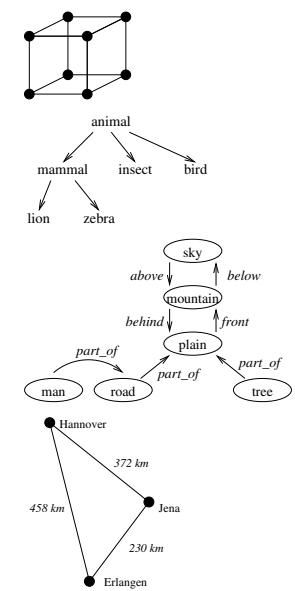
Knoten ohne Nachfolgerknoten heißen **Blattknoten**.



Taxonomie geometrischer Objekte

Graphen

Ungerichtet · Gerichtet · Markiert · Gewichtet



Ungerichteter Graph $\mathcal{G} = (U, L)$

$U \triangleq$ Knotenmenge

$L \triangleq$ Kantenmenge, $L \subseteq \{\{u, v\} \mid u, v \in U\}$

Gerichteter Graph $\mathcal{G} = (U, L)$

$U \triangleq$ Knotenmenge

$L \triangleq$ Kantenmenge,

$L \subseteq \{(u, v) \mid u, v \in U\} = U \times U$

Markierter Graph $\mathcal{G} = (U, L, \ell)$

$A \triangleq$ Symbolvorrat, Alphabet der Markierungen

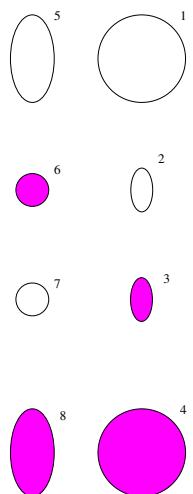
$\ell \triangleq$ Kantenmarkierungsfunktion, $\ell : L \mapsto A$

Gewichteter Graph $\mathcal{G} = (U, L, w)$

$w \triangleq$ Kantengewichtungsfunktion, $w : L \mapsto \mathbb{R}$

Listen

Geordnete Folge von (1) Listen oder (2) Symbolen aus Alphabet \mathcal{A}



Objektrepräsentationen

```
object 1: ( (shape circle) (size large) (color white) )
object 2: ( (shape ellipse) (size small) (color white) )
object 3: ( (shape ellipse) (size small) (color pink) )
object 4: ( (shape circle) (size large) (color pink) )
object 5: ( (shape ellipse) (size large) (color white) )
object 6: ( (shape circle) (size small) (color pink) )
object 7: ( (shape circle) (size small) (color white) )
object 8: ( (shape ellipse) (size large) (color pink) )
```

Verschachtelte Darstellungen

```
( (object1 ( (shape circle) (size large) (color white) )
  (object2 ( (shape ellipse) (size small) (color white) )
  (object3 ( (shape ellipse) (size small) (color pink) )
  (object4 ( ... ... ...
  .... ))
```

Spezialfälle

Bäume \triangleq Listen ohne Nachfolgerordnung

Zeichenketten \triangleq flache Listen

„Sein oder Nichtsein ...“ oder „GACTTTATAGCT...“

Logische Repräsentationen

Aussagenlogik · Prädikatenlogik · Modal- und Zeitlogik

Hornklausel

(Disjunktive) Klausel mit *höchstens einem* positiven Literal

$$\neg P_1 \vee \dots \vee \neg P_m \vee Q \quad \text{oder} \quad \neg P_1 \vee \dots \vee \neg P_m$$

Schreibweise: «Kopf» \leftarrow «Rumpf»

$Q \leftarrow P_1, P_2, \dots, P_m$	(allg.)
$\leftarrow P_1, P_2, \dots, P_m$	(Zielklausel)
$Q \leftarrow$	(Faktenklausel)
\leftarrow	(leere Klausel)

Beispiel

```
female(angela)
male(franz)
mutual_love(franz, angela)
can_marry(x1, x2)      ← mutual_love(x1, x2), female(x1), male(x2)
```

Prozedurale Repräsentationen

Imperative Formen · „if/then“-Regeln · Produktionsregeln

Beispiel

Imperative Darstellung einer Objektbeschreibung der Robotik:
„die kleine rote Schachtel steht auf der großen schwarzen Schachtel“

```
make_on (x,y) {
    cleartop (x);
    cleartop (y);
    puton (x,y);
}
puton (x,y) {
    STORE <on (x,y)>;
}
cleartop (x) {
    for all y DELETE <on (y,x)>;
}
```

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Was ist Data Mining ?

... und warum wird seit Beginn des Jahrtausends so viel darüber geredet ?

*„Data Mining is the exploration and analysis,
by automatic or semi-automatic means,
of large quantities of data
in order to discover meaningful patterns and rules.“*

Woher kommt der aktuelle Boom ?

- Massenproduktion von Daten
- Präsentation in *data warehouses*
- Rechnerleistung verfügbar
- Kommerzielle Datamining-Software erhältlich
- Starker Konkurrenzdruck

KDD — Knowledge Discovery in Databases

„We are drowning in information, but we are starving for knowledge.“ (John Naisbett 1996)

Was sind Daten?

- einzelne Objekte
 - individuelle Merkmale
 - riesige Fallzahlen
 - verwirrende Vielfalt
 - preiswert zu beschaffen
- \ominus Voraussagen

Tycho Brahe (1546–1601)

Massendatensammlung zu den Umlaufbahnen der Himmelskörper unseres Planetensystems
geozentrische Koordinaten

Was ist Wissen?

- Klassen von Objekten
 - globale Muster
 - allgemeine Gesetze
 - einfache Prinzipien
 - schwer zu bekommen
- \oplus Voraussagen

Johannes Kepler (1571–1630)

1. Umlaufbahnen sind elliptisch
2. Laufzeit \propto Sektorfläche
3. Umlaufperiode² \propto Großradius³

Typische Datenquellen

Industrielle Prozeßdaten

Analyse der Altpapieraufbereitung bei Kübler+Niethammer
8 Deinkingzellen à 54 Sensoren à 9000 Meßwerte/Tag \Rightarrow 3.888.000 Mw/T

Umsatzdatenbanken

Warenkorbanalyse für die Scannerkassen bei WalMart
20 Millionen Transaktionen/Tag \Rightarrow Datenbank 24 Terabytes

Molekularbiologie

Human Genome Database Project
Entschlüsselung des genetischen Codes des Menschen
60 000–80 000 Gene \Rightarrow 3 Milliarden DNA-Basen

Visuelle Daten

NASA *Earth Observing System* sammelt
Oberflächenbilder tieffliegender Satelliten \Rightarrow 50 Gigabytes/Stunde

Textinformationen

Ca. 10 Milliarden HTML-Seiten im *World Wide Web*
Suchmaschinen, Indexierer, Extrahierer, Emailfilter

Was ist das Analyseziel ?

Abstrakter Datensatz $\hat{=}$ Relation (Objekte \times Attribute)

Gruppierung

Partitionierung der Datenobjekte in Häufungsgebiete

Klassifikation

Zuordnung von Datenobjekten zu Kategorien

Dependenzstruktur

Aufdecken der Abhängigkeiten zwischen den Objektattributen

Prädiktion

Vorhersage (noch) nicht verfügbarer Objektattribute

Selektion und Assoziation

Erkennung von Auffälligkeiten & Regelmäßigkeiten

Anwendungsbedarf nach Industriezweigen

Großhandel · Finanzen · Telekommunikation · Verkehr · Gesundheit

Fälschungssicherheit

Mobilfunk — 'cloning' der Gerätetrennung
Kreditkartenmißbrauch — physikalisch/elektronisch
Rechnermißbrauch — Angriff, Einbruch

Kreditwesen

Kreditwürdigkeit, Zahlungsfähigkeit
Risikokapital, Unternehmenssolvenz
Anlageberatung

Kundenbetreuung

Kundenbindung (Beispiel: 5% Reduktion der Fluktuation \Rightarrow 200% Gewinn)
Direktmarketing (Handel, Bank, Versicherung)
Warenkorbanalyse im Einzelhandel

Beispiel Prozeßautomatisierung

Industrielle Herstellung von ICE-Türen aus Verbundwerkstoffen

Fertigungszelle

Prozeßkettenmodell $\hat{=}$ Workflow mit aktiven & passiven Komponenten:

- Meßwerte erfassen + auswerten  Sensoren
 - Stellgrößen berechnen + anlegen  Aktoren

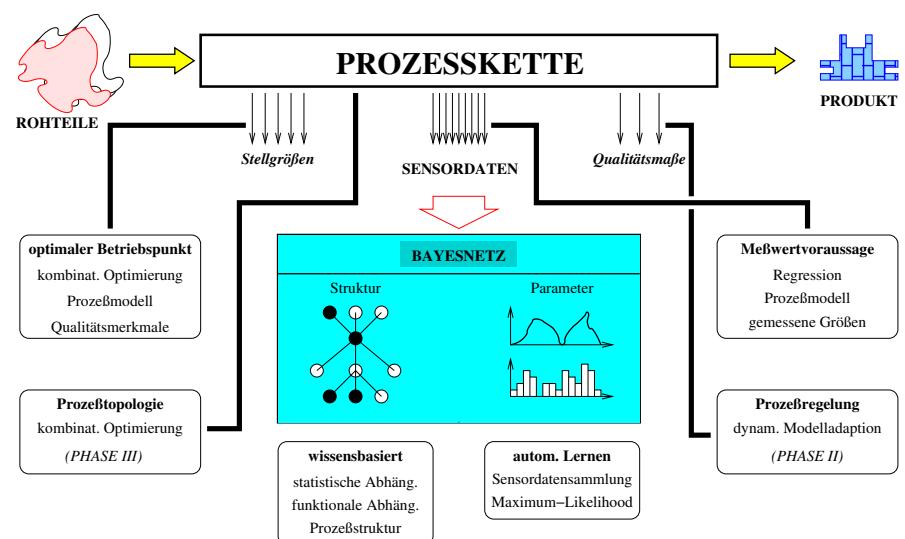
Produktionsoptimierung

Statt Erfahrung. Daumenregel und Intuition ...

- Prozeßvisualisierung
 - Entscheidungsunterstützung
 - Automatische (adaptive) Regelung
 - Optimale Strukturierung der Prozeßkette

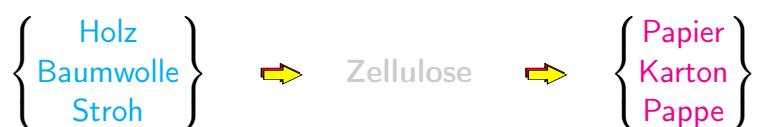
Beispiel Prozeßautomatisierung

Stochastischer Abhangigkeitsgraph zur Vorhersage optimaler Stellgroen



Beispiel Prozeßautomatisierung

Automatisierung in der Papierindustrie



Industrielle Arbeitsschritte

- | | |
|--------------------|---------------------------------|
| 1. Kocher | chemischer Aufschluß, Bleichung |
| 2. Flotationszelle | lösen, vorsortieren, entfärbten |
| 3. Refiner | Fasern mahlen |
| 4. Pulper | Wasser zusetzen (Suspension) |
| 5. Trockner | Bandsieb, Pressung (Tambouren) |
| 6. Cutter | zuschneiden, aufstapeln |

Prozeßdatenerhebung

Automatisierung in der Papierindustrie

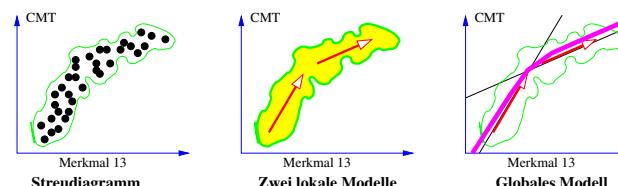
Zielgröße Papierqualität

Concora Medium Test

CMT $\stackrel{\text{def}}{=}$ „Gewicht“ / „Festigkeit“

26 Stellgrößen und Meßwerte

Druck, Temperatur, Menge, Gewicht, Qualität von Rohstoffen und Zwischenprodukten



Elliptotype-Cluster mit $x_{27} = 1.56 \cdot x_{13} + 0.32$ und $x_{27} = 0.60 \cdot x_{13} + 0.48$

Ablauf des Datamining-Prozesses

Automatisierung in der Papierindustrie

(Algorithmus)

0 LAUFZEITBEREINIGUNG

Transformation **physischer** Zeit t an Prozeßstation P_i via $\tau = t + \Delta t$; Meßwertvektoren $\tilde{x}_t \in \mathbb{R}^{27} \rightsquigarrow$ Fälle $x_\tau \in \mathbb{R}^{27}$ mit **synchronisierter** Referenzzeit

1 DATENSATZBEREINIGUNG

Ungültige Einträge markieren · Ausreißer nach 4σ -Regel markieren
Fälle mit markierten Werten tilgen

2 NORMIERUNG

Jedes der 27 Merkmale wird auf $\mathcal{N}(0, 1)$ normiert.

3 DEPENDENZANALYSE

Untersuche Abhängigkeiten der Form (x_i, x_{27}) und (x_i, x_j, x_{27}) .

4 REGRESSIONSANSATZ

Linear oder stückweise linear · zwei Elliptotype-Cluster

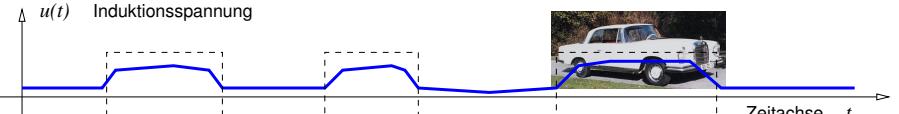
5 REGELERZEUGUNG

Überlagerung lokaler Modelle · Zugehörigkeitsfunktion \rightsquigarrow Regelprämisse

(summing up)

Beispiel Verkehrsplanung und -lenkung

Dienstgüteanalyse der Verkehrszustände auf Autobahnstrecken



Meßverfahren

- **Meßwertreihe $u(t)$**

Impulsfunktion der Induktionsschleife auf der Fahrbahn

Induktionsspannung

- **Verkehrsstärke q**

Zählung der Anzahl q von Impulsen (in $[1/h]$)

Fahrzeuge/Stunde

- **Streckenbelegung β**

Summe der Impulsbreiten $\beta = \frac{1}{u_{\max} \cdot \Delta T} \int_T^{T+\Delta T} u(t) dt$

Zeitanteil

- **Verkehrsdichte ρ**

$\rho \approx \rho_{\max} \cdot \beta$ und gleichzeitig auch $q \approx \bar{v} \cdot \rho$, aber ρ_{\max} und \bar{v} unbekannt

Fahrzeuge/Kilometer

Vernetzte Systeme

Datenanalyse in granularen Transportsystemen

Aufgabenstellungen

- **Monitoring** · Erfassung des aktuellen Zustandes
- **Modellierung** · Gesetzmäßigkeiten in Transportströmen
- **Prognose** · Vorhersage der Netzbelaistung
- **Routing** · Bestimmung optimaler Wege
- **Optimierung** · Verbesserung des Netzzustandes/Netzflusses

Anwendungsgebiete

- Güter- und Personenverkehr
- Telekommunikation
- Energieversorgung
- Rohstoffzufuhr im Fertigungsprozeß

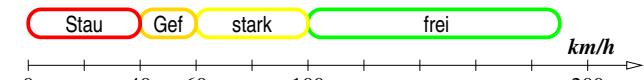
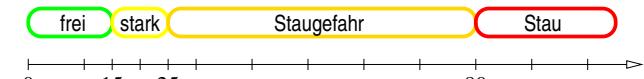
Beispiel Verkehrsplanung und -lenkung

Verkehrsflussmodell und Dienstgütestufen

Mathematisches Verkehrsflussmodell

Den Idealfall einer funktionalen Abhängigkeit $q(\rho) = v(\rho) \cdot \rho$ liefert:

$$v(\rho) = v_0 \cdot \rho \cdot \left(1 - (\rho / \rho_{\max})^{\ell-1}\right)^{\frac{1}{1-\ell}}$$

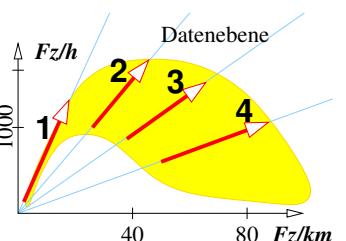
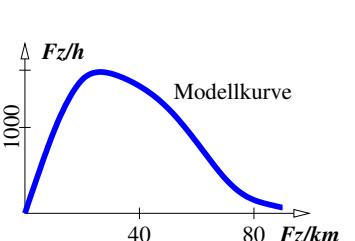


Dienstgütestufen („levels of service“)

- 1 freier Verkehr · 2 starker Verkehr · 3 Staugefahr · 4 Stau

Beispiel Verkehrsplanung und -lenkung

Modellierung und Interpretation der Meßdatensätze



Tagesgangkurven

Viertelstündige Verkehrsstärkemessung

Mediangußtung · Datensätze für Wochenkerntage

Clustering in drei prototypische Gruppen:

1 Urlaubstag · 2 Durchschnittstag · 3 Großveranstaltungstag

96 Werte/Tag

$M = 5$; Mo,Di,Mi,Do

Struktur der (ρ, q) -Datenebene

Konzentrische Geradenstücke $\hat{=}$ Verkehrssituationen gleicher Geschwindigkeit

4 Dienstgüten \Rightarrow konzentrische Längscluster

Beispiel Marketing

Welche Datamining-Methoden für welche Fragestellung ?

Segmentierung

Welche Idealtypen von Kunden besitzt die Firma?

Klassifikation

Ist die konkrete Person ein potentieller Neukunde?

Konzeptualisierung

Welche Attribute charakterisieren ein Kundensegment?

Prädiktion

Welcher Umsatz ist im Folgejahr zu erwarten?

Deviation

Wo und warum ist Kundenverhalten verändert?

Dependenz

Wie beeinflußt eine Marketingaktion das Kundenverhalten?

Beispiel Marketing

Aktive Orientierung an Kundenwünschen \rightsquigarrow Wettbewerbsvorteil

Relationale Datenbank eines Versandhauses

Kundentabelle

KuNr, PLZ, GJ (Geburtsjahr), ...

Umsatztabelle

BestNr, KuNr, Betrag, ...

Datamining-Schritte

Clusteranalyse der Verbundtabelle

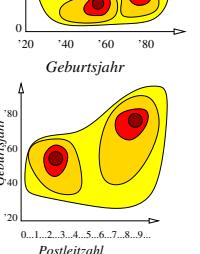
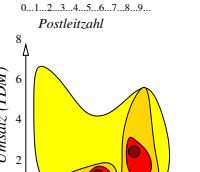
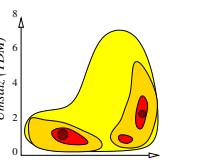
$$(PLZ, GJ, Umsatz) \in \mathbb{R}^3$$

Gewichteter euklidischer Abstand
 $\mathbf{g} = (10^{-5}, 10^{-2}, 10^{-4})$

$$\mu^{(1)} = \begin{pmatrix} 27374 \\ 1954.16 \\ 1122.44 \end{pmatrix}, \quad \mu^{(2)} = \begin{pmatrix} 86356 \\ 1969.35 \\ 1618.99 \end{pmatrix}$$

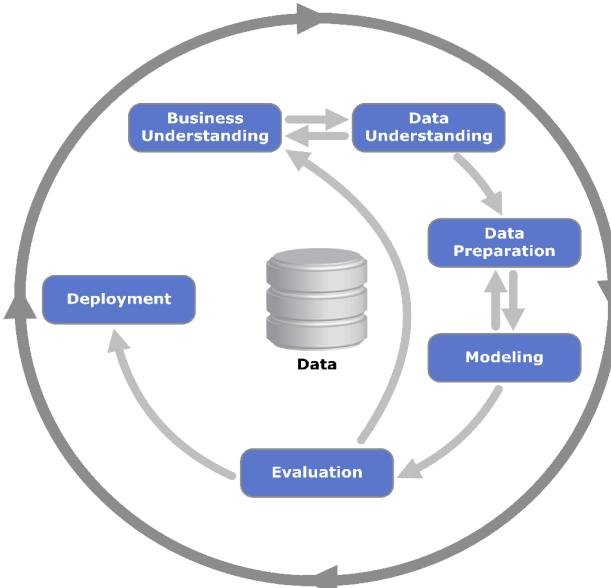
Risiken und Nebenwirkungen

„Alter“ \Rightarrow „Geburtsdatum“ \Rightarrow „1.1.1970“



Cross-Industry Standard Process for Datamining

CRISP-DM (NCR & Daimler & SPSS/IBM)



SEMMA (SAS)

Sample
Explore
Modify
Model
Assess

WEKA et al.

data acquisition
data preprocessing
data modeling
data evaluation

Datamining-Projekte

Arbeitsphasen & Grundbausteine eines Datamining-Prozesses

Materialbeschaffung (I)

Planung
Datensammlung
Merkmalsberechnung
Datenauswahl

Vorverarbeitung (II)

Normierung
Säuberung
Filterung
Ergänzung
Korrektur



Auswertung (IV)

Visualisierung
Interpretation
Dokumentation

Strukturanalyse (III)

Korrelation
Regression
Modellierung
Klassifikation
Gruppierung



(Kommerzielle) Softwaresysteme

Allroundpakete — nicht anwendungsspezifisch, viele Werkzeuge

Paket (Anbieter)	Implementierte Methoden
Clementine (SPSS & IBM)	EB Reg MLP Rul kNN SOM Clus
Enterprise Miner (SAS)	EB Reg MLP Rul Seq Clus
Darwin (Thinking Machines)	EB MLP kNN
WEKA (OSS/FSW) 'R'-Projekt (OSS/FSW)	EB Reg MLP Rul SOM Clus ... das alles und noch viel mehr ...

EB Statistische Entscheidungsbäume (CART)

Reg Regressionsmodelle für Vorhersage & Kategorisierung

MLP Mehrschichtenperzeptron

Rul Assoziations- und Fuzzyregelsysteme

kNN *k*-nächster-Nachbar Klassifikation

SOM Selbstorganisierende Merkmalkarten

Clus (Hierarchische) Gruppierungsverfahren

Seq Statistische Zeitreihenanalyse

Kommerzielle Softwaresysteme

Anwendungsspezifische Werkzeuge — integrierte Speziallösungen

Fälschungsschutz

HNC Falcon/Eagle, Neuraltech Nestor/Minotaur, Nestor

Kreditkontrolle

FairIsaac, Sigma Analytics, Neuraltech Decider

Kundenbindung

SLP InfoWare, Neuraltech Churn Manager

Kundenprofil

HNC ProfitMax, Neuraltech Gold, RightPoint, AppliedMetrix

Kommerzielle Softwaresysteme

Methodenspezifische Werkzeuge — die Welt sieht aus wie ein Nagel ...

(Tiefe) Neuronale Netze

PittNet, NN/XNN, SNNS; TensorFlow, Caffe, Torch

Nächster-Nachbar-Klassifikator

SGI MLC++, Condor PEMLS

Abhängigkeitsanalyse

SGI MineSet, XPertRule Miner

Graphische Modelle

LEDA, LINK, ViCLAS, Precision Crimelink

Was ist (maschinelles) Lernen ?

Beispielanwendungen

Repräsentationsformalismen

Data Mining

Zusammenfassung

Zusammenfassung (1)

1. **Maschinelles Lernen** verknüpft empirische *Beobachtungen*, menschliches *Vorwissen* und überlegene *Rechnerleistung* zu einer neuen Qualität intelligenter Informationsverarbeitung.
2. **Induktives Lernen**, die Verallgemeinerung auf Basis von Einzelfällen, ist eine unverzichtbare, gleichwohl unbeweisbare Schlußtechnik.
3. Die **Lernbeispiele** zu einem **Begriff** und ihre **Etikettierung** werden vom **Lehrer** und/oder dem **Lernprogramm** vorgegeben.
4. Die Frage nach einer (geeigneten) **Repräsentation** stellt sich bei den präsentierten **Datenobjekten**, den zugrundeliegenden **Begriffen** („*Konzepten*“) und den zu lernenden **Hypothesen**.
5. Die Objektrepräsentation umfasst **numerische**, **symbolische**, **prozedurale**, **relationale** und **metrisch-topologische** Darstellungen.
6. Zur Lösung der Lernaufgabe wird ein **Erfolgskriterium** optimiert.
7. **Datamining** ist die (oft interaktive) Anwendung von ML-, Statistik- und Visualisierungsmethoden auf **große Datenbestände**.
8. Das Anliegen ist das Aufdecken von **Gruppenstrukturen** und **Abhängigkeiten**, das Ermitteln von **Kategoriezugehörigkeiten** sowie Vorhersage und Abgleich zukünftiger oder unzugänglicher **Attributwerte**.
9. Datamining ist ein **zyklischer Prozess** der Schritte **Akquisition**, **Bereinigung**, **Modellierung** und **Evaluierung**.

MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2020

Prof. E.G. Schukat-Talamazzini

Stand: 10. September 2020

Teil II

Datenaufbereitung

Werteskalen Relationen Skalenkonversion Ausreißer Imputation Σ

Grundbegriffe des Data Mining

Datensätze mit expliziter oder impliziter Objektcharakterisierung

Datensatz

Menge oder **Folge** von **Objekten** („*Instanzen*“) des Aufgabenbereichs Ω mit ihren Eigenschaften und/oder Beziehungen

Attribut

Objekteigenschaft $\hat{=}$ Element eines **Wertebereichs** \mathcal{X} („*Skala*“)

Beziehung

Relation $\mathcal{R} \subset \Omega \times \Omega$ zwischen Objekten oder ...

Abstand/Ähnlichkeit $d : \Omega \times \Omega \rightarrow \mathbb{R}$ zwischen Objekten

	o_1	o_2	o_3	o_4
o_1		∞		
o_2			∞	
o_3	∞			
o_4		∞		

	o_1	o_2	o_3	o_4
o_1	0	3	8	15
o_2	3	0	5	12
o_3	8	5	0	7
o_4	15	12	7	0

	x_1	x_2	x_3	x_4
o_1	+	low	1.2	+4
o_2	-	hi	0.5	-3
o_3	+	hi	2.3	-3
o_4	+	med	2.1	-7

Werteskalen Relationen Skalenkonversion Ausreißer Imputation Σ

Datenmatrix $\hat{=}$ Objekte \times Attribute

Objekteigenschaften erster Ordnung

Definition

Sind $\mathcal{X}_1, \dots, \mathcal{X}_N$ die Attribute eines Objekts, so heißen die Elemente

$$\mathbf{x} = (x_1, \dots, x_N)^\top \in \mathcal{X}_1 \times \dots \times \mathcal{X}_N =: \mathcal{X}$$

Datenvektoren

des Objekts.

Die Menge \mathcal{X} heißt **Wertebereich** des Objekts.

Eine (Multi-)Menge $\{\mathbf{x}_1, \dots, \mathbf{x}_T\} \subset \mathcal{X}$ oder eine Folge $(\mathbf{x}_1, \dots, \mathbf{x}_T) \in \mathcal{X}^T$ bezeichnen wir als **Datenmatrix** oder ggf. als **Meßreihe**.

Schreibweise

Datenvektoren

Meßwerte

$$\text{Reelle Datenmatrix } \begin{pmatrix} x_{1,1} & \dots & x_{1,N} \\ \vdots & \ddots & \vdots \\ x_{T,1} & \dots & x_{T,N} \end{pmatrix} \quad \mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)^\top \in \mathbb{R}^{T \times N}$$

Variable

$$\left\{ \begin{array}{l} \text{Typ} \\ \text{Name} \\ \text{Wert} \end{array} \right\}$$

Datum

$$\left\{ \begin{array}{l} \text{Attribut} \\ \text{Objekt} \\ \text{Eintrag} \end{array} \right\}$$

$$\mathbf{x}_t = (x_{t,1}, \dots, x_{t,N})^\top$$

Werteskalen

Relationen und Distanzen

Skalenkonversion

Detektion von Ausreißern

Imputation von Fehlanzeigen

Zusammenfassung

Attribute und ihre Skalentypen

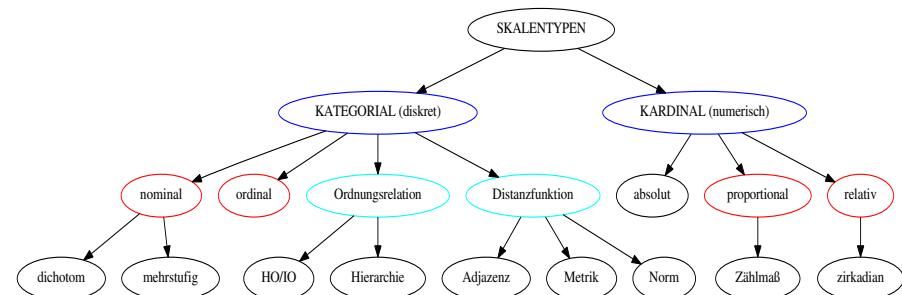
Was bedeuten die Spalteneinträge einer Datenmatrix ?

Beispiel

	vorbestraft	Partei	Abinote	Geburt	Spenden
Angela	F	CDU	gut	1954	$345 \cdot 10^3$
Guido	F	FDP	ausreichend	1961	$137 \cdot 10^3$
Roland	T	CDU	gut	1958	$3.6 \cdot 10^6$
Gregor	F	PDS	sehr gut	1948	NA
Linus	F	Pirat	NA	1969	0
Bill	F	Rep	mangelhaft	1955	$-4.2 \cdot 10^9$
Roman	T	1933	...
:	:	:	:	:	:
:					
	$\{T, F\}$	$\{\pi_1, \dots, \pi_9\}$	$\{\nu_1, \dots, \nu_5\}$	$\mathbb{Z} \subset \mathbb{R}$	\mathbb{R}

Skalentypen

Objektattribut $\hat{=} (\text{Wertebereich}, \text{Operatorenmenge})$



Diskrete Skala

Endlicher Wertebereich

$$\mathcal{X} = \{\xi_1, \xi_2, \dots, \xi_K\}$$

Numerische Skala

Kontinuierlicher Wertebereich

$$\mathcal{X} \subseteq \mathbb{R}$$

Typische Wertebereiche

Nominalskala

- Dichotomien
 $\{0, 1\}, \{T, F\}, \{+, -\}, \{m, f\} \dots$
- Zeichensätze
 $\{C, G, A, T\}$
- Farben
„red“, green, blue)
- Gruppen & Prädikate

Ordinalskala

- Notenskala
„sehr gut“, „gut“, „befriedigend“, ...
- Unscharfe Prädikate
„kalt“, „kühl“, „lau“, „warm“, „heiß“
- Eingefrorene Quantitäten
„2-türig“, „4-türig“, „5-türig“

Intervallskala (relativ)

- Temperaturen
 $20^\circ\text{C}, 451^\circ\text{F}$
- Zeitangaben
1066, 2001/09/11, 469 v.Chr., ...

Verhältnisskala (absolut/proport.)

- absolut. Temperatur
 273°K
- Dauer
 $45 \text{ min}, 13.7 \cdot 10^9 \text{ Jahre}$
- Mengenangaben
 $C = 2.98, 17 \text{ cm}, 8 \mu\text{g}, \dots$

Binäre Skalenoperationen

Vergleichsoperationen $\mathcal{X} \times \mathcal{X} \rightarrow \{T, F\}$ · Rechenoperationen $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

Nominalskala

= Gleichheitstest

Alle Attributwerte ξ_ℓ sind gleichberechtigt.

Ordinalskala

\prec Vergleichbarkeit

Abschnittsbildung nach Totalordnung:
 $\{\xi | \xi \preceq \xi_\ell\}$

Intervallskala (relativ)

- Differenzbildung

Unterschiede sind durch $x_1 - x_2$ quantifizierbar.

Verhältnisskala (absolut/proport.)

\div Quotientenbildung

Wohldefiniert:
„Nullpunkt“, „doppelt“, „Drittel“

Durchschnittswerte

Wie berechnet man/frau einen für $(x_1, \dots, x_T) \in \mathcal{X}^T$ „(proto)typischen“ Wert ?

Nominalskala

Modus — der häufigste Wert:
 $\mu^{\text{mod}} = \xi_{\ell^*}$ mit

$$\ell^* = \operatorname{argmax}_\ell N_\ell$$

mit den absoluten Häufigkeiten
 $N_\ell = \sum_{t=1}^T \delta_{x_t, \xi_\ell}$

Ordinalskala

Median — der mittlere Wert:
 $\mu^{\text{med}} = \xi_{\ell^*}$ mit

$$\sum_{k=1}^{\ell^*-1} N_k \leq \frac{T}{2} \leq \sum_{k=1}^{\ell^*} N_k$$

falls das Inventar \mathcal{X} geordnet ist:
 $\xi_1 < \xi_2 < \dots < \xi_\ell < \xi_{\ell+1} < \dots < \xi_L$

Intervallskala (relativ)

Arithmetisches Mittel

$$\mu^{\text{mean}} = \frac{1}{T} \cdot \sum_{t=1}^T x_t = \frac{1}{T} \cdot \sum_{\ell=1}^L N_\ell \cdot \xi_\ell$$

Verhältnisskala (absolut/proport.)

Geometrisches Mittel

$$\mu^{\text{geo}} = \sqrt[T]{\prod_{t=1}^T x_t} = \sqrt[T]{\prod_{\ell=1}^L \xi_\ell^{N_\ell}}$$

Durchschnittswerte

Verallgemeinerung auf Metriken und normierte Vektorräume

Beispiel

Für die Wertemenge $\{1, 1, 1, 2, 2, 5, 9\}$ gilt:

$$\mu^{\text{mod}} = 1, \quad \mu^{\text{med}} = 2, \quad \mu^{\text{mean}} = 3, \quad \mu^{\text{geo}} = 2.0998$$

Definition

In einem metrischen Raum (\mathcal{X}, d) heißt der Wert

$$\mu^{\text{zen}} = \operatorname{argmin}_{z \in \mathcal{X}} \left(\sum_{t=1}^T d(z, x_t) \right)$$

das **Zentroid** der (Multi-)Menge $\{x_1, \dots, x_T\}$.

Lemma

- (1) Es ist $\mu^{\text{mean}}(\cdot)$ das Zentroid zur euklidischen Metrik $d(y, z) = (y - z)^2$.
- (2) Es ist $\mu^{\text{med}}(\cdot)$ das Zentroid zur Betragsmetrik $d(y, z) = |y - z|$.
- (3) Es ist $\mu^{\text{mod}}(\cdot)$ das Zentroid zur diskreten Metrik $d(y, z) = 1 - \delta_{y,z}$.

Durchschnittswerte

Verallgemeinerung von (endlichen) Wertemengen auf diskrete Verteilungen

Definition

Es sei \mathbb{X} eine diskrete Zufallsvariable über dem Wertebereich $\mathcal{X} \subset \mathbb{R}$ mit der Wahrscheinlichkeitsfunktion $P(\cdot)$. Dann heißt

$$\mu(\mathbb{X}) = \mathbb{E}[\mathbb{X}] \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} x \cdot P(\mathbb{X} = x)$$

der **Erwartungswert** von \mathbb{X} , es heißt

$$\mu^{\text{med}}(\mathbb{X}) = \xi \quad \text{mit} \quad \sum_{x < \xi} P(\mathbb{X} = x) \leq \frac{1}{2} \leq \sum_{x \leq \xi} P(\mathbb{X} = x)$$

der **Median** von \mathbb{X} , und es heißt

$$\mu^{\text{mod}}(\mathbb{X}) \stackrel{\text{def}}{=} \operatorname{argmax}_{x \in \mathcal{X}} P(\mathbb{X} = x)$$

der **Modus** von \mathbb{X} .

Durchschnittswerte

Verallgemeinerung von (endlichen) Wertemengen auf stetige Verteilungen

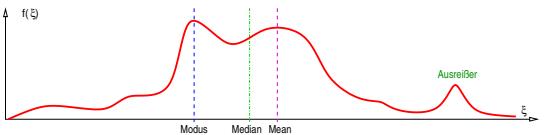
Definition

Für eine kontinuierliche Zufallsvariable über dem Wertebereich $\mathcal{X} = \mathbb{R}$ mit der Wahrscheinlichkeitsdichtefunktion $f_{\mathbb{X}}(\cdot)$ gilt entsprechend:

$$\begin{aligned}\mu(\mathbb{X}) &\stackrel{\text{def}}{=} \mathbb{E}[\mathbb{X}] = \int_{\mathbb{R}} x \cdot f_{\mathbb{X}}(x) dx \\ \mu^{\text{med}}(\mathbb{X}) &\stackrel{\text{def}}{=} \xi \quad \text{mit} \quad \int_{-\infty}^{\xi} f_{\mathbb{X}}(x) dx = \frac{1}{2} \\ \mu^{\text{mod}}(\mathbb{X}) &\stackrel{\text{def}}{=} \underset{x \in \mathbb{R}}{\operatorname{argmax}} f_{\mathbb{X}}(x)\end{aligned}$$

Bemerkung

Die Mediandefinition erfordert eine stetige und streng monotone Wahrscheinlichkeitsverteilungsfunktion.



Relationen auf diskreten Attributen

Spezialfall: Objekte besitzen genau ein Attribut \mathcal{X}

Adjazenz

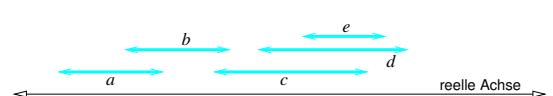
Die Matrix $A \in \{0,1\}^{L \times L}$ repräsentiert eine **(Objekt)nachbarschaft**.

- räumliche Nähe, Verwandtschaft, Interaktion ...
- „Elter-von“, Einflußnahme, ...

Präferenz

Die Relation $\mathcal{R} \subset \mathcal{X} \times \mathcal{X}$ repräsentiert eine (nicht notwendig totale) **Ordnung**.

- Halbordnung, Verband, Boolesche Algebra
- Turnier, (echte) Intervallordnung



Bemerkung

Zyklus: $c \preceq b \prec d \preceq c$
 \neg -transitiv: $c \preceq b \preceq a$

Ordnung ist das halbe Leben ...

Relationeneigenschaften zur Charakterisierung ausgefilterter Ordnungsbegriffe

Binäre Relationen

Infixschreibweise für Relationen:

Unkonventionell (nicht kleiner/gleich):

R reflexiv	\Leftrightarrow	$\forall a: a \prec a$
R irreflexiv	\Leftrightarrow	$\forall a: a \not\prec a$
R symmetrisch	\Leftrightarrow	$\forall a, b: a \prec b \Rightarrow b \prec a$
R antisymmetrisch	\Leftrightarrow	$\forall a, b: a \prec b \wedge b \prec a \Rightarrow a = b$
R asymmetrisch	\Leftrightarrow	$\forall a, b: a \not\prec b \vee b \not\prec a$
R vollständig	\Leftrightarrow	$\forall a \neq b: a \prec b \vee b \prec a$
R streng vollständig	\Leftrightarrow	$\forall a, b: a \prec b \vee b \prec a$
R transitiv	\Leftrightarrow	$\forall a, b, c: a \prec b \prec c \Rightarrow a \prec c$
R negativ transitiv	\Leftrightarrow	$\forall a, b, c: a \not\prec b \not\prec c \Rightarrow a \not\prec c$
R semitransitiv	\Leftrightarrow	$\forall a, b, c, e: a \prec b \prec c \Rightarrow a \prec e \vee e \prec c$
R Ferrer-transitiv	\Leftrightarrow	$\forall a, b, c, d: a \prec b, c \prec d \Rightarrow a \prec d \vee c \prec b$

... nach der Ordnung wollen wir streben.

M. Roubens, Ph. Vincke: „Preference Modelling“ (1985)

Einige ausgewählte Typen von Ordnungsrelationen

ORDNUNGSBEGRIFF	SYMMETRIE	VERGLEICH	TRANSFER	REFLEX
Turnierordnung	asymmetrisch	vollständig		
totale Ordnung	antisymmetrisch	streng vollständig	transitiv	
strenge totale O.	asymmetrisch	vollständig	transitiv	
schwache Ordnung		streng vollständig	transitiv	
strenge schwache O.	asymmetrisch		negativ transitiv	
schwache Halbordn.	asymmetrisch		transitiv	
Halbordnung	antisymmetrisch		transitiv	reflexiv
Quasiordnung			transitiv	reflexiv
Äquivalenzrelation	symmetrisch		transitiv	reflexiv
Intervallordnung	antisymmetrisch	Ferrer-transitiv	transitiv	irreflexiv
Semiordnung	antisymmetrisch	semitransitiv	transitiv	irreflexiv

Definition

Eine Abstandsfunktion $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{IR}$ heißt **Metrik** auf \mathcal{X} , wenn $d(\cdot, \cdot)$ für alle $x, y, z \in \mathcal{X}$ die drei Eigenschaften

$$1. d(x, y) = 0 \Leftrightarrow x = y$$

$$2. d(x, y) = d(y, x)$$

$$3. d(x, z) \leq d(x, y) + d(y, z)$$

besitzt.

Definitheit

Symmetrie

Dreiecksungleichung

Bemerkungen

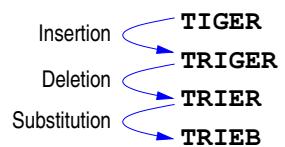
1. Jede Vektorraumnorm $\|\cdot\| : \mathcal{X} \rightarrow \mathbb{IR}$ definiert eine Metrik $d(x, y) = \|x - y\|$.
2. Jedes innere Produkt $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{IR}$ definiert eine VR-Norm $\|x\| = \sqrt{\langle x, x \rangle}$.
3. Distanzen transformieren in Ähnlichkeiten $s(x, y) = \exp(-d(x, y)/2\sigma^2)$.
4. Ähnlichkeiten transformieren in Distanzen $d(x, y) = -2\sigma^2 \cdot \log(s(x, y))$.

Spezialfall Zeichenketten

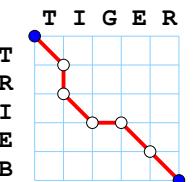
Attribute mit einem diskreten Wertebereich $\mathcal{X} \subset \mathcal{A}^*$

Elementare Operationen auf Zeichenketten

- Ersetzung eines Zeichens durch ein anderes
- Löschung eines Zeichens
- Einfügung eines Zeichens



substitution
deletion
insertion



Definition

Ist \mathcal{A} ein endliches Alphabet und sind v, w zwei Zeichenfolgen aus \mathcal{A}^* , so bezeichnet der **Levenshtein-Abstand** $d^{\text{lev}}(v, w)$ die minimale Anzahl von Elementaroperationen, mit denen v in w überführt werden kann.

Spezialfall Zeichenketten

Zeichenkettenattribute sind metrisch und erlauben die Durchschnittsbildung

Lemma

Der Levenshtein-Abstand $d^{\text{lev}} : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{IR}$ über dem Alphabet \mathcal{A} ist eine definite, symmetrische Distanzfunktion und erfüllt die Dreiecksungleichung — $(\mathcal{A}^*, d^{\text{lev}})$ ist folglich ein **metrischer Raum**.

Definition

Sei (\mathcal{X}, d) ein metrischer Raum und $(w_1, \dots, w_T) \in \mathcal{X}^T$ eine Auswahl (Multimenge) von Elementen. Der Wert

$$\mu^{\text{mid}} = \underset{z \in \{w_1, \dots, w_T\}}{\operatorname{argmin}} \left(\sum_{t=1}^T d(z, w_t) \right)$$

heißt das **Medoid** der Menge bezüglich der Metrik $d(\cdot, \cdot)$.

Bemerkung

Das Medoid einer Wortmenge w_1, \dots, w_T mit maximaler Wortlänge N_{\max} lässt sich mit Aufwand $O(T^2 N_{\max}^2)$ berechnen.

Werteskalen

Relationen und Distanzen

Skalenkonversion

Detektion von Ausreißern

Imputation von Fehlanzeigen

Zusammenfassung

Konversion der Attributskalen — wozu ?

Datensatz mit Attributen unterschiedlichen Skalentyps

Traditionelle Modellierungsverfahren erfordern einheitliche Skalen:

- **Numerische Skalen**
Multivariate Normalverteilung

$$\mathcal{X} = \underbrace{\mathbb{R} \times \dots \times \mathbb{R}}_{N\text{-mal}} = \mathbb{R}^N$$

$$f(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \mathbf{S}) = |2\pi\mathbf{S}|^{-1/2} \cdot \exp \left\{ -\frac{1}{2} \cdot (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

- **Diskrete Skalen**
 N -dimensionale Wahrscheinlichkeitstabelle
mit $\mathcal{X}_n = \{1, \dots, \ell_n\}$

$$P(\mathbf{x}) = p_{x_1, \dots, x_N} \quad \text{mit dem Tensor} \quad \mathbf{p} \in [0, 1]^{\ell_1 \times \ell_2 \times \dots \times \ell_N}$$

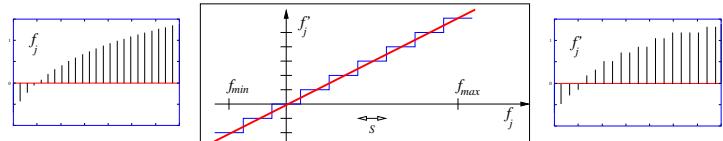
Option auf robusteres Datenmodell

Sind die Attributwerte wirklich normalverteilt?

Kann ich mir eine Tabelle mit $\prod_n \ell_n$ Einträgen leisten?

Diskretisierung numerischer Attribute (unüberwacht)

(kardinal \Rightarrow ordinal)



Äquidistante Intervalle

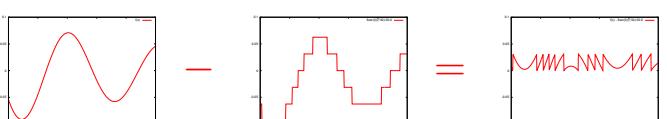
Mißachtet Datenverteilung \rightsquigarrow unglm. Zellenbesetzung & Übersteuern

Äquifrequente Intervalle

Histogrammegalisierung \rightsquigarrow konstante Zellenbesetzung T/L

Nichtlinearer Skalarquantisierer

Minimiert den Störabstand (SNR): mittlerer quadratischer Quantisierungsfehler



Faustregel
 $L = \sqrt{T}$

Nominalisierung ordinaler Attribute

(ordinal \Rightarrow nominal)

Problem

Die Quantisierung numerischer Skalen liefert konstruktionsbedingt Werte einer **ordinalen Skala**.

Die immanente Reihenfolgeinformation wird aber von den einschlägigen Datenmodellen (W-Tabellen, lineare Modelle, Entscheidungsbäume) nicht genutzt.

Ordinalle Entflechtung

Die ordinale Skala mit (sortiertem) Wertebereich $\mathcal{X} = \{\xi_1, \dots, \xi_\ell\}$ wird auf einen Komplex **binärer** Attribute $\mathcal{X}_i = \{0, 1\}$, $i = 1, \dots, \ell - 1$, abgebildet:

$$\phi(\xi_j) = (\underbrace{0, \dots, 0}_{(j-1)\text{-mal}}, \underbrace{1, \dots, 1}_{(\ell-j)\text{-mal}}) \in \{0, 1\}^{\ell-1}$$

Für jede \mathcal{X} -Stufe $\xi - j$ gilt also:
 $\phi_i(\xi_j) = 1 \Leftrightarrow i \geq j$.

Fall $\ell = 3$

\mathcal{X}	ξ_1	ξ_2	ξ_3
ξ_1	1	1	1
ξ_2	0	1	1
ξ_3	0	0	1

Fall $\ell = 5$

\mathcal{X}	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
ξ_1	1	1	1	1	1
ξ_2	0	1	1	1	1
ξ_3	0	0	1	1	1
ξ_4	0	0	0	1	1
ξ_5	0	0	0	0	0

Kardinalisierung nominaler Attribute

(nominal \Rightarrow numerisch)

Problem

Zahlreiche Methoden (k -nächste-Nachbarn, Bayesregel, Trennfunktionen) der Klassifikation und Vorhersage benötigen Objektabstände oder numerische, besser noch gaußverteilte Objektattribute.

Nominale Entflechtung

Die nominale Skala mit Wertebereich

$\mathcal{X} = \{\xi_1, \dots, \xi_\ell\}$ wird auf einen Komplex

reellwertiger Attribute $\mathcal{X}_i = \{0, 1\}, i = 1, \dots, \ell$, abgebildet:

$$\phi(\xi_j) = (\underbrace{0, \dots, 0}_{(j-1)\text{-mal}}, 1, \underbrace{0, \dots, 1}_{(\ell-j)\text{-mal}}) \in \mathbb{R}^\ell$$

Für diese Darstellung gilt die Äquidistanzeigenschaft

$$d(\phi(\xi_i), \phi(\xi_j)) = \|\phi(\xi_i) - \phi(\xi_j)\| = \begin{cases} 0 & \xi_i = \xi_j \\ \sqrt{2} & \xi_i \neq \xi_j \end{cases}$$

Fall $\ell = 5$

\mathcal{X}	$\mathbb{R}\mathbb{I}_1$	$\mathbb{R}\mathbb{I}_2$	$\mathbb{R}\mathbb{I}_3$	$\mathbb{R}\mathbb{I}_4$	$\mathbb{R}\mathbb{I}_5$
ξ_1	1	0	0	0	0
ξ_2	0	1	0	0	0
ξ_3	0	0	1	0	0
ξ_4	0	0	0	1	0
ξ_5	0	0	0	0	1

Kontrastmatrizen

Auch im $\mathbb{R}^{\ell-1}$ ist genug Platz für ξ_1, \dots, ξ_ℓ

Ursprung & Einheiten

Einer-gegen-alle: treatment

ξ_1	0	0	0	0
ξ_2	1	0	0	0
ξ_3	0	1	0	0
ξ_4	0	0	1	0
ξ_5	0	0	0	1

Spaltenmittelwertfrei

Einer-gegen-alle: sum

ξ_1	-1	-1	-1	-1
ξ_2	1	0	0	0
ξ_3	0	1	0	0
ξ_4	0	0	1	0
ξ_5	0	0	0	1

Distanzen 0, $\sqrt{2}$, aber auch 1

Gestaffelt

Gegen-Anfangsparte: helmert

ξ_1	-1	-1	-1	-1
ξ_2	1	-1	-1	-1
ξ_3	0	2	-1	-1
ξ_4	0	0	3	-1
ξ_5	0	0	0	4

Distanzen 0, $\sqrt{2}$, aber auch $\sqrt{\ell+2}$

Äquidistant

Orthonormalpolynome: poly

ξ_1	$p_1(r_1)$	$p_1(r_2)$	$p_1(r_3)$	$p_1(r_4)$
ξ_2	$p_2(r_1)$	$p_2(r_2)$	$p_2(r_3)$	$p_2(r_4)$
ξ_3	$p_3(r_1)$	$p_3(r_2)$	$p_3(r_3)$	$p_3(r_4)$
ξ_4	$p_4(r_1)$	$p_4(r_2)$	$p_4(r_3)$	$p_4(r_4)$
ξ_5	$p_5(r_1)$	$p_5(r_2)$	$p_5(r_3)$	$p_5(r_4)$

Distanzen 0 und viele andere ...

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 - 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|^2 = 2$$

Redundante Kardinalisierung

Fehlererkennende und fehlerkorrigierende Codes

Indexcodierung

Holzhammermethode: $\phi(\xi_j) = j$

$$\phi : \{\xi_1, \dots, \xi_\ell\} \rightarrow \mathbb{R}^1$$

Dualcodierung

(Hamming)abstände „fehleranfällig“

$$\phi : \{\xi_1, \dots, \xi_\ell\} \rightarrow \mathbb{R}^{\lceil \log_2 \ell \rceil}$$

Vollständige Korrekturcodes

Erkennt und kompensiert

$$\phi : \{\xi_1, \dots, \xi_\ell\} \rightarrow \mathbb{R}^L, L = 2^{\ell-1} - 1$$

Fehler in einer Komponente

(interessant ab $\ell = 4$)

ξ_1	1	1	1	1	1	1	1
ξ_2	0	0	0	0	1	1	1
ξ_3	0	0	1	0	0	0	1
ξ_4	0	1	0	1	0	0	1

Beinhaltet alle $\{0, 1\}^\ell$ -Spalten außer Komplementen und den uninformativen Attributen 0, 1.

Konversion von Distanzfunktionen

Nachbarschaft — Metrik — normierter Vektorraum

Metrik \Rightarrow symmetrische Nachbarschaft

Global operierende Schwellwertoperation ($0 < \delta_{\max} \in \mathbb{R}$)

$$\xi_i \propto \xi_j \Leftrightarrow d(\xi_i, \xi_j) \leq \delta_{\max}$$

Metrik \Rightarrow nichtsymmetrische Nachbarschaft

Lokale Umgebungsdefinition (k nächste Nachbarn, $k \in \mathbb{N}$)

$$\xi_i \propto \xi_j \Leftrightarrow \xi_j \in \mathcal{U}_\mathcal{X}^{(k)}(\xi_i)$$

Adjazenz \Rightarrow Metrik

Geodätische Abstände (minimale Pfadlängen im Adjazenzgraphen)

Metrik \Rightarrow (euklidischer) Vektorraum

Nicht jede metrische Distanz $D \in \mathbb{R}^{L \times L}$ ist im \mathbb{R}^{L-1} repräsentierbar.

Floyd-Warshall-Algorithmus

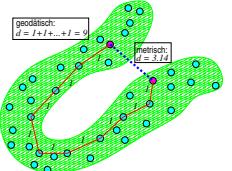
Schnelle Berechnung geodätischer Distanzen mittels dynamischer Programmierung

1 INITIALISIERUNG

$$\text{Setze } D_{ij} = \begin{cases} 0 & i=j \\ 1 & \xi_i, \xi_j \text{ adjazent} \\ \infty & \text{sonst} \end{cases}$$

Wirkungsweise

Der FWA erzwingt in $O(L^3)$ Schritten die Gültigkeit der Dreiecksungleichung.



2 REKURSION

Für alle $k, i, j \in \{1, \dots, L\}$:

$$D_{ij} \leftarrow \min \{D_{ij}, D_{ik} + D_{kj}\}$$

3 TERMINIERUNG

Die Matrix D enthält alle minimalen Wegelängen zwischen Elementen ξ_i, ξ_j .

Bemerkung

Der Algorithmus ist auch anwendbar für gewichtete und nichtsymmetrische Adjazenzen.

Kardinalisierung von Präferenzrelationen

Schwache Ordnungsrelation (\mathcal{X}, \prec) \Leftrightarrow ein, zwei, mehrere relative Attribute

Intervallordnung

Repräsentation durch $\mathcal{X}_1 \times \mathcal{X}_2 = \mathbb{R}^2$ mit

$$a \prec b \Leftrightarrow a_2 < b_1$$

Inklusionsfreie Intervallordnung

Repräsentation durch $\mathcal{X}_1 = \mathbb{R}^1$ mit $\delta \in \mathbb{R}_+$ und

$$a \prec b \Leftrightarrow a_1 + \delta < b_1$$

Endliche Halbordnung

Repräsentation durch $\mathcal{X}_1 \times \dots \times \mathcal{X}_L = \mathbb{R}^L$ mit

$$a \prec b \Leftrightarrow \forall \ell = 1, \dots, L: a_\ell < b_\ell$$

Standardisierung numerischer Skalen

Vereinheitlichung von Wertebereichen u/o Dynamikeigenschaften

Min-Max-Normierung

$$f : \begin{cases} \mathbb{R} & \rightarrow [0, 1] \\ x & \mapsto \frac{x - x_{\min}}{x_{\max} - x_{\min}} \end{cases}, \quad f^{-1}(x) = (x_{\max} - x_{\min}) \cdot x + x_{\min}$$

Statistische Normierung

$$f : \begin{cases} \mathbb{R} & \rightarrow [\mu - C\sigma, \mu + C\sigma] \\ x & \mapsto \frac{x - \mu}{\sigma} \end{cases}, \quad f^{-1}(x) = \sigma \cdot x + \mu$$

Reziproke Transformation

$$f : \begin{cases} \mathbb{R} \setminus \{0\} & \rightarrow \mathbb{R} \setminus \{0\} \\ x & \mapsto 1/x \end{cases}, \quad f^{-1}(x) = 1/x$$

Standardisierung numerischer Skalen

Vereinheitlichung von Wertebereichen u/o Dynamikeigenschaften

Wurzel-Transformation

$$f : \begin{cases} (C, \infty) & \rightarrow \mathbb{R}^+ \\ x & \mapsto \sqrt[B]{x - C} \end{cases}, \quad f^{-1}(x) = x^B + C$$

Logarithmus-Transformation

$$f : \begin{cases} (C, \infty) & \rightarrow \mathbb{R} \\ x & \mapsto \log_B(x - C) \end{cases}, \quad f^{-1}(x) = B^x + C$$

Fisher-Transformation

$$f : \begin{cases} (-1, +1) & \rightarrow \mathbb{R} \\ x & \mapsto \frac{1}{2} \log_e \frac{1+x}{1-x} \end{cases}, \quad f^{-1}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Werteskalen

Relationen und Distanzen

Skalenkonversion

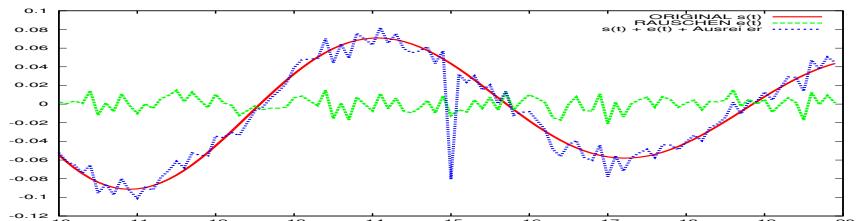
Detektion von Ausreißern

Imputation von Fehlanzeigen

Zusammenfassung

Meßfehler & Erhebungsfehler

Die „Rohdaten“ sind oft fehlerbehaftet, verrauscht, verzerrt



Zufällige Fehler

- Meßgenauigkeit
- Übertragungsstrecke
- Modell *additives Rauschen*: $y_n = x_n + e_n, e_n \sim \mathcal{N}(0, \sigma^2)$

Ausreißer

Systematische Fehler

- Kalibrierung
- Skalierung
- Trend, Drift, Saisoneffekt
- Ausreißer
- Ausreißer

Ausreißerdetektion

Was ist ein Ausreißer und wie erkenne ich ihn ?

Vertikale Detektion

Ein Wert x_{ij} fällt aus dem Rahmen seines **Attributs** \mathcal{X}_j .

Kategoriale Attribute bieten *keine Handhabe* !

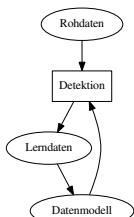
$$\mathcal{X}_j = \{m, f\}$$

Horizontale Detektion

Ein Wert x_{ij} fällt aus dem Rahmen seines **Objekts** \mathbf{o}_t .

Werden Objekte durch Ausreißer erst *interessant* ?

$$\mathbf{o}_t = (\text{„kath.“, „verh.“})$$



Teufelskreis

vertikale Detektion \Rightarrow Attributmodell

horizontale Detektion \Rightarrow Objektmodell

Satz (Tschebyscheff)

Ist \mathbb{X} eine kontinuierliche Zufallsvariable mit dem Erwartungswert μ und der Varianz σ^2 , so gilt für jede Konstante $C > 0$ die Ungleichung:

$$P\left(\left|\frac{\mathbb{X} - \mu}{\sigma}\right| \geq C\right) \leq \frac{1}{C^2}$$

Beispiel

Zweiseitige Streuungswahrscheinlichkeiten m/o NV-Annahme:

	σ	2σ	3σ	4σ	5σ
Tschebyscheff	≤ 1	≤ 0.25	≤ 0.11	≤ 0.063	≤ 0.040
$\mathcal{N}(\mu, \sigma)$	= 0.323	= 0.065	= 0.003	= 0.001	≤ 0.0001

\Rightarrow ein „zahnloser“ Test ohne Kenntnis der Dichtefunktion !

Hypothesentests für Ausreißer

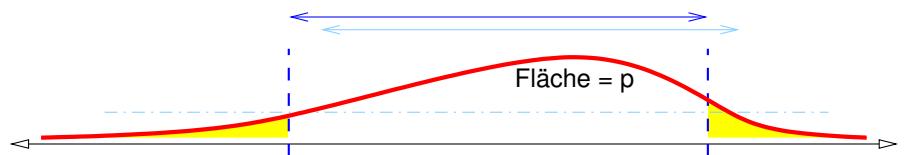
... bei bekannter unimodaler Verteilungsdichtefunktion

Definition (Quantilmethode)

Ein Wert $x_q \in \mathbb{R}$ heißt **q -Quantil** der Dichtefunktion $f_{\mathbb{X}}(\cdot)$ genau dann, wenn gilt:

$$F_{\mathbb{X}}(x_q) = P(\mathbb{X} \leq x_q) = q$$

Ein Wert $x \in \mathbb{R}$ heißt **Ausreißer** der Verteilung zum **Niveau** $p \in [0, 1]$, wenn er außerhalb des Akzeptanzintervalls $[x_{1/2-p/2}, x_{1/2+p/2}]$ liegt.



Bemerkungen

- Für **symmetrische** Dichtefunktionen gilt für jedes $q \in [0, 1]$ die Identität $f_{\mathbb{X}}(x_q) = f_{\mathbb{X}}(x_{1-q})$. $[\mu - C\sigma, \mu + C\sigma]$
- Für **mehrmodale** Dichtefunktionen ergibt das definierte Akzeptanzintervall keinen Sinn.

Hypothesentests für Ausreißer

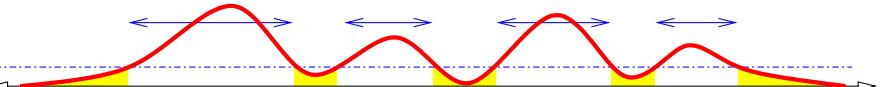
... bei bekannter multimodaler Verteilungsdichtefunktion

Definition (Bayesträgermethode)

Die Wertemenge $\mathcal{B}_c = \{x \mid f_{\mathbb{X}}(x) \geq c\}$ heißt **Bayesträger** der Verteilung $f_{\mathbb{X}}(\cdot)$ zum **Niveau** $p \in [0, 1]$, wenn gilt:

$$\int_{\mathcal{B}_c} f_{\mathbb{X}}(\xi) d\xi = p$$

Jeder Wert $x \in \mathbb{R}$ mit $f_{\mathbb{X}}(x) < c$ heißt **Ausreißer** der Verteilung zum **Niveau** p .



Bemerkungen

- Für **symmetrisch-unimodale** Dichtefunktionen stimmen Bayesträger und Akzeptanzintervall überein.
- Nicht verwechseln mit **Bayesintervall**, dem kürzesten Intervall mit Fläche p .

Faustregeln zur Ausreißerdetektion

Treffer als Fehlanzeige (NA=„not available“) markieren

Unimodal

- Normalverteilung**

$$|x - \mu| > C \cdot \sigma$$

- Gleichverteilung**

$$|x - \mu| > p\text{-Niveau}$$

- Empirischer Trimm**

$$x \notin [x_{1/2-p/2}, x_{1/2+p/2}]$$

Multimodal

- Tschebyscheff**

$$\frac{|x - \mu|}{\sigma} > \sqrt{\frac{1}{1-p}}$$

- Gauß-Mischung**

$$(\forall \ell) |x - \mu_\ell| > C \cdot \sigma_\ell$$

- Lonesome Cowperson**

$$|x - k\text{-NN}(x)| > d_{\max}$$

Teufelskreis Parameterschätzung

Ausreißer verändern die genutzten Verteilungsparameter

Modellrechnung für die $C\sigma$ -Regel

- Datensatz**

Eine $\mathcal{N}(\mu, \sigma)$ -verteilte Probe der Größe T zuzüglich M^+ Ausreißer der Gestalt $a^+ = \mu + c\sigma$ zuzüglich M^- Ausreißer der Gestalt $a^- = \mu - c\sigma$ ($T' = T + M^+ + M^-$)

- Geschätzter Erwartungswert**

(Im Fall $M^+ = M^-$ gilt einfach $\hat{\mu} = \mu$.)

$$\hat{\mu} = \mu + \frac{M^+ - M^-}{T'} \cdot c\sigma$$

- Geschätzte Varianz**

Gilt mit $M := M^+/2 = M^-/2$ wegen

$$\frac{1}{T'} \sum_{x \in \omega'} x^2 = \mu^2 + \sigma^2 + \frac{M}{T+M} (c^2 - 1)\sigma^2$$

und der Abkürzung $r := M/(T+M)$.

$$\hat{\sigma} = \sigma \sqrt{1 + c^2 r - r}$$

Für eine nicht verschwindende Anzahl ($r \gg 0$) von markanten Ausreißern ($c \gg 1$) dominiert $c^2 r$ den Wurzelausdruck und die $C\sigma$ -Regel ist wegen $\hat{\sigma} \propto c$ entschärft!

Fehlanzeigen (a.k.a. „not available“)

Nicht zugängliche Attributwerte in der Datenmatrix

Werteskalen

Relationen und Distanzen

Skalenkonversion

Detektion von Ausreißern

Imputation von Fehlanzeigen

Zusammenfassung

Fehlanzeige als Unfall

Sensorkomponente hat versagt
Erhebungsprotokoll unvollständig
Markierte Ausreißer

Fehlanzeige als Regelfall

Verzicht aus Kostengründen
Nichthomogenes Warehousing
Dünnbesetzung anwendungsbedingt
z.B. Bewertungssysteme für Musik, Bücher,
Restaurants, Webseiten, Bordellbetriebe, ...

Fehlanzeigenbehandlung

- Objekt löschen**
Können wir uns das leisten?
- Eintrag markieren**
und auf spezielle Weise weiterverarbeiten.
- Imputieren**
Leerstelle mit geeignetem Wert auffüllen.

Imputationstechniken

Welcherart Zusatzinformation wird zur Wertergänzung genutzt ?

$$\begin{array}{cccccc} M_1 & M_2 & \dots & M_n & \dots & M_N \\ \hline x_1 & x_2 & \dots & x_n & \dots & x_N \end{array}$$

Kontextfrei (MCAR)
Attributstatistik (& Ausreißer)

- Ersetzen durch Datenmittel $\hat{\mu}$
- durch x_{\min} bzw. x_{\max}
- durch nächsten Nachbarn

$$x_n^* \stackrel{\text{def}}{=} \underset{\xi \in \bar{\omega}}{\operatorname{argmin}} d(x_n, \xi)$$

„Missing (Completely) At Random“

Interpolation (MAR)
Zeile/Spalte $\hat{=}$ Meßreihe

- Linearer Ausgleich, z.B.
- Polynome, Splines (nl.)
- Glättungsfilter

$$x_n^* \stackrel{\text{def}}{=} (x_{n-1} + x_{n+1}) / 2$$

Regression (MAR)
Probabilistisches Datenmodell

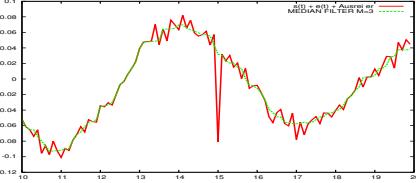
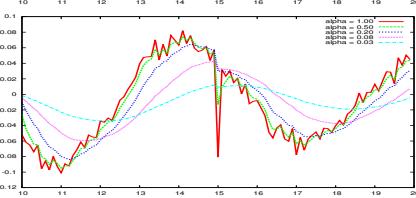
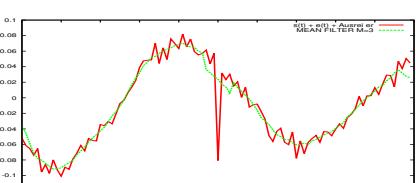
$$x_n^* = \mathcal{E}[\mathbb{X}_n | \dots, x_{n-1}, x_{n+1}, \dots]$$

Matrixapproximation (MAR)
Lückenhafte (num.) Datenmatrix

$$\mathbf{X} \xrightarrow{\text{NA}} \mathbf{V}^\top \mathbf{D} \mathbf{U}$$

Glättungsfilter für Meßreihenfehler

Imputation $\hat{=}$ kontextfrei Ersetzen + Filtern



Gleitender Mittelwert
der Ordnung $q = 2p + 1$, $p \in \mathbb{N}$:

$$\hat{x}_n = \frac{1}{q} \cdot \sum_{\ell=n-p}^{n+p} x_\ell$$

↳ Ausreißer, ↳ Phasentreue

Exponentialfilter
mit Abklingparameter $\alpha \in [0, 1]$:

$$\hat{x}_n = \hat{x}_{n-1} + \alpha \cdot (x_n - \hat{x}_{n-1})$$

↳ Ausreißer,
↳ Phasentreue/Nivellierung

Gleitendes Medianfilter
der Ordnung $q = 2p + 1$, $p \in \mathbb{N}$:

$$\hat{x}_n = \mu^{\text{med}}(x_{n-p}, \dots, x_n, \dots, x_{n+p})$$

Regression für nominale Datensätze

Beispielszenario: drei Attribute X_1, X_2, X_3 mit 2, 3 bzw. 4 Wertestufen

(Algorithmus)

1 ABSOLUTE HÄUFIGKEITEN

Erstelle Tabelle $f \in \mathbb{N}^{2 \cdot 3 \cdot 4}$ mit den 24 Auftretenzahlen f_{ijk} der Ereignisse $(x_1, x_2, x_3) = (\xi_i, \eta_j, \zeta_k)$.

2 EREIGNISWAHRSCHENLICHKEITEN

Erstelle Tabelle der 24 ML-Schätzwerte $\hat{p}_{ijk} = f_{ijk}/T$.

3 BEDINGTE ATTRIBUTWAHRSCHENLICHKEITEN

$$q_{i|jk}^{(1|23)} = \frac{\hat{p}_{ijk}}{\sum_\ell p_{\ell jk}}, \quad q_{j|ik}^{(2|13)} = \frac{\hat{p}_{ijk}}{\sum_\ell p_{i\ell k}}, \quad q_{k|ij}^{(3|12)} = \frac{\hat{p}_{ijk}}{\sum_\ell p_{ij\ell}}$$

4 IMPUTATION DES BEDINGTEN MODUS

$$\mu_{jk}^{(23)} = \operatorname{argmax}_i q_{i|jk}^{(1|23)}, \quad \mu_{ik}^{(13)} = \operatorname{argmax}_j q_{j|ik}^{(2|13)}, \quad \mu_{ij}^{(12)} = \operatorname{argmax}_k q_{k|ij}^{(3|12)}$$

(zumdInogA)

Zusammenfassung (2)

1. Ein Datensatz besteht aus **Objekten**, die explizit durch eine Reihe von **Attributwerten** oder implizit durch Beziehungen wie **Abstand**, **Adjazenz** oder **Präferenz** charakterisiert sind.
2. Attribute besitzen eine **diskrete Skala** (**nominal** oder **ordinal**) oder eine **numerische Skala** (**relativ** oder **proportional**).
3. Die Skalen unterscheiden sich hinsichtlich ihres **Wertebereichs**, ihrer **Verknüpfungsoperationen** und ihrer **Durchschnittswertbildung**.
4. Auf **Zeichenketten** ist mit dem Levenshteinabstand eine **Metrik** und mit dem **Medoid** ein Durchschnitt definiert.
5. Skalen lassen sich nötigenfalls mittels **Quantisierung** (numerisch \rightsquigarrow ordinal), **Entflechtung** (ordinal \rightsquigarrow nominal) bzw. **Kontrastmatrizen** (nominal \rightsquigarrow numerisch) konvertieren.
6. Aus **Adjazenzen** leiten sich **geodätische Distanzen** her, aus **Präferenzen** ein oder zwei **Ordinalskalen**.
7. **Ausreißer** werden durch einen der attributbezogenen **Hypothesentests** detektiert.
8. Als **Ersatzwerte** für Ausreißer und andere **Fehlanzeigen** dienen Mittel- und Extremwerte; wenn möglich, imputieren wir durch **Interpolation** oder **Regression**.

Werteskalen

Relationen und Distanzen

Skalenkonversion

Detektion von Ausreißern

Imputation von Fehlanzeigen

Zusammenfassung

MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2020

Prof. E.G. Schukat-Talamazzini

Stand: 21. November 2020

Teil III

Visualisierung von Massendaten

PCA MDS NMF ICA Faktoren SOM/MLP EDA

Datenvisualisierung

Mensch-Maschine-Interaktion zur explorativen Datenanalyse

Das menschliche Auge

ist das beste bekannte Werkzeug zur explorativen Datenanalyse

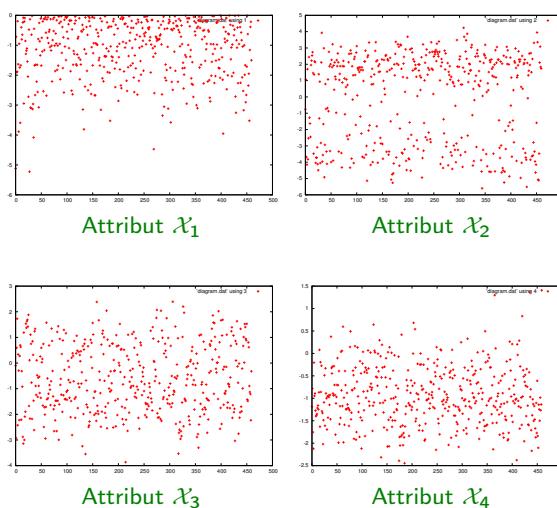
Aufgaben der Visualisierung

Graphische Darstellung der Datenvektoren eines Datensatzes

- Welche Objekte werden dargestellt ?
- Welche Objektattribute werden dargestellt ?
- Wie werden Attributwerte ergonomisch befriedigend dargestellt ?
- Was geschieht mit Abhängigkeiten (höherer Ordnung) ?

PCA MDS NMF ICA Faktoren SOM/MLP EDA

1D-Diagramme

$$\mathcal{G}_i(\omega) = \{(t, x_{t,i}) \mid t = 1, \dots, T\} \text{ für } 1 \leq i \leq N$$


Achsenparallele
Projektion

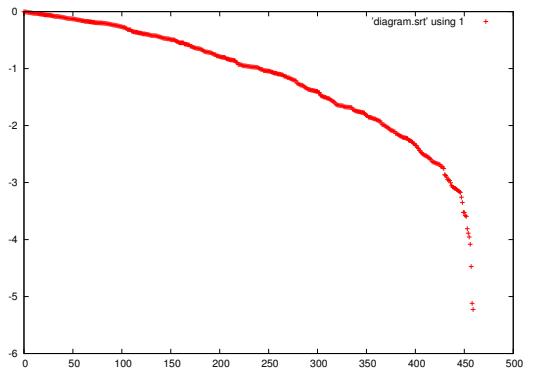
N separate
Diagramme für
Einzelattribute

t -Achse nur bei
Zeitreihen
informativ

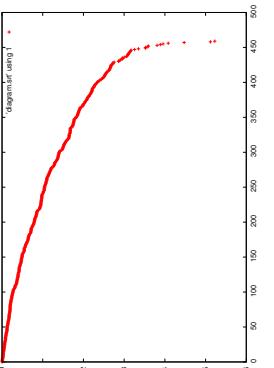
Grobe
Häufungseigen-
schaften

1D-Diagramme mit sortiertem Attributwert

Attribut \mathcal{X}_1 absteigend sortiert



Drehung 90°



Bemerkungen

- Über die nicht bedeutungstragende t -Achse wird durch Sortierung verfügt.
- Nach Vierteldrehung ergibt sich die *empirische kumulative Verteilungsfunktion*.
- Achtung — keine Dichtefunktion nach Ableitung ... !

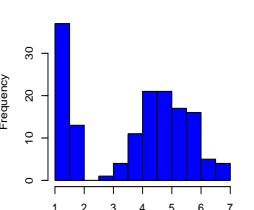
1D-Diagramme

Histogramme mit Absoluthäufigkeiten oder Punktdichten

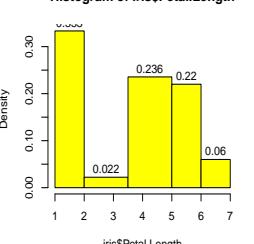
-Code

```
layout (matrix (
  1:4, 2, 2, TRUE
))
brx <- c (
  1,2,3,5,5,6,7
)
hist (iris$Petal.Length,
  col="blue")
hist (iris$Petal.Length,
  freq=FALSE,
  col="yellow")
hist (iris$Petal.Length,
  breaks=brx,
  labels=TRUE,
  col="yellow")
hist (iris$Petal.Length,
  breaks=brx, freq=TRUE,
  labels=TRUE,
  col="blue")
```

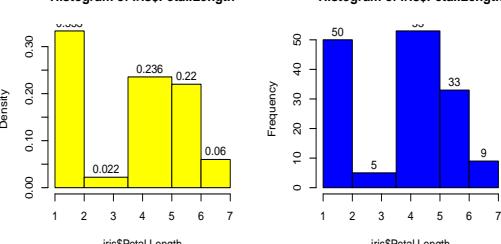
Histogramm von iris\$Petal.Length



Histogramm von iris\$Petal.Length

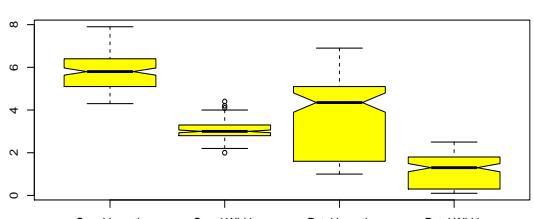
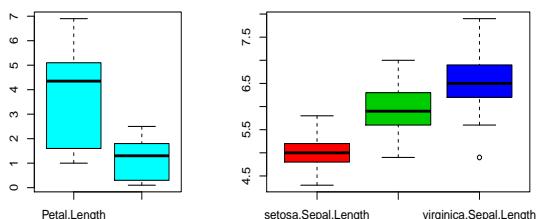


Histogramm von iris\$Petal.Length



1D-Diagramme

Box-Whisker-Plots · Quartile, Ausreißer, Signifikanzintervall



-Code

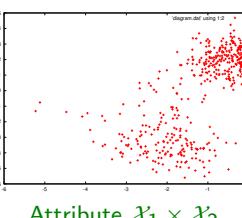
```
layout (matrix (
  c(1,2,3,3), 2, 2, TRUE
), width=c(2,3))
boxplot (iris[3:4],
  col="cyan")
boxplot (
  sapply (split (
    iris[1], iris[[5]]
  ), as.vector),
  col=2:4
)
boxplot (iris[1:4],
  notch=TRUE,
  col="yellow")
```

Wespentaille

mit der Breite:
 $\pm 1.58 \cdot IQR / \sqrt{N}$
 95%-signifikant
 unterschiedlicher Median
 bei fehlender
 Überschneidung

2D-Diagramme (Streudiagramme)

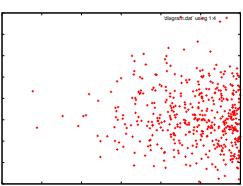
$\mathcal{G}_{ij}(\omega) = \{(x_{t,i}, x_{t,j}) \mid t = 1, \dots, T\}$ für $1 \leq i < j \leq N$



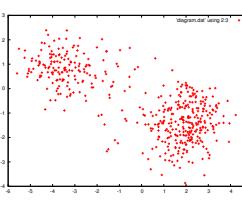
Attribute $\mathcal{X}_1 \times \mathcal{X}_2$



Attribute $\mathcal{X}_1 \times \mathcal{X}_3$



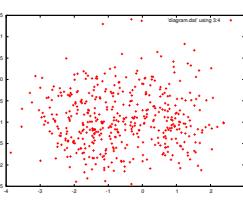
Attribute $\mathcal{X}_1 \times \mathcal{X}_4$



Attribute $\mathcal{X}_2 \times \mathcal{X}_3$



Attribute $\mathcal{X}_2 \times \mathcal{X}_4$



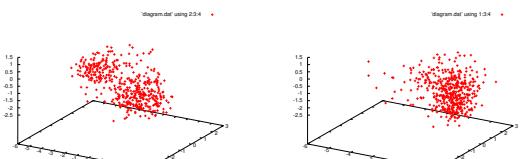
Attribute $\mathcal{X}_3 \times \mathcal{X}_4$

Achsenparallele Projektion mit $\binom{N}{2}$ separaten Diagrammen für Attributpaare

Statistische Abhängigkeiten \triangleq asphärische Ballungsform

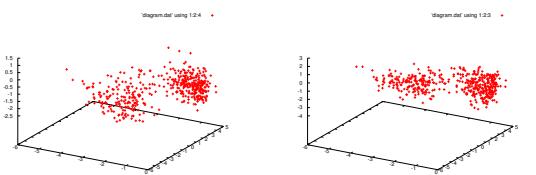
3D-Diagramme (Streuungswürfel)

$$\mathcal{G}_{ijk}(\omega) = \{(x_{t,i}, x_{t,j}, x_{t,k}) \mid t = 1, \dots, T\} \text{ für } 1 \leq i < j < k \leq N$$



Attribute $\mathcal{X}_2 \times \mathcal{X}_3 \times \mathcal{X}_4$

Attribute $\mathcal{X}_1 \times \mathcal{X}_3 \times \mathcal{X}_4$



Attribute $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_4$

Attribute $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$

Achsenparallele
Projektion

$\binom{N}{3}$ separate
Diagramme für
Attributtripel

Im Prinzip
Dreiwege-
abhängigkeiten

Ansicht-
bedingte
Darstellungs-
defizite

Erster Fluch der
Dimension

Mehrdimensionale
Wertetupel sind
kognitiv nur
unzureichend erfaßbar!

Zweiter Fluch der
Dimension

Die relevanten
Strukturen des
Datensatzes beinhalten
oft Abhängigkeiten
höherer (≥ 3)
Ordnung!

Dimensionsreduzierende Abbildungen
Kanonische Projektionen $\binom{N}{M}$ Auswahlmengen

$$\phi : (x_1, \dots, x_N) \mapsto (x_{i_1}, \dots, x_{i_M})$$

Lineare Projektionen Achsrichtungskoeffizienten

$$\phi : \mathbf{x} = (x_1, \dots, x_N)^\top \mapsto \Phi \mathbf{x}, \quad \Phi \in \mathbb{R}^{M \times N}$$

Nichtlineare Abbildungen arithm. Funktionsstruktur

$$\phi : \mathbf{x} \in \mathbb{R}^N \mapsto (\phi_1(\mathbf{x}), \dots, \phi_M(\mathbf{x}))$$

Punktuelle Abbildungen keine Neuzugänge

$$\phi : \{\mathbf{x}_1, \dots, \mathbf{x}_T\} \subset \mathbb{R}^N \mapsto \{\mathbf{y}_1, \dots, \mathbf{y}_T\} \subset \mathbb{R}^M$$

Visualisierung durch Dimensionsreduktion

Gütekriterien zur Dimensionsreduktion

1. Varianz

Die visuellen Achsen y_i besitzen maximale Streuung.

2. Reproduktion

Die Originale \mathbf{x}_t sind aus den \mathbf{y}_t (fast) wiederherstellbar.

3. Distanztreue

Paarweise Abstände $d(\mathbf{x}_s, \mathbf{x}_t)$ und $d(\mathbf{y}_s, \mathbf{y}_t)$ sind etwa gleich.

4. Häufung

Die Originale \mathbf{x}_t sind K Prototypen zuzuordnen.

5. Unabhängigkeit

Die visuellen Achsen y_i sind unabhängig oder unkorreliert.

Nichtnumerische Attribute

- Spezialgrafiken
- Vorabkonversion
- Metrische Reduktion

Hauptachsentransformation

Mehrdimensionale Skalierung

Nichtnegative Matrixfaktorisierung

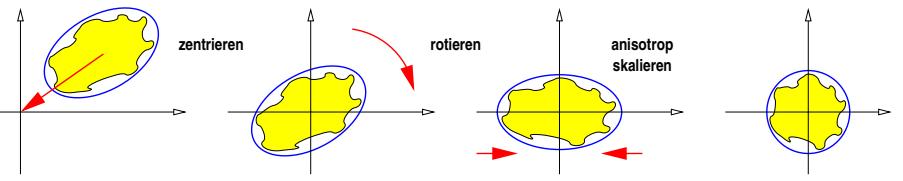
Independent Component Analysis

Faktorenanalyse

Merkmalkarte und autoassoziatives MLP

Explorative Datenanalyse & Grafik

PCA — Principal Component Analysis



Vollständige Reihenentwicklung (Karhunen-Loëve)

Translation · Rotation · Achsenstreckung/-stauchung

$$\mathbf{x} \mapsto \mathbf{x} - \boldsymbol{\mu} \mapsto \mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu}) \mapsto \mathbf{D}^{-1} \mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu})$$

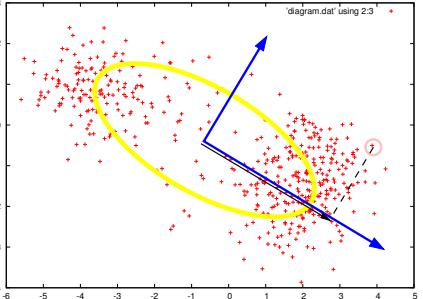
Unvollständige Reihenentwicklung

Dimensionsreduktion auf wenige ($M \ll N$) hochvariante Achsen:

$$\Phi : \begin{cases} \mathbb{R}^N & \rightarrow \mathbb{R}^M, M \ll N \\ \mathbf{x} & \mapsto (y_1, \dots, y_M)^\top \end{cases}, \quad y_m = \frac{1}{d_m} \cdot \mathbf{u}_m^\top (\mathbf{x} - \boldsymbol{\mu})$$

Raumrichtungen mit maximaler Datenvarianz

Über Datenwolken $\{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ des \mathbb{R}^N und ihre Hauptachsen $\mathbf{u}_1, \dots, \mathbf{u}_N$



Projektion von \mathbf{x} auf \mathbf{u} :

$$\frac{\mathbf{x}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} \cdot \mathbf{u} \text{ bzw. } \mathbf{x}^\top \mathbf{u} \cdot \mathbf{u}$$

Varianzanteil von \mathbf{x} auf \mathbf{u} :

$$\|\mathbf{x}^\top \mathbf{u} \cdot \mathbf{u}\|^2 = (\mathbf{x}^\top \mathbf{u})^2$$

(für normierte Achse \mathbf{u})

Varianzmaximierung

Die erste Hauptachse \mathbf{u}_1 mittelwertfreier Daten maximiert:

$$\sigma_u^2 = \frac{1}{T} \sum_{t=1}^T (\mathbf{x}_t^\top \mathbf{u})^2 = \frac{1}{T} \sum_{t=1}^T \mathbf{u}^\top \mathbf{x}_t \cdot \mathbf{x}_t^\top \mathbf{u} = \mathbf{u}^\top \cdot \mathbf{S} \cdot \mathbf{u}$$

Für weitere Hauptachsen \mathbf{u}_m verwende Restkomponenten

$$\mathbf{x}_t^{(m)} = \mathbf{x}_t - \sum_{i=1}^{m-1} \mathbf{x}_t^\top \mathbf{u}_i \cdot \mathbf{u}_i.$$

Eigenzerlegung der Datenkovarianz

Definition

Besitzt der Datensatz $\omega = \{\mathbf{x}_1, \dots, \mathbf{x}_T\} \subset \mathbb{R}^N$ den Mittelwertvektor $\boldsymbol{\mu}$ und die empirische Kovarianzmatrix \mathbf{S} , so heißt die (normierte) Drehverschiebung

$$\mathbf{x} \mapsto \mathbf{U}^\top \cdot (\mathbf{x} - \boldsymbol{\mu}) \quad \text{bzw.} \quad \mathbf{x} \mapsto \mathbf{D}^{-1} \mathbf{U}^\top \cdot (\mathbf{x} - \boldsymbol{\mu})$$

mit der Orthogonalmatrix \mathbf{U} aus der nach absteigenden Eigenwerten sortierten Eigenzerlegung $\mathbf{S} = \mathbf{U} \mathbf{D}^2 \mathbf{U}^\top$ die **Karhunen-Loëve-** oder **Hauptachsentransformation** der Punktmenge ω .

Lemma (Dekorrelation & Varianzmaximierung)

(1) Unter der Karhunen-Loëve-Transformation wird die Kovarianzmatrix von ω zur Diagonalmatrix $\tilde{\mathbf{S}} = \mathbf{D}^2$.

(2) Die unvollständige KLT der Ordnung $M < N$ erzielt unter allen normierten M -dimensionalen Reihenentwicklungen die maximale Gesamtvarianz $d_1^2 + \dots + d_M^2$ des Datensatzes ω .

Beachte: Die Datenvarianz ist rotationsinvariant.

Hauptachsen und Hauptkomponenten

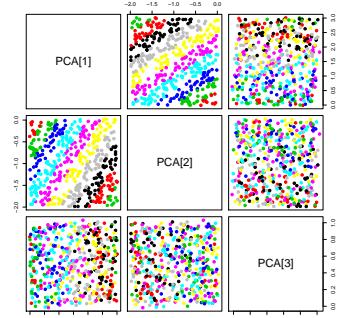
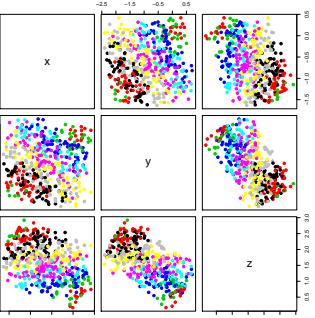
PCA: Punkte des \mathbb{R}^N werden durch wenige Hauptkomponenten repräsentiert

Definition

Für $i = 1, \dots, N$ heißt \mathbf{u}_i die i -te **Hauptachse** der KLT und es heißt $y_i = \mathbf{u}_i^\top \mathbf{x}$ die i -te **Hauptkomponente** des Punktes $\mathbf{x} \in \mathbb{R}^N$.

Beispiel

Gleichverteilte Punkte aus $[0, 1] \times [0, 2] \times [0, 3] \subset \mathbb{R}^3$ · Einfärbung gemäß $[y + z]$
Links: zufällig rotierter Würfel · Rechts: Hauptkomponentendarstellung



Algorithmen zur PCA

Voraussetzung: die Mittelwertfreiheit des Datensatzes

Standardberechnung

- Berechne die Kovarianzmatrix

$$\mathbf{S} = \frac{1}{T} \cdot \mathbf{X}^\top \mathbf{X}$$

- Berechne $M < N$ Hauptachsen

$$\mathbf{U}_{(M)} = (\mathbf{u}_1, \dots, \mathbf{u}_M)$$

- Berechne die Zielvektoren

$$\mathbf{y}_t = \mathbf{U}_{(M)}^\top \mathbf{x}_t \quad (\forall t)$$

Bemerkung

Für $N > T$ (für $T > N$) ist der duale (primäre) Algorithmus die bessere Wahl.

Duale Berechnung

- Berechne die Gramsche Matrix

$$\mathbf{G} = \mathbf{X} \mathbf{X}^\top$$

- Berechne $M < N$ Eigenvektoren

$$\mathbf{V}_{(M)} = (\mathbf{v}_1, \dots, \mathbf{v}_M)$$

und Eigenwerte d_1^2, \dots, d_M^2 .

- Berechne Zielkomponenten

$$y_{t,i} = d_i \cdot v_{i,t} \quad (\forall t, i)$$

Beweis.

Die Datenmatrix $\mathbf{X} \in \mathbb{R}^{T \times N}$ enthalte die Datenvektoren $\mathbf{x}_1, \dots, \mathbf{x}_T$ zeilenweise und besitze die Singulärwertzerlegung

$$\mathbf{X} = \mathbf{V} \mathbf{D} \mathbf{U}^\top, \quad \mathbf{V} \in \mathbb{R}^{T \times r}, \mathbf{D} \in \mathbb{R}^{r \times r}, \mathbf{U} \in \mathbb{R}^{N \times r}, r \leq \min(T, N).$$

Wegen der Mittelwertfreiheit der Daten gilt einfach

$$\mathbf{S} = \frac{1}{T} \sum_{t=1}^T (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top = \frac{1}{T} \mathbf{X}^\top \mathbf{X} = \frac{1}{T} \mathbf{U} \mathbf{D}^2 \mathbf{U}^\top = \mathbf{U} \frac{1}{T} \mathbf{D}^2 \mathbf{U}^\top;$$

also ist \mathbf{U} die PCA-Basis und $\mathbf{X} \mathbf{U}_{(M)}$ enthält die Zielvektoren \mathbf{y}_t in den Zeilen. \square

Beweis.

Für die Gramsche Matrix $\mathbf{G} \in \mathbb{R}^{T \times T}$ der paarweisen Skalarprodukte $G_{st} = \mathbf{x}_s^\top \mathbf{x}_t$ gilt

$$\mathbf{G} = \mathbf{X} \mathbf{X}^\top = \mathbf{V} \mathbf{D} \mathbf{U}^\top \cdot \mathbf{U} \mathbf{D} \mathbf{V}^\top = \mathbf{V} \mathbf{D}^2 \mathbf{V}^\top;$$

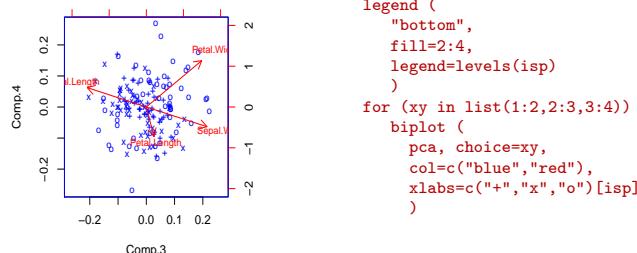
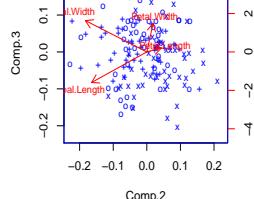
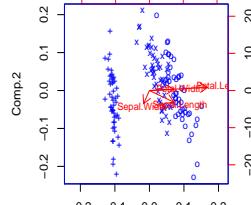
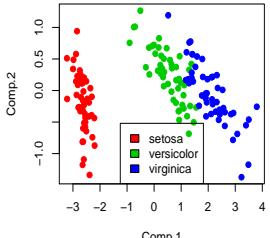
der Algorithmus berechnet also die ersten M Rechtssingulärvektoren von \mathbf{X} . Wegen der Orthogonalität von \mathbf{U} gilt die Identität $\mathbf{X} \mathbf{U} = \mathbf{V} \mathbf{D}$, und wir berechnen die Zielkomponenten

$$y_{ti} = (\mathbf{X} \mathbf{U})_{ti} = (\mathbf{V} \mathbf{D})_{ti} = (d_i \mathbf{v}_i)_t = d_i \cdot v_{it}$$

für alle $t = 1, \dots, T$ und einige wenige $i = 1, \dots, M$. \square

Scatterplot der Hauptkomponenten #1 und #2

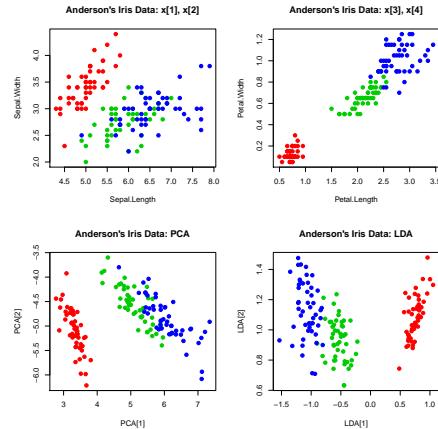
Gemeinsame Grafik für Hauptkomponenten und Hauptachsen (Gabriel & Odoroff 1990)



R-Code

```
layout (matrix (
  1:4, 2, 2, TRUE
))
isp <- iris$Species
pca <- princomp (
  ~.-Species,
  iris,
  cor=FALSE
)
plot (
  pca$scores[,1:2],
  pch=19,
  col=1+unclass (isp)
)
legend (
  "bottom",
  fill=2:4,
  legend=levels(isp)
)
for (xy in list(1:2,2:3,3:4))
  biplot (
    pca, choice=xy,
    col=c("blue","red"),
    xlabs=c("+","x","o") [isp]
  )
```

Skalierungsabhängigkeit & Klassentrennung



Skalierung

PCA \hookrightarrow Achsenkalierung
Verwende **Korrelationsmatrix** statt Kovarianzmatrix!

Nominale Attribute

Visualisierung durch Farbgebung u/o Ikonen.

Konvertieren nominal/numerisch?

Fisherdiskriminanten

Klassenseparation entschärft Skalierungsproblem.

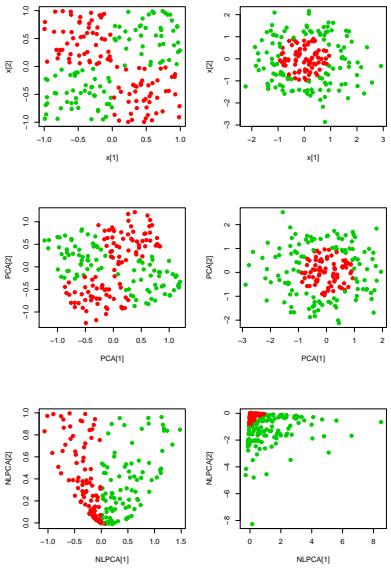
Definition (FDA oder überwachte PCA)

Sei $\mathbf{x}_1, \dots, \mathbf{x}_T$ ein Datensatz mit Klassenkennung c_1, \dots, c_T und \mathbf{S}_W bzw. \mathbf{S}_B die Inner- und Außerklassenstreuungsmatrix.

Die Projektionen $y_{ti} = \mathbf{u}_i^\top \mathbf{x}_t$ auf die Rechtseigenvektoren der Kitanomatrix $\mathbf{S}_W^{-1} \mathbf{S}_B$ heißen **lineare- oder Fisherdiskriminanten** von \mathbf{x}_t .

Nichtlineare PCA für „krumme“ Hauptachsen

Originaldaten $x \Rightarrow$ Termexpansion $\phi(x) \Rightarrow$ Hauptkomponenten $U_\phi^\top \phi(x)$



Polynomansatz

$$\phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2 \\ \sqrt{2} \cdot x_1 x_2 \\ x_2^2 \end{pmatrix}$$

Radialansatz

$$\phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2 \\ \frac{1}{2} \cdot x_1 x_2 \\ x_2^2 \end{pmatrix}$$

Kernel Trick ?

Berechne $G_\phi = X_\phi X_\phi^\top$.
Nutze Kernoperator
 $K(\mathbf{x}, \mathbf{y}) = \langle \phi \mathbf{x}, \phi \mathbf{y} \rangle$.
Expandierte Daten mittelwertfrei ?!?

Hauptachsentransformation

Mehrdimensionale Skalierung

Nichtnegative Matrixfaktorisierung

Independent Component Analysis

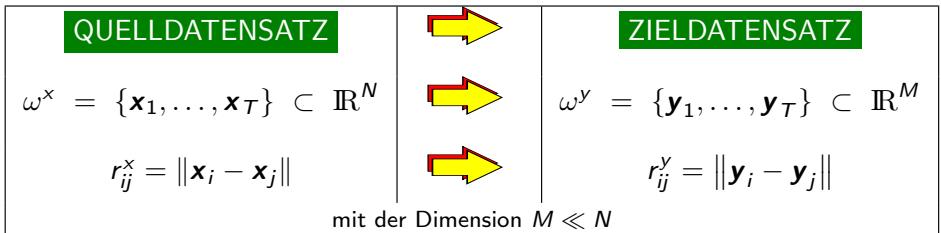
Faktorenanalyse

Merkmalkarte und autoassoziatives MLP

Explorative Datenanalyse & Grafik

Mehrdimensionale Skalierung (MDS)

Distanzerhaltende Repräsentation der Datenobjekte in einem visualisierbaren Raum \mathbb{R}^M



GESUCHT:

Niederdimensionale Punktmenge des \mathbb{R}^M mit geringstmöglicher Abweichung der paarweisen Distanzen

$$r_{ij}^x \approx r_{ij}^y \quad (\forall i, j = 1, \dots, T)$$

Es werden nur **Objektabstände** erwähnt, nicht jedoch **Objektattribute**.
(↔ Visualisierung *metrischer* Datensätze)

Reproduktionsfehler für die Objektdistanzen

Gütemaße für die „metrische“ Skalierung

Absoluter quadratischer Fehler

$$\varepsilon_a(\mathbf{R}^x, \mathbf{R}^y) = \frac{1}{\sum_{i < j} (r_{ij}^x)^2} \cdot \sum_{i < j} (r_{ij}^y - r_{ij}^x)^2$$

Relativer quadratischer Fehler

$$\varepsilon_r(\mathbf{R}^x, \mathbf{R}^y) = \frac{1}{\sum_{i < j} (r_{ij}^x)^2} \cdot \sum_{i < j} \left(\frac{r_{ij}^y - r_{ij}^x}{r_{ij}^x} \right)^2$$

Sammon-Fehler (der Kompromiß)

$$\varepsilon_s(\mathbf{R}^x, \mathbf{R}^y) = \frac{1}{\sum_{i < j} (r_{ij}^x)^2} \cdot \sum_{i < j} \frac{(r_{ij}^y - r_{ij}^x)^2}{r_{ij}^x}$$

Bemerkung

Der Fehler wird durch Gradientenabstieg bezüglich der M Koordinaten der Repräsentanten $\mathbf{y}_1, \dots, \mathbf{y}_T$ minimiert.

Auch im Falle euklidischer Datenobjekte $\mathbf{x}_t \in \mathbb{R}^N$ kennen wir keinerlei geschlossene Lösungen für diese drei Fehlermaße.

Der Sammonsche MDS-Algorithmus

(Algorithmus)

1 GEGEBEN:

Datensatz $\{x_1, \dots, x_T\} \subset \mathbb{R}^N$, Dimension $M < N$, Schrittweite $\eta > 0$, Schwelle θ .

2 INITIALISIERUNG:

Wähle zufällige Punkte $y_1, \dots, y_T \in \mathbb{R}^M$
Berechne alle Distanzen $r_{ij}^x = \|x_i - x_j\|$, $i \neq j$.

 $O(T^2N)$

3 ITERATIONSSCHRITT:

Berechne alle Distanzen $r_{ij}^y = \|y_i - y_j\|$, $i \neq j$
sowie die partiellen Ableitungen und setze

 $O(T^2MI_{\max})$

$$y_{t,m} \leftarrow y_{t,m} - \eta \cdot \frac{\partial \varepsilon}{\partial y_{t,m}}$$

für alle $1 \leq m \leq M$ und alle $1 \leq t \leq T$.

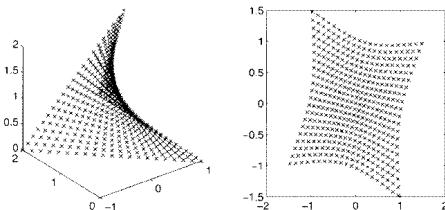
 $O(T^2MI_{\max})$

4 ABBRUCHBEDINGUNG:

Falls $\|\nabla \varepsilon\| \leq \theta$ dann \rightsquigarrow ENDE sonst \rightsquigarrow 3.

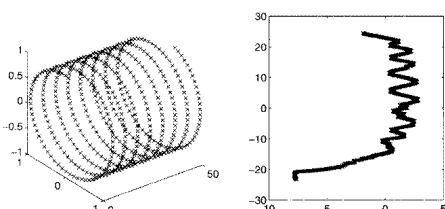
(zum dargestellten)

Synthetische Anwendungsbeispiele

Vom \mathbb{R}^3 in den \mathbb{R}^2 mit dem Sammon-AlgorithmusQuelldaten ω^x Zieldaten ω^y

3D-Torsionsfläche
 $\{(x_1, x_2, x_3)^\top \mid x_1 = (x_2 - 1)(x_3 - 1)\}$

- ⊕ Gitterstruktur
- ⊕ Äquidistanztendenz



3D-Spirale
 $\{(x, \sin(x), \cos(x))^\top \mid x \in [0, 50]\}$

- ⊕ Linientopologie
- ⊗ Welligkeit
- ⊖ Symmetrien
(kein globales Minimum)

Lemma

Im eindimensionalen Fall $y_1, \dots, y_T \in \mathbb{R}^1$ lauten die partiellen Ableitungen des Sammon-Fehlers

$$\frac{\partial \varepsilon_s}{\partial y_t} = \frac{1}{\sum(r_{ij}^x)^2} \cdot \sum_{s \neq t} \frac{r_{ts}^y - r_{ts}^x}{r_{ts}^x} \cdot \frac{y_t - y_s}{r_{ts}^y}$$

und im M -dimensionalen Fall analog (ersetze oben y_t durch $y_{t,i}$).

Beweis.

Es seien die Zielpunkte skalare Größen $y_1, \dots, y_T \in \mathbb{R}$. Sei $t \in \{1, \dots, T\}$.
In den Abständen r_{ij}^x kommt y_t gar nicht vor, folglich ist trivialerweise $\frac{\partial r_{ij}^x}{\partial y_t} = 0$.

Für die Abstände r_{ij}^y sind drei Fälle zu unterscheiden:

$$\frac{\partial r_{ij}^y}{\partial y_t} = \begin{cases} +(y_i - y_j)/r_{ij}^y & t = i \\ -(y_i - y_j)/r_{ij}^y & t = j \\ 0 & \text{sonst} \end{cases}$$

Dieses Resultat ergibt sich aus

$$r_{ij}^y = \|y_i - y_j\| = \sqrt{(y_i - y_j)^2} = ((y_i - y_j)^2)^{1/2}$$

durch wiederholte Anwendung der Kettenregel zur Differentiation. Aus den Distanzableitungen gewinnt man die Ableitungen der Fehlermaße ε_a , ε_r , ε_s auf geradem Wege. □

Punktweise Fortsetzung der MDS-Transformation

Berechnung von y_{T+1} für ein „neues“ Objekt x_{T+1}

Problem

Die MDS berechnet Repräsentanten y_1, \dots, y_T , aber weder eine lineare noch eine nichtlineare **explizite** Abbildungsvorschrift $\psi : \mathcal{X} \rightarrow \mathbb{R}^M$.

Lösung

1. Nächste-Nachbar-Regression

$$y_{T+1} = y_{s^*}, \quad s^* = \underset{s=1..T}{\operatorname{argmax}} d(x_{T+1}, x_s)$$

2. Konvexe Fortsetzung

$$y_{T+1} = \sum_{s=1}^T \lambda_s \cdot y_s, \quad \lambda_s \propto \exp\left(-\frac{1}{2} \cdot \|x_{T+1} - x_s\|^2\right)$$

3. Tabula-Rasa-Rekonfiguration

Anwendung der MDS auf den erweiterten Datensatz x_1, \dots, x_{T+1} .

Klassische mehrdimensionale Skalierung

Abstandstreue Transformation für Vektoren mit Skalarprodukt

Definition

Eine Abbildung der Vektoren $\mathbf{x}_1, \dots, \mathbf{x}_T \in \mathbb{R}^N$ auf Vektoren $\mathbf{y}_1, \dots, \mathbf{y}_T \in \mathbb{R}^M$ mit minimalem Fehler

$$\varepsilon_k(\mathbf{X}, \mathbf{Y}) = \|\mathbf{G}^y - \mathbf{G}^x\|_{\text{Frob}}^2 = \sum_{s,t=1}^T (\mathbf{y}_s^\top \mathbf{y}_t - \mathbf{x}_s^\top \mathbf{x}_t)^2$$

heißt **klassische Skalierung** des Datensatzes mit Dimension M .

Beachte: gleiche wechselseitige Skalarprodukte \Leftrightarrow gleiche wechselseitige Abstände

Lemma

Ist $\mathbf{G}^x = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top$ die Eigenzerlegung der Gramschen Matrix von \mathbf{X} , so bilden die Zeilenvektoren der Matrix

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{V}}\tilde{\mathbf{D}} \quad \text{mit } \tilde{\mathbf{V}} = (\mathbf{v}_1, \dots, \mathbf{v}_M), \quad \tilde{\mathbf{D}} = \text{diag}(d_1, \dots, d_M)$$

eine klassische Skalierung von \mathbf{X} mit Dimension M .

Eigenschaften der klassischen MDS

- Der **Rechenaufwand** $O(T^3)$ ist unter Umständen ($T \gg N$) hoch.
- KMDS ist eine **punktweise** Abbildung $\mathbf{x}_t \mapsto \mathbf{y}_t$; es gibt keine Vorschrift für neue Objekte \mathbf{x}_{T+1} .
- KMDS operiert auf **Skalarprodukten**, nicht auf **Objekten** und nicht auf **Distanzen**.

$$d^2(\mathbf{x}_s, \mathbf{x}_t) = \|\mathbf{x}_s - \mathbf{x}_t\|^2 = \langle \mathbf{x}_s, \mathbf{x}_s \rangle - 2 \cdot \langle \mathbf{x}_s, \mathbf{x}_t \rangle + \langle \mathbf{x}_t, \mathbf{x}_t \rangle$$

4. Translation, Rotation, Orientierung

Auch für $M = N$ keine exakte Reproduktion von \mathbf{X} durch \mathbf{Y} .

Klassische / nichtlineare / metrische Skalierung

$$\left\{ \begin{array}{l} \text{Vektoren } \mathbf{x}_t \in \mathbb{R}^N \\ \text{Produkte } \langle \mathbf{x}_s, \mathbf{x}_t \rangle \in \mathbb{R} \\ \text{Distanzen } d(\mathbf{x}_s, \mathbf{x}_t) \in \mathbb{R}_0^+ \end{array} \right\} \Rightarrow \mathbb{R}^T \Rightarrow \mathbb{R}^M$$

Beweis.

Die Gramsche Matrix $\mathbf{G}^x = \mathbf{X}\mathbf{X}^\top$ ist symmetrisch und positiv-semidefinit; ihre Eigenwerte können also stets als Quadrate geschrieben werden.

1. Ohne Dimensionsreduktion ($M = T$)

Für die Datenmatrix $\mathbf{Y} = \mathbf{V}\mathbf{D}$ gilt

$$\mathbf{G}^y = \mathbf{Y}\mathbf{Y}^\top = (\mathbf{V}\mathbf{D})(\mathbf{V}\mathbf{D})^\top = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top = \mathbf{G}^x,$$

also sind die wechselseitigen Skalarprodukte $\mathbf{x}_s^\top \mathbf{x}_t$ des Quellraums und $\mathbf{y}_s^\top \mathbf{y}_t$ des Zielraums exakt gleich, der Fehler $\varepsilon_k(\mathbf{X}, \mathbf{Y})$ folglich unbedingt minimal.

2. Mit Dimensionsreduktion ($M < T$)

Die Gramschen Matrizen für die Quellvektoren \mathbf{x}_t und für die dimensionsreduzierten Vektoren $\tilde{\mathbf{y}}_t$ lassen sich in Summenform

$$\mathbf{G}^x = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top = \sum_{t=1}^T d_t^2 \mathbf{v}_t \mathbf{v}_t^\top \quad \text{bzw.} \quad \mathbf{G}^{\tilde{y}} = \tilde{\mathbf{V}}\tilde{\mathbf{D}}^2\tilde{\mathbf{V}}^\top = \sum_{m=1}^M d_m^2 \mathbf{v}_m \mathbf{v}_m^\top$$

formulieren. Wegen der Orthogonalität der Eigenvektoren $\mathbf{v}_1, \dots, \mathbf{v}_T$ und wegen der Eigenwertsortierung $d_1^2 \geq d_2^2 \geq \dots \geq d_T^2$ stellt $\mathbf{G}^{\tilde{y}}$ die beste M -dimensionale Approximation von \mathbf{G}^x im Sinne des mittleren quadratischen Fehlers (Frobeniusnorm!) dar.

□

Klassische MDS und Hauptkomponentenanalyse

Singularwertzerlegung	$\mathbf{X} = \mathbf{V} \mathbf{D} \mathbf{U}^\top$
Momentenmatrix	$\mathbf{T} \mathbf{R} = \mathbf{X} \mathbf{X}^\top = \mathbf{U} \mathbf{D} \mathbf{U}^\top$
Rechts/Links Eigenvektoren	$\mathbf{X} = \mathbf{V} \mathbf{D} \mathbf{U}^\top$
Gramsche Matrix	$\mathbf{G}^x = \mathbf{X} \mathbf{X}^\top = \mathbf{V} \mathbf{D} \mathbf{V}^\top$

Lemma (MDS $\hat{=}$ PCA)

Sind die Quelldaten $\mathbf{x}_1, \dots, \mathbf{x}_T \in \mathbb{R}^N$ mittelwertfrei, so minimieren die Hauptkomponentenvektoren $\mathbf{y}_t = \mathbf{U}_{(M)}^\top \mathbf{x}_t$ das Fehlermaß der klassischen Skalierung zur Dimension M .

Beweis.

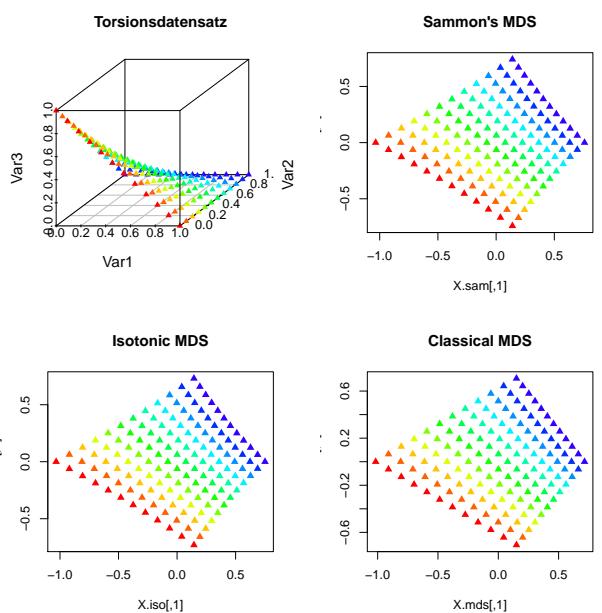
Ist die Datenmatrix mittelwertfrei, so $\mathbf{S}^x = \mathbf{R}^x - \mathbf{O}\mathbf{O}^\top = \mathbf{X}^\top \mathbf{X}$.

Im Vollrangfall ist $\mathbf{Y} = \mathbf{V}\mathbf{D} = \mathbf{X}\mathbf{U}$ bis auf den Faktor \sqrt{T} die (vollständige) PCA.

□

Beispiel Torsionsdatensatz

ISO-MDS — Absolutfehler plus isotone Distanzenverzerrung

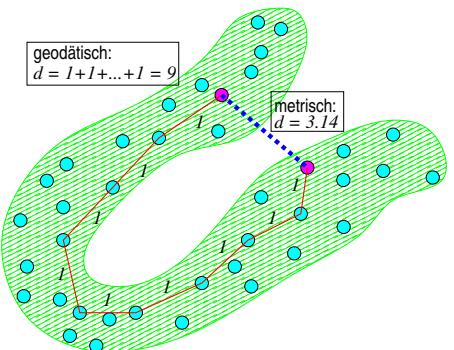


R-Code

```
layout (matrix (1:4,2,2,TRUE))
n <- 12
x <- seq (0, 1, length=n)
xy <- as.matrix (
  expand.grid (x, x))
X <- data.frame (
  xy,
  Var3=(1-xy[,1])*(1-xy[,2]))
o.col <- rep (
  rainbow (n, end=.7), each=n)
require (scatterplot3d)
scatterplot3d (
  X, pch=17, color=o.col,
  main="Torsionsdatensatz")
require (MASS)
X.sam <- sammon (
  dist(X)
  )$points
plot (X.sam, pch=17, col=o.col,
  main="Sammon's MDS")
X.iso <- isoMDS (
  dist(X), p=2
  )$points
plot (X.iso, pch=17, col=o.col,
  main="Isotonic MDS")
X.mds <- cmdscale (
  dist(X))
plot (X.mds, pch=17, col=o.col,
  main="Classical MDS")
```

Geodätische Skalierung

Nichtlineare Datenprojektion $\mathbb{R}^N \rightarrow \mathbb{R}^M$ für gekrümmte-zerklüftete Datenwolken



CCA (Curvilinear Component Analysis)

MDS für alle Abstände
 $d(\mathbf{x}_s, \mathbf{x}_t) \geq \theta$

ISOMAP

(Tenenbaum, 2000)
Geodätischer Abstand \sim
klassische MDS

GeoNLM

(Zha & Zhang, 2004)
Geodätischer Abstand \sim
Sammon-MDS

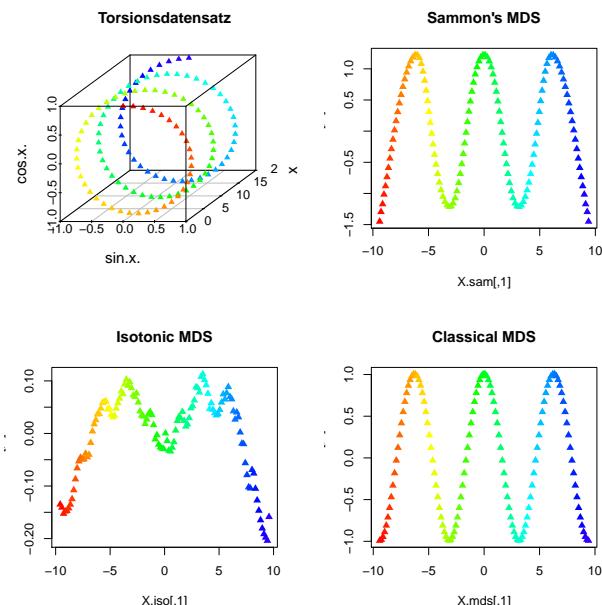
Definition (Knüllpapiermetapher)

Es sei $\mathcal{C} \subseteq \mathbb{R}^N$ zusammenhängend und abgeschlossen.

Die Länge $d_{\mathcal{C}}^{\text{geo}}(\mathbf{x}_s, \mathbf{x}_t)$ einer kürzesten Verbindungskurve zwischen \mathbf{x}_s und \mathbf{x}_t , die ausschließlich in \mathcal{C} verläuft, heißt der **geodätische Abstand** von \mathbf{x}_s und \mathbf{x}_t bezüglich \mathcal{C} .

Beispiel Spiralendatensatz

Sammon verwendet klassische MDS als Iterationsanfang!



R-Code

```
layout (matrix (1:4,2,2,TRUE))
n <- 124
xmax <- 3*2*pi
x <- seq (0, xmax, length=n)
o.col <- rainbow (n, end=0.7)
X <- data.frame (
  sin(x), x, cos(x)
  )
require (scatterplot3d)
scatterplot3d (
  X, pch=17, color=o.col,
  main="Torsionsdatensatz")
require (MASS)
X.sam <- sammon (
  dist(X)
  )$points
plot (X.sam, pch=17, col=o.col,
  main="Sammon's MDS")
X.iso <- isoMDS (
  dist(X), p=2
  )$points
plot (X.iso, pch=17, col=o.col,
  main="Isotonic MDS")
X.mds <- cmdscale (
  dist(X))
plot (X.mds, pch=17, col=o.col,
  main="Classical MDS")
```

Kürzeste Wege in Nachbarschaftssystemen

Diskretisierung des Begriffs geodätischer Abstände

Nachbarschaftsrelation in $\{\mathbf{x}_1, \dots, \mathbf{x}_T\}$

Adjazenzmatrix für eng beieinanderliegende Datenpunkte

- **Globale Nachbarschaft** (\oplus symmetrisch)

$$A_{st} = 1 \quad \Leftrightarrow \quad \|\mathbf{x}_s - \mathbf{x}_t\| < \delta$$

- **Lokale Nachbarschaft** (\ominus symmetrisch)

$$A_{st} = 1 \quad \Leftrightarrow \quad \mathbf{x}_t \in k\text{NN}(\mathbf{x}_s)$$

$$\delta \in \mathbb{R}^+$$

$$k \in \mathbb{N}$$

Floyd-Warshall-Algorithmus

(effiziente Pfadminimierung)

Feinjustierung problematisch (NBS-Grad k / Aktionsradius δ)

- **Small-World Problem**

kürzeste Verbindungskurven durch das Niemandsland

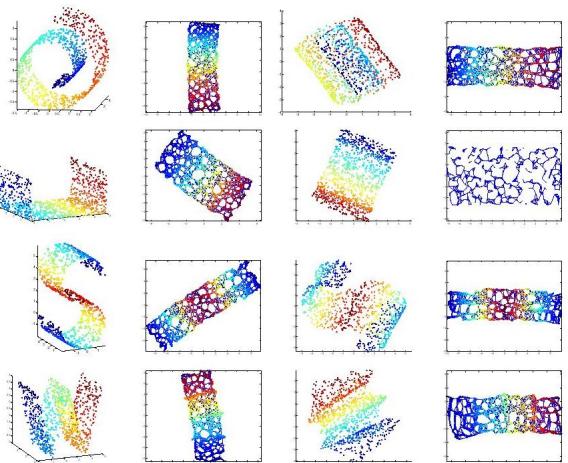
- **Emmenthalersyndrom**

schaumartige Maschenstruktur behindert geradlinige Wege

Beispiele zur geodätischen Skalierung

Beispiel A

„Gekrümmte Flächen im Raum“



Quelldaten Geo/MDS Sammon/MDS ISOMAP

Beispiel B

„Elektronische Spürnase“



Parfumdesign

$T = 300$ Rezepturen
 $N = 32$ Attribute
 ↗ Adjazenzurteile ↗
 $k = 7$ Nachbarn
 ↗ Stopelpfade ↗
 $M = 6$ Duftnoten

Beweis.

Es seien $\mathbf{X}_i = \sqrt{T_i} \cdot \mathbf{V}_i \mathbf{D}_i \mathbf{U}_i^\top$ die Singulärwertzerlegungen der Datensätze; dann gilt für $i = 1, 2$:

$$\mathbf{V}_i = \mathbf{X}_i \mathbf{U}_i \mathbf{D}_i^{-1} / \sqrt{T_i}$$

$$\mathbf{G}_i = \mathbf{X}_i \mathbf{X}_i^\top = T_i \cdot \mathbf{V}_i \mathbf{D}_i^2 \mathbf{V}_i^\top$$

$$\mathbf{U}_i = \mathbf{X}_i^\top \mathbf{V}_i \mathbf{D}_i^{-1} / \sqrt{T_i}$$

$$T_i \cdot \mathbf{R}_i = \mathbf{X}_i^\top \mathbf{X}_i = T_i \cdot \mathbf{U}_i \mathbf{D}_i^2 \mathbf{U}_i^\top$$

Wegen der geforderten Übereinstimmung der Momente gilt

$\mathbf{R}_1 = \mathbf{S}_1 + \mu_1 \mu_1^\top = \mathbf{S}_2 + \mu_2 \mu_2^\top = \mathbf{R}_2$ und damit auch $\mathbf{U}_1 = \mathbf{U}_2$ sowie $\mathbf{D}_1 = \mathbf{D}_2 =: \mathbf{D}$. Nach Einsetzen obiger Ausdrücke für \mathbf{V}_2 und \mathbf{U}_1 ergibt sich das behauptete Resultat:

$$\begin{aligned} \mathbf{V}_2 &= \mathbf{X}_2 \mathbf{U}_2 \mathbf{D}^{-1} \frac{1}{\sqrt{T_2}} = \mathbf{X}_2 \mathbf{U}_1 \mathbf{D}^{-1} \frac{1}{\sqrt{T_2}} = \mathbf{X}_2 \cdot \mathbf{X}_1^\top \mathbf{V}_1 \mathbf{D}^{-1} \frac{1}{\sqrt{T_1}} \cdot \mathbf{D}^{-1} \frac{1}{\sqrt{T_2}} \\ &= \mathbf{X}_2 \mathbf{X}_1^\top \cdot \mathbf{V}_1 \mathbf{D}^{-2} \cdot \frac{1}{\sqrt{T_1 T_2}} \end{aligned}$$

□

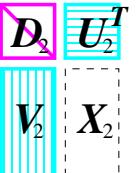
Bemerkungen

1. Sowohl die Berechnung von \mathbf{V}_1 und \mathbf{D} als auch die Auswertung der Approximationsformel nutzen die Datenobjekte ausschließlich in Form wechselseitiger innerer Produkte. Kerneltrick!
2. Wenn beide Datensätze mittelwertfrei sind, dann entsprechen die \mathbf{R}_i den Kovarianzmatrizen und die Zeilen der Matrizen \mathbf{V}_i enthalten die Hauptkomponentenvektoren (PCA) der Datenobjekte.

Nyström-Approximation

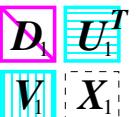
Problem

Für einen umfangreichen Datensatz $\mathbf{X} \in \mathbb{R}^{T \times N}$ sollen die Eigenwerte und Eigenvektoren der Gramsschen Matrix $\mathbf{G} = \mathbf{X} \mathbf{X}^\top$ berechnet werden. Aufwand $O(T^3)$



Lösung

Wir würfeln einen kleinen, repräsentativen Teildatensatz \mathbf{X}' aus und wenden den nachfolgenden Satz an:



Lemma (Williams & Seeger, 2001)

Es seien \mathbf{X}_i , $i = 1, 2$ zwei Datensätze der Größe T_i mit identischen ersten und zweiten Momenten.

Dann besitzen die skalierten Gramsschen Matrizen $\mathbf{G}_i / T_i = \mathbf{X}_i \mathbf{X}_i^\top / T_i$ dieselben (positiven) Eigenwerte d_1^2, \dots, d_r^2 und für die Eigenvektoren gilt

$$\mathbf{V}_2 = \mathbf{G}^* \cdot \mathbf{V}_1 \cdot \mathbf{D}^{-2} / \sqrt{T_1 T_2}, \quad \mathbf{G}^* = \mathbf{X}_2 \mathbf{X}_1^\top.$$

Nyström-Algorithmus

0 GEGEBEN

$$\mathbf{X}, \mathbf{Z}, K(\cdot, \cdot)$$

Datensatz $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{T_x})^\top$ und Teildatensatz $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_{T_z})^\top$ („Landmarkensatz“).

1 BERECHNE PARTIELLE KERNMATRIX

$$\mathbf{G}_{st} = K(\mathbf{z}_s, \mathbf{z}_t) \text{ für alle } s, t = 1, \dots, T_z$$

2 LÖSE EIGENWERTAUFGABE

$$\mathbf{G}\mathbf{v} = T_z \lambda \mathbf{v} \text{ mit Eigenwerten } T_z \lambda \in \mathbb{R} \text{ und Eigenvektoren } \mathbf{v} \in \mathbb{R}^S$$

3 BERECHNE EXTRAPOLATIONSMATRIX

$$\mathbf{G}_s^* = K(\mathbf{x}_s, \mathbf{z}_t) \text{ für alle } s = 1, \dots, T_x \text{ und } t = 1, \dots, T_z$$

4 EXTRAPOLIERE

$$\hat{\mathbf{v}}_m = \mathbf{G}^* \mathbf{v}_m / (d_m^2 \sqrt{T_z T_x}) \in \mathbb{R}^T$$

Nyström-Inkrementierung

Approximiere die Hauptkomponenten eines „Neuzugangs“

Aufgabenstellung

Für die Objekte $\mathbf{z}_1, \dots, \mathbf{z}_T \in \mathcal{X}$ wurden die Gramsche Matrix \mathbf{G}_ϕ sowie Eigenvektoren $\mathbf{v}_1, \dots, \mathbf{v}_M \in \mathbb{R}^T$ und Eigenwerte $\lambda_1, \dots, \lambda_M \in \mathbb{R}^+$ berechnet.

Gesucht ist der Zielvektor für ein neues Objekt $\mathbf{x} \in \mathcal{X}$.

Lösung ohne Zentrierung

Berechne den Vektor $\mathbf{g}(\mathbf{x})$ mit $g_t(\mathbf{x}) = K(\mathbf{z}_t, \mathbf{x})$ und setze

$$y_m = \frac{1}{\lambda_m} \cdot \mathbf{v}_m^\top \mathbf{g}(\mathbf{x}), \quad m = 1, \dots, M.$$

Lösung mit Zentrierung

An Stelle von \mathbf{x} ist das zentrierte Objekt (in \mathbb{H}) zu verwenden.

$$\langle \phi\mathbf{x} - \mu_\phi, \phi\mathbf{z}_t \rangle = \langle \phi\mathbf{x}, \phi\mathbf{z}_t \rangle - \frac{1}{T} \sum_{s=1}^T \langle \phi\mathbf{z}_s, \phi\mathbf{z}_t \rangle$$

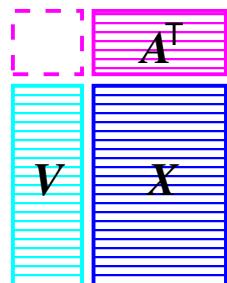
Datenanalyse als Matrixfaktorisierung

Objekt $\mathbf{x}_t \doteq$ Zutaten \mathbf{A} & Rezept \mathbf{v}_t

Lineares Erzeugungsmodell

Verborgene Faktoren + Zufallskomponenten

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A} \cdot \mathbf{v}_t + \boldsymbol{\varepsilon}_t \\ &= \sum_{m=1}^M v_{tm} \cdot \mathbf{a}_m + \boldsymbol{\varepsilon}_t \end{aligned}$$



- Ladungsvektoren $\mathbf{a}_1, \dots, \mathbf{a}_M \in \mathbb{R}^N$
- Latente Variablen $(v_{t1}, \dots, v_{tM})^\top = \mathbf{v}_t$
- Zufällige Fehler $(\varepsilon_{t1}, \dots, \varepsilon_{tN})^\top = \boldsymbol{\varepsilon}_t$
- Abhängige Variablen $(x_{t1}, \dots, x_{tN})^\top = \mathbf{x}_t$

Uneindeutigkeit:
 $\mathbf{V}' = \mathbf{V}\mathbf{D}$, $\mathbf{A}' = \mathbf{A}\mathbf{D}^{-1} \Rightarrow \mathbf{V}\mathbf{A}^\top = \mathbf{V}'\mathbf{A}'^\top$

Näherung
durch ein
rangdefizites
Produkt

$$\begin{aligned} \mathbf{X} &\approx \mathbf{V} \cdot \mathbf{A}^\top \\ \mathbf{X} &\in \mathbb{R}^{T \times N} \\ \mathbf{V} &\in \mathbb{R}^{T \times M} \\ \mathbf{A} &\in \mathbb{R}^{N \times M} \end{aligned}$$

Hauptachsentransformation

Mehrdimensionale Skalierung

Nichtnegative Matrixfaktorisierung

Independent Component Analysis

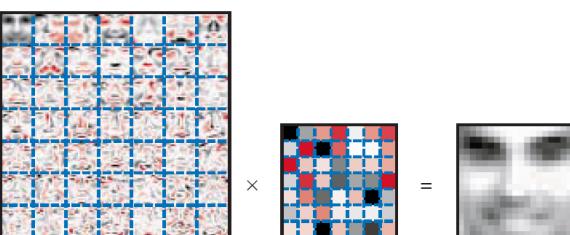
Faktorenanalyse

Merkmalkarte und autoassoziatives MLP

Explorative Datenanalyse & Grafik

Zwei Spezialfälle

Gesichtidentifikation — serialisierte Bildmatrix als Objekt



PCA
Hauptkomponenten

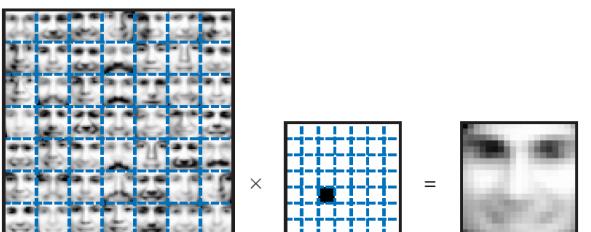
$$\begin{aligned} \mathbf{V}^\top \mathbf{V} &= \mathbf{E} \\ \mathbf{A}^\top \mathbf{A} &= \mathbf{D}^2 \end{aligned}$$

Lin.komb. „oszilliert“

VQ
Vektorquantisierung

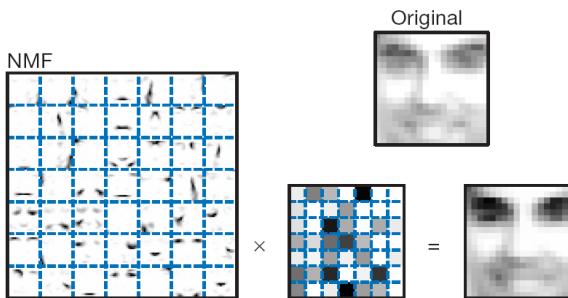
$$\mathbf{v}_t \in \{\mathbf{e}_1, \dots, \mathbf{e}_M\}$$

„winner takes all“



Nichtnegative Matrixfaktorisierung

(Lee & Seung, 1999)



NMF
Nichtnegative
Faktorisierung

Euklidische vs. statistische Faktorisierung

Definition

Es seien $T, N, M \in \mathbb{N}$ und $\mathbf{X} \in \mathbb{R}^{T \times N}$. Das Matrizenprodukt $\mathbf{V} \cdot \mathbf{A}^\top$ heißt **nichtnegative Faktorisierung** von \mathbf{X} mit Rang M bezüglich der **euklidischen-** bzw. der **Kullback-Leibler-Divergenz**, falls

$$\|\mathbf{X} - \mathbf{VA}^\top\|_{\text{Frob}}^2 \quad \text{bzw.} \quad \mathcal{D}(\mathbf{X} \parallel \mathbf{VA}^\top)$$

minimal ist unter den Bedingungen

$$\mathbf{V} \in \mathbb{R}^{T \times M}, \mathbf{A} \in \mathbb{R}^{N \times M}, \quad \mathbf{V} \geq \mathbf{0}, \mathbf{A} \geq \mathbf{0}.$$

Vorteile konvexer Zerlegung

- Ladungsvektoren sind als **Grauwertbildprototypen** interpretierbar.
- Linearkombination sind als deren **Überlagerung** interpretierbar.
- Optimalzerlegung tendiert zur **Dünnbesetzung** (von \mathbf{A} und \mathbf{V}).
- ☞ **NMF besitzt** $\left\{ \begin{array}{l} \text{dünne} \\ \text{verteilte} \end{array} \right\}$ Repräsentation — im Gegensatz zu $\left\{ \begin{array}{l} \text{PCA} \\ \text{VQ} \end{array} \right\}$

Euklidische Distanz
Symmetrischer Abstand

$$\underbrace{\sum_{i,j} (P_{ij} - Q_{ij})^2}_{\|\mathbf{P} - \mathbf{Q}\|_{\text{Frob}}^2}$$

Kullback-Leibler-Distanz
Nichtsymmetrischer Abstand

$$\underbrace{\sum_{i,j} \left(P_{ij} \cdot \log \frac{P_{ij}}{Q_{ij}} - P_{ij} + Q_{ij} \right)}_{\mathcal{D}(\mathbf{P} \parallel \mathbf{Q})}$$

Gradientenabstiegsverfahren I

Lemma (Lee & Seung)

Die euklidische Distanz $\|\mathbf{X} - \mathbf{VA}^\top\|^2$ ist schwach fallend unter den Auffrischungsregeln

$$V_{tm} \leftarrow V_{tm} \cdot \frac{(\mathbf{XA})_{tm}}{(\mathbf{VA}^\top \mathbf{A})_{tm}} \quad \text{und} \quad A_{nm} \leftarrow A_{nm} \cdot \frac{(\mathbf{X}^\top \mathbf{V})_{nm}}{(\mathbf{AV}^\top \mathbf{V})_{nm}} ;$$

Invarianz herrscht genau an den stationären Punkten (\mathbf{V}, \mathbf{A}).

Bemerkungen

- Die Frobeniusdistanz ist *konvex* in \mathbf{V} , in \mathbf{A} , aber *nicht* in (\mathbf{V}, \mathbf{A}) .
- Ex. i.a. mehrere lokale Minima ⇔ Optimierungsverfahren startwertabhängig.
- Der Rechenaufwand beträgt $O(TNM) + O(M^2N) + O(TM^2)$ (je Iterationsschritt), wenn $\mathbf{V} \cdot (\mathbf{A}^\top \mathbf{A})$ und $\mathbf{A} \cdot (\mathbf{V}^\top \mathbf{V})$ gerechnet wird.

Gradientenabstiegsverfahren II

Lemma (Lee & Seung)

Die Divergenz $\mathcal{D}(\mathbf{X} \parallel \mathbf{VA}^\top)$ ist schwach fallend unter den Auffrischungsregeln

$$V_{tm} \leftarrow V_{tm} \cdot \frac{\sum_n A_{nm} \cdot X_{tn}}{\sum_n (\mathbf{AV}^\top)_{nt}} \quad \text{und} \quad A_{nm} \leftarrow A_{nm} \cdot \frac{\sum_t V_{tm} \cdot X_{tn}}{\sum_t (\mathbf{AV}^\top)_{nt}} ;$$

Invarianz herrscht genau an den stationären Punkten (\mathbf{V}, \mathbf{A}).

Bemerkungen

- Auch für die Divergenz ist das Optimierungsverfahren i.a. startwertabhängig.
- Der Rechenaufwand beträgt $O(TNM)$ je Iterationsschritt.

Beweis.

BEWEISIDEE. Die Auffrischungsformeln selbst erhält man durch Ausrechnen der partiellen Ableitungen, Hinschreiben der Gradientenabstiegsformel und Wahl einer speziellen Schrittweite. Der Beweis der Abstiegseigenschaft basiert auf der Definition einer Hilfsfunktion — ähnlich wie beim EM-Prinzip — und ist nicht ganz kurz.

DISTANZ

$$V_{tm} \leftarrow V_{tm} + \eta_{tm} \cdot \{(\mathbf{XA})_{tm} - (\mathbf{VA}^\top \mathbf{A})_{tm}\}$$

Nach diagonaler Reskalierung der Variablen und mit den nachfolgenden Schrittweiten ergibt sich die Auffrischungsformel des ersten Lemmas.

$$\eta_{tm} \stackrel{\text{def}}{=} V_{tm} / (\mathbf{VA}^\top \mathbf{A})_{tm}$$

DIVERGENZ

$$V_{tm} \leftarrow V_{tm} + \eta_{tm} \cdot \left\{ \sum_n \frac{A_{nm} \cdot X_{tn}}{(\mathbf{AV}^\top)_{nt}} - \sum_n A_{nm} \right\}$$

Nach diagonaler Reskalierung der Variablen und mit den nachfolgenden Schrittweiten ergibt sich die Auffrischungsformel des zweiten Lemmas.

$$\eta_{tm} \stackrel{\text{def}}{=} V_{tm} / \sum_n A_{nm}$$

□

Beispiel — Information Retrieval

Objekte \times Attribute $\hat{=}$ Dokumente \times Worteinträge

Vektorraumrepräsentation(en)

$$\mathbf{X} = \begin{pmatrix} & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & \dots & w_{4711} \\ \hline d_1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ d_2 & 0 & 2 & 0 & 1 & 1 & 0 & 0 & \dots & 1 \\ d_3 & 0 & 0 & 0 & 1 & 2 & 0 & 1 & \dots & 0 \\ d_4 & 0 & 0 & 3 & 1 & 0 & 8 & 0 & \dots & 0 \\ d_5 & 1 & 0 & 0 & 0 & 4 & 2 & 0 & \dots & 0 \\ d_6 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & \dots & 0 \\ d_7 & 5 & 1 & 0 & 3 & 0 & 1 & 2 & \dots & 0 \\ d_8 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ d_9 & 0 & 0 & 1 & 0 & 2 & 3 & 0 & \dots & 0 \\ d_{10} & 0 & 0 & 1 & 1 & 1 & 0 & 1 & \dots & 0 \\ d_{11} & 0 & 1 & 0 & 0 & 2 & 0 & 1 & \dots & 1 \\ d_{12} & 1 & 0 & 0 & 1 & 1 & 4 & 2 & \dots & 0 \\ \vdots & \dots & \vdots \\ d_{882} & 0 & 1 & 1 & 0 & 7 & 3 & 1 & \dots & 0 \\ d_{883} & 0 & 0 & 1 & 0 & 1 & 2 & 0 & \dots & 0 \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} q_1=0 \\ q_2=0 \\ q_3=0 \\ \vdots \\ q_{1492}=0 \\ q_{1493}=1 \\ q_{1494}=0 \\ \vdots \\ q_{1967}=0 \\ q_{1968}=1 \\ q_{1969}=0 \\ \vdots \\ q_{4710}=0 \\ q_{4711}=0 \end{pmatrix}$$

Beispiel — Information Retrieval

Faktorisierungstechnik und Dokumentrepräsentation

Boolesches Retrieval

- IR sucht nach Stichworten oder Stichwortkombination.
- Scheitert an *Polysemen* und *Pleonasmen*.

Thematische Dokumentengruppen

- Die Dokumente sind *monothematisch*.
- a_m beschreibt prototypische Worthäufigkeitsverteilung.

Latente semantische Indexierung

- Repräsentation v_t enthält „*semantische Variablen*“.
- Negative Wortwerte & negative Themenbeiträge? — kontraintuitiv!

Dünne Themencharakterisierung

- Jedes Thema ist charakterisiert durch „dünnen“ Teilwortschatz.
- Jedes Dokument überstreicht nur wenige treffsichere Themen.

BIR

VQ

PCA

NMF

Gegeben

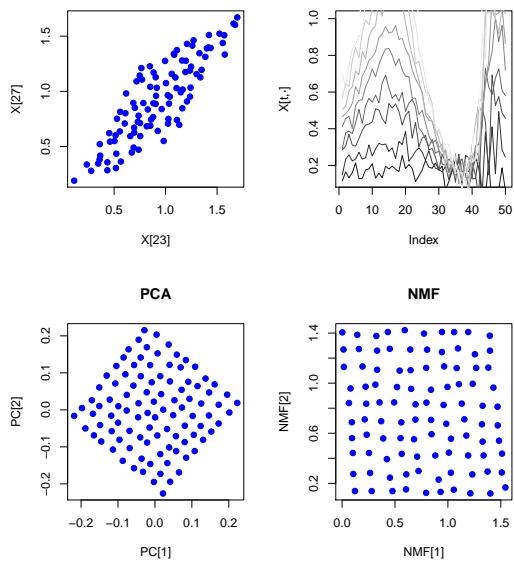
- die Rang- M -Faktorisierung $\mathbf{V} \cdot \mathbf{A}^\top$ eines Datensatzes \mathbf{X}
- ein neuer Datenvektor $\mathbf{x} \in \mathbb{R}^N$

Gesucht

Nichtnegativer Gewichtvektor $\mathbf{v} \in \mathbb{R}^M$ mit minimalem Rekonstruktionsfehler $\|\mathbf{x} - \mathbf{Av}\|^2$ bzw. $\mathcal{D}(\mathbf{x} \| \mathbf{Av})$

-
- | | |
|--|--|
| <p>1 INITIALISIERUNG
Berechne $\mathbf{A}^\top \mathbf{A}$ und $\mathbf{A}^\top \mathbf{x}$.
Setze $\mathbf{v}^{(0)} = \frac{1}{M} \mathbf{1}$.</p> <p>2 ITERATIONSSCHRITT
Komponentenweise multipl.:
$\mathbf{v}^{(i+1)} \leftarrow \mathbf{v}^{(i)} \cdot (\mathbf{A}^\top \mathbf{x}) / (\mathbf{A}^\top \mathbf{A} \cdot \mathbf{v}^{(i)})$</p> <p>3 TERMINIERUNG
Eind. Lösung $\mathbf{v}^{(\infty)}$ (Konvexität)</p> | <p>1 INITIALISIERUNG
Berechne $\mathbf{A}^\top \mathbf{1}$ und $\mathbf{v}^{(0)} = \frac{1}{M} \mathbf{1}$.</p> <p>2 ITERATIONSSCHRITT
Berechne für alle m:
$v_m^{(i+1)} \leftarrow v_m^{(i)} \cdot \left(\frac{\mathbf{a}_m}{\mathbf{A} \mathbf{v}^{(i)}} \right)^\top \mathbf{x} / (\mathbf{A}^\top \mathbf{1})_m$</p> <p>3 TERMINIERUNG
Eind. Lösung $\mathbf{v}^{(\infty)}$ (Konvexität)</p> |
|--|--|
-

Beispiel — stöchiometrische Gemengelagen



Synthese-Daten

Zwei Vektoren (\mathbb{R}^{50}) mit stark verrauschten Sinuskurven werden 100-mal überlagert:

$$\mathbf{x}_{10i+j} = i \cdot \mathbf{a}_1 + j \cdot \mathbf{a}_2$$

Problem

Wer findet die latenten Faktoren i, j ?

PCA

2D-Unterraum korrekt

NMF

2D-Unterraum korrekt
Achsen $\{y_1 = i\}$ korrekt
 $\{y_2 = j\}$

Probabilistische Deutung der Datenmatrix

Zweistufiger, diskreter Mischverteilungsprozeß

$$\bullet \quad \mapsto \quad \mathbb{A}_n \in \{\alpha_1, \dots, \alpha_M\} \quad \mapsto \quad \mathbb{X}_n \in \{\xi_1, \dots, \xi_T\}$$

Attribute als Objektselektoren

Spaltenweise stochastische Datenmatrix mit Wahrscheinlichkeiten

$$\underbrace{P(\mathbb{X}_n = \xi_t)}_{x_{tn}} = \sum_{m=1}^M P(\mathbb{A}_n = \alpha_m, \mathbb{X}_n = \xi_t) \\ = \sum_{m=1}^M \underbrace{P(\mathbb{A}_n = \alpha_m)}_{= a_{nm}} \underbrace{P(\mathbb{X}_n = \xi_t | \mathbb{A}_n = \alpha_m)}_{= w_{ntm} \approx v_{tm}} \stackrel{!}{=} (\mathbf{V} \mathbf{A}^\top)_{tn}$$

1. Zeilennormiertes \mathbf{A} mit N Faktormischungen.
2. Spaltennormiertes \mathbf{V} mit M Objektverteilungen.

Konvexe Zerlegung mit dem EM-Prinzip

Maximum-Likelihood-Schätzung der \mathbf{V} - und \mathbf{A} -Einträge

Minimiere Kreuzentropie zwischen \mathbf{X} und $\mathbf{P} = \mathbf{VA}^\top$

$$\mathcal{H}(\mathbf{X}, \mathbf{P}) = -\log P(\mathbf{X}|\mathbf{P}) = -\sum_{t=1}^T \sum_{n=1}^N x_{tn} \cdot \log \underbrace{\sum_{m=1}^M a_{nm} v_{tm}}_{p_{tn} = (\mathbf{V} \mathbf{A}^\top)_{tn}}$$

Bemerkungen

1. $\mathcal{H}(\mathbf{X}, \mathbf{P})$ bildet die Summe aller attributbezogenen Kreuzentropien.
2. Die ML-Güefunktion ist konvex und garantiert ein globales Optimum.
3. Iterative Optimierung mit *Expectation-Maximization*-Algorithmus

Objekte als Attributselektoren

Duales Mischverteilungsmodell nach Rollentausch

$$\underbrace{P(\mathbb{X}_t = \xi_n)}_{x_{tn}} = \sum_{m=1}^M \underbrace{P(\mathbb{X}_t = \xi_n | \mathbb{V}_n = \phi_m)}_{= b_{tnm} \approx a_{nm}} \cdot \underbrace{P(\mathbb{V}_t = \phi_m)}_{= v_{tm}} \stackrel{!}{=} (\mathbf{V} \mathbf{A}^\top)_{tn}$$

Mischungsidentifikation mit EM-Algorithmus

(Algorithmus)

0 INITIALISIERUNG

Normiere \mathbf{X} spaltenweise stochastisch.
Berechne Startwerte \mathbf{V}, \mathbf{A} (Seung-Algorithmus).

1 EXPECTATION

Berechne a posteriori Verteilung der Variablen \mathbb{A}_n

$$\gamma_t^{(n)}(m) = P(\mathbb{A}_n = m | \mathbb{X}_n = t) = \frac{a_{nm} \cdot v_{tm}}{\sum_{\mu=1}^M a_{n\mu} \cdot v_{t\mu}}$$

2 MAXIMIZATION

Neuschätzung der Wahrscheinlichkeitsparameter

$$\hat{a}_{nm} \propto \sum_{t=1}^T x_{tn} \cdot \gamma_t^{(n)}(m) \quad \text{und} \quad \hat{v}_{tm} \propto \sum_{n=1}^N x_{tn} \cdot \gamma_t^{(n)}(m)$$

3 TERMINIERUNG

Abbruch oder weiter bei Schritt 1.

(summingA)

Hauptachsentransformation

Mehrdimensionale Skalierung

Nichtnegative Matrixfaktorisierung

Independent Component Analysis

Faktorenanalyse

Merkmalkarte und autoassoziatives MLP

Explorative Datenanalyse & Grafik

Independent Component Analysis

Definition

Es sei $\mathbf{A} \in \mathbb{R}^{N \times N}$ invertierbar und es bezeichne \mathbb{V} ein Tupel von N statistisch unabhängigen Zufallsvariablen $\mathbb{V}_1, \dots, \mathbb{V}_N$. Das lineare Mischungssystem

$$\mathbf{X} = \mathbf{A} \cdot \mathbb{V}$$

heißt **ICA-Modell mit den latenten Variablen \mathbb{V}** .

Das Umkehrsystem

$$\mathbb{V} = \mathbf{W} \cdot \mathbf{X}, \quad \mathbf{W} \stackrel{\text{def}}{=} \mathbf{A}^{-1}$$

heißt **Quellentrennungsmodell**.

Bemerkungen

1. ICA-Modell in Matrixschreibweise:
 $\mathbf{X} = \mathbf{V}\mathbf{A}^\top$ mit $M = N$
2. ICA fordert die **Unabhängigkeit** der \mathbb{V} -Spalten, PCA fordert die **Unkorreliertheit**.

Mehrdeutigkeit der ICA-Zerlegung

Reihenfolge der \mathbb{V}_m
Varianzen der \mathbb{V}_m
Vorzeichen der \mathbb{V}_m

Das „Cocktailparty-Problem“

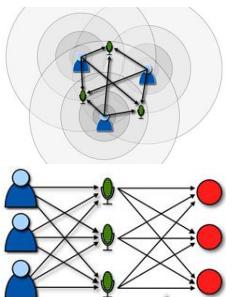
Selektives Anbaggen nach Signalentflechtung mit Hilfe des binauralen Gehörs

Blinde Quellentrennung

$$\begin{aligned} x_1(t) &= a_{11} \cdot v_1(t) + a_{12} \cdot v_2(t) \\ x_2(t) &= a_{21} \cdot v_1(t) + a_{22} \cdot v_2(t) \end{aligned}$$

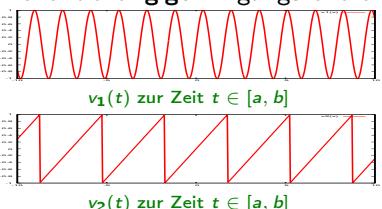
bzw.

$$\mathbf{x}(t) = \underbrace{\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}}_{\mathbf{A}} \cdot \mathbf{v}(t)$$



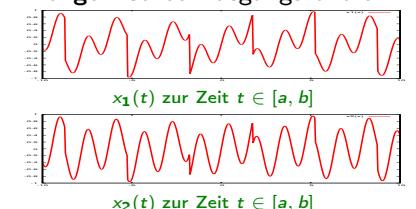
GESUCHT

Zwei **unabhängige** Eingangskanäle



GEGEBEN

Zwei **gemischte** Ausgangskanäle

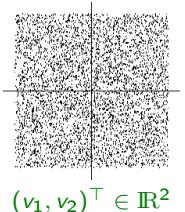


Beispiel — ICA versus PCA

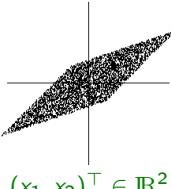
PCA berücksichtigt nur statistische Abhängigkeiten zweiter Ordnung

Uniforme Quellenverteilung & Mischungsmatrix

$$f_{\mathbb{V}_1}(r) = f_{\mathbb{V}_2}(r) = \begin{cases} 1/\sqrt{12} & |r| \leq \sqrt{3} \\ 0 & \text{sonst} \end{cases}, \quad \mathbf{A} = \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}$$



$$(\mathbb{V}_1, \mathbb{V}_2)^\top \in \mathbb{R}^2$$



$$(\mathbf{x}_1, \mathbf{x}_2)^\top \in \mathbb{R}^2$$

Bemerkungen

1. Der Wert $1/\sqrt{12}$ ergibt sich aus der Normierungsbedingung, die Grenzen $\pm\sqrt{3}$ des Verteilungsträgers garantieren $\mu = 0$ und $\sigma^2 = 1$.
2. Offensichtlich ist eine PCA nicht in der Lage, aus (x_1, x_2) die latenten Variablen $(\mathbb{V}_1, \mathbb{V}_2)$ als Hauptachsen zu extrahieren.

Was ist statistische Unabhängigkeit ?

Unabhängig $\hat{=}$ faktorisierbar

$$f_{\mathbb{V}}(\mathbf{v}) = \prod_m f_{\mathbb{V}_m}(v_m)$$

$$\mathcal{E}\left[\prod_m h_m(\mathbb{V}_m)\right] = \prod_m \mathcal{E}[h_m(\mathbb{V}_m)]$$

Momente q -ter Ordnung
unabhängig $\begin{cases} \Rightarrow \\ \neq \end{cases}$ unkorreliert

Spezialfall Normalverteilung
unabhängig \Leftrightarrow unkorreliert

Normale Faktoren ?!
Latente Variablen

$$\mathbb{V}_m \sim \mathcal{N}(0, 1)$$

Gemeinsame Verteilung

$$\mathbb{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{E})$$

Mischung

$$\mathbb{X} = \mathbf{A}\mathbb{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{AA}^\top)$$

Zusätzliche Rotationen

$$\mathbf{B} = \mathbf{AU} \quad \mathbf{BB}^\top = \mathbf{AA}^\top$$

sind grundsätzlich nicht identifizierbar!

Beweis.

Für den maximal nicht-normalen Faktor y gilt die Darstellung:

$$y = \mathbf{c}^\top \mathbf{x} = \mathbf{c}^\top \mathbf{Av} = \mathbf{z}^\top \mathbf{v} \quad \text{mit} \quad \mathbf{z} \stackrel{\text{def}}{=} \mathbf{A}^\top \mathbf{c}$$

Es ist y also eine Linearkombination der latenten Variablen v_n . Weil jede Linearkombination mindestens so normal ist wie jeder ihrer Summanden (ZGWS), muß eine maximal nicht-normale Kombination entartet sein:

- ⇒ Es ist $y = \alpha \cdot v_\ell$ für ein $\ell \in \{1, \dots, N\}$.
- ⇒ Es gilt $\mathbf{z} = \alpha \cdot \mathbf{e}_\ell$, wobei \mathbf{e}_n den n -ten Einheitsvektor des \mathbb{R}^N bezeichne.
- ⇒ Es ist $\mathbf{c} = \frac{1}{\sqrt{\alpha}} \cdot \mathbf{b}_\ell$, und \mathbf{b}_ℓ ist die ℓ -te Spalte von $(\mathbf{A}^\top)^{-1}$.
- ⇒ Also ist \mathbf{b}_ℓ die ℓ -te Zeile von \mathbf{A}^{-1} und damit von \mathbf{W} .

□

Bemerkung

Der Koeffizientenvektor \mathbf{w}_ℓ ist bestenfalls (Reihenfolge, Betrag) bis auf das Vorzeichen eindeutig. Für die Matrix \mathbf{W} gibt es also mindestens 2^N lokale Optima!

Das ist eine schlechte Nachricht für alle Gesamtschrittverfahren („symmetrische Dekorrelation“).

Die ICA-Heuristik

Zentraler Grenzwertsatz

Die Summe unabhängiger ZV ist **asymptotisch** normalverteilt.
Die Summe unabhängiger ZV ist „normaler“ als jede Komponente.

Lemma

Sei $\mathbb{X} = \mathbf{A} \cdot \mathbb{V}$ ein ICA-Modell und sei $\mathbf{c} \in \mathbb{R}^N$ ein Koeffizientenvektor, so daß die Linearkombinationen

$$y = \mathbf{c}^\top \mathbf{x}$$

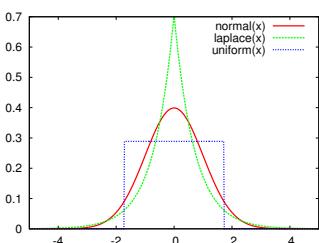
von \mathbb{X} -Realisierungen **maximal nicht-normal** verteilt sind.

Dann gleicht \mathbf{c} (dem Vielfachen) einer Zeile der Quellentrennungsmatrix $\mathbf{W} = \mathbf{A}^{-1}$.

-
- Algorithmus**
- 0 GEGEBEN Datenmatrix $\mathbf{X} \in \mathbb{R}^{T \times N}$, Startindex $n \leftarrow 1$.
 - 1 OPTIMALE PROJEKTION Berechne maximal nicht-normales \mathbf{w}_n .
 - 2 DEFLATIEREN Hierarchische Dekorrelation $\mathbf{X} \leftarrow \mathcal{D}(\mathbf{X}; \mathbf{w}_1, \dots, \mathbf{w}_n)$
 - 3 TERMINIERUNG Setze $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_N)^\top$ und $\mathbf{V} = \mathbf{X} \mathbf{W}^\top$ bzw. setze $n \leftarrow n + 1$ und \rightsquigarrow 1.
-
- lumdringA

Kriterien für Nicht-Normalität

Kumulant — Mittel, Streuung, Schiefe und Abklingverhalten



Bemerkung

Neben Mittelwert $\mu = \mathcal{E}[\mathbb{Y}]$ und Varianz $\sigma^2 = \mathcal{E}[(\mathbb{Y} - \mu)^2]$ bietet auch die Schiefe

$$\text{skew}(\mathbb{Y}) \stackrel{\text{def}}{=} \mathcal{E}[\tilde{\mathbb{Y}}^3], \quad \tilde{\mathbb{Y}} = \frac{\mathbb{Y} - \mu}{\sigma},$$

ein momentbezogenes Gestaltmerkmal; für gipfelsymmetrische Verteilungen wie $\mathcal{N}(\mu, \sigma^2)$ ist ihr Wert Null.

Definition

Es sei \mathbb{Y} eine Zufallsvariable. Das zentrierte Moment

$$\text{kurt}(\mathbb{Y}) \stackrel{\text{def}}{=} \mathcal{E}[\tilde{\mathbb{Y}}^4] - 3 \cdot (\mathcal{E}[\tilde{\mathbb{Y}}^2])^2$$

heißt **Kurtosis** von \mathbb{Y} .

Die Zufallsvariable heißt **supernormal** (leptokurt) bzw. **subnormal** (platykurt), wenn $\text{kurt}(\mathbb{Y})$ größer bzw. kleiner als Null ist.

Eigenschaften der Kurtosis

Die Normalverteilungsdichte ist normal!

Lemma

1. Für normalverteiltes $\mathbb{X} \sim \mathcal{N}(0, \sigma^2)$ gilt $\mathbb{E}[\mathbb{X}^4] = 3\sigma^4$. $\text{kurt}(\mathbb{X}) = 0$
2. Die Laplace-Dichte $\mathcal{L}(\lambda) = \frac{1}{2\lambda} e^{-\lambda|x|}$ ist supernormal.
3. Die uniforme Dichte ist subnormal. $\sigma^2 = 1 \Rightarrow \text{kurt}(\mathbb{X}) = -\frac{6}{5}$
4. Sind \mathbb{X}, \mathbb{Y} statistisch unabhängig, so gilt:

$$\text{kurt}(\mathbb{X} + \mathbb{Y}) = \text{kurt}(\mathbb{X}) + \text{kurt}(\mathbb{Y})$$

5. Für $\alpha \in \mathbb{R}$ gilt $\text{kurt}(\alpha \cdot \mathbb{X}) = \alpha^4 \cdot \text{kurt}(\mathbb{X})$.

Maximierung der Nicht-Normalität

Betont Dichteränder, daher empfindlich gegenüber Ausreißern:

$$\text{kurt}(y) = \text{kurt}(\mathbf{c}^\top \mathbf{x}) = \text{kurt}(\mathbf{z}^\top \mathbf{v}) = \sum_{n=1}^N z_n^4 \cdot \text{kurt}(\mathbb{V}_n) \xrightarrow{\text{!}} \text{MAX/MIN}$$

Beweis.

(zur Subnormalität der Gleichverteilung)

Die $(0,1)$ -uniforme Dichte besitzt die Trägermenge $[-\sqrt{3}, +\sqrt{3}]$ und dort den Dichtewert $1/\sqrt{12}$. Wir berechnen das vierte Moment:

$$\begin{aligned} \mathbb{E}[\mathbb{Y}^4] &= \int_{-\infty}^{+\infty} y^4 f_u(y) dy = \int_{-\sqrt{3}}^{+\sqrt{3}} \frac{y^4}{\sqrt{12}} dy = \left. \frac{y^5}{10\sqrt{3}} \right|_{-\sqrt{3}}^{+\sqrt{3}} \\ &= \frac{2 \cdot (\sqrt{3})^5}{10 \cdot \sqrt{3}} = \frac{(\sqrt{3})^4}{5} = \frac{9}{5} \end{aligned}$$

Der Wert 1.8 ist kleiner als $3 \cdot (\sigma^2)^2 = 3$, also ist die Kurtosis negativ. \square

Bemerkung

Zur IC-Optimierung ist die y -Kurtosis möglichst weit von Null zu entfernen; es ist also der Betrag $|\text{kurt}(y)|$ oder das Quadrat $(\text{kurt}(y))^2$ zu maximieren.

In beiden Fällen ist die Zielfunktion ein Polynom **achten Grades** in den Datenwerten!

Die Darstellung von $\text{kurt}(y)$ als Kombination der z_n^4 zeigt, daß mit zahlreichen lokalen Minima/Maxima zu rechnen ist.

Differentielle Entropie

Die Gaußglocke — „gerechte“ Verteilung auf der reellen Achse

Definition

Sei \mathbb{Y} eine Zufallsvariable mit der Verteilungsdichtefunktion $f_{\mathbb{Y}}(\cdot)$. Der Erwartungswert

$$\mathcal{H}(\mathbb{Y}) \stackrel{\text{def}}{=} \mathbb{E}[-\log f(\mathbb{Y})] = - \int f(y) \cdot \log f(y) dy$$

heißt **(differentielle) Entropie** von \mathbb{Y} bzw. von $f_{\mathbb{Y}}$.

Lemma

Unter allen Verteilungsfunktionen mit Erwartungswert μ und Varianz σ^2 hat die Normalverteilungsdichte $\mathcal{N}(\mu, \sigma^2)$ die maximale Entropie.

Bemerkung

Wegen der inhärenten Skalierungsabhängigkeit dieser Aussage reicht es nicht (ganz) aus, zwecks Maximierung der Nicht-Normalität die Entropie zu minimieren.

Negative Entropie

Definition

Sei \mathbb{Y} eine Zufallsvariable mit der Varianz σ_y^2 . Dann heißt

$$\mathcal{J}(\mathbb{Y}) \stackrel{\text{def}}{=} \mathcal{H}(\mathcal{N}(0, \sigma_y^2)) - \mathcal{H}(\mathbb{Y})$$

die **negative Entropie** von \mathbb{Y} .

Lemma

1. Die negative Entropie $\mathcal{J}(\mathbb{Y})$ ist invariant gegenüber linearen Transformationen von \mathbb{Y} .
2. Für unkorrelierte Zufallsvariablen ist die negative Entropie antiton zur Transformation:

$$\mathfrak{S}(\mathbb{Y}_1, \dots, \mathbb{Y}_N) = \mathcal{J}(\mathbb{Y}) - \sum_{n=1}^N \mathcal{J}(\mathbb{Y}_n) + \frac{1}{2} \log \frac{\prod_{n=1}^N S_{nn}^Y}{\det \mathbf{S}^Y}$$

\Rightarrow (1) Achsen normieren (2) Hauptkomponenten (3) Neg. Entropie maximieren

Näherungsformeln für $\mathcal{J}(\mathbb{Y})$

Problem

Es ist $\mathcal{H}(\mathcal{N}(0, \sigma_y^2)) = \frac{1}{2} \cdot \log(2\pi e \sigma^2)$, aber wie groß ist $\mathcal{H}(\mathbb{Y})$?

Approximation durch Kumulanten

Empirische dritte und vierte Kumulanten \rightsquigarrow Polynom 8. Grades

$$\mathcal{J}(\mathbb{Y}) \approx \frac{1}{12} (\text{skew}(\mathbb{Y}))^2 + \frac{1}{48} (\text{kurt}(\mathbb{Y}))^2$$

Proportionalität mit Glockenformdevianz

Schwerpunktvergleich zwischen verzerrter Gauß- und Datenverteilung

$$\mathcal{J}(\mathbb{Y}) \propto [\mathcal{E}[g(\tilde{\mathbb{Y}})] - \mathcal{E}[g(\mathbb{E})]]^2, \quad \tilde{\mathbb{Y}} = \frac{\mathbb{Y} - \mu_{\mathbb{Y}}}{\sigma_{\mathbb{Y}}}, \quad \mathbb{E} \sim \mathcal{N}(0, 1)$$

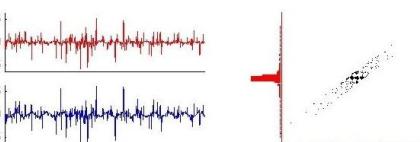
Das trifft zu für praktisch alle nichtquadratischen Funktionen, z.B.:

$$g_1(u) = \log \cosh(a_1 u), \quad g_2(u) = -\exp(-a_2 \cdot \frac{u^2}{2})$$

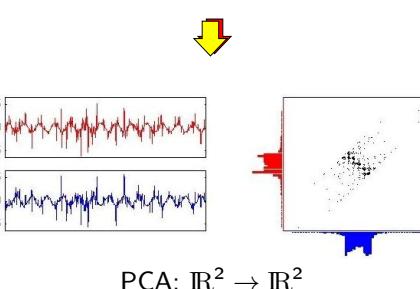
Der FastICA-Algorithmus

Beispielentflechtung für $N = M = 2$ — aus: Hyvärinen & Oja

Dekorrelierung der Quelle

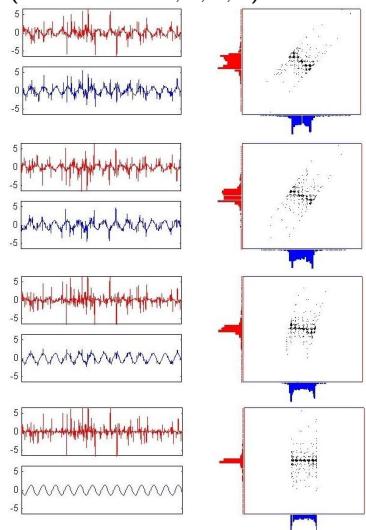


Quellen $x_1(t)$, $x_2(t)$ und \mathbb{R}^2 -Dichte



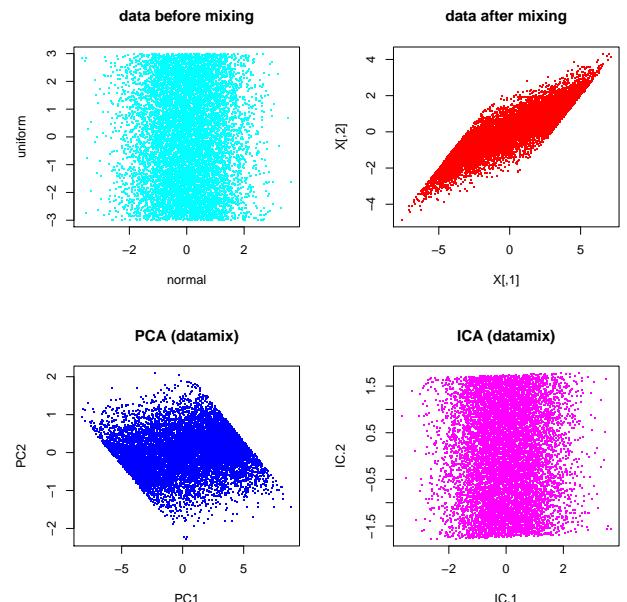
PCA: $\mathbb{R}^2 \rightarrow \mathbb{R}^2$

Gesamtschritt-Iteration (Schritte $i = 1, 2, 3, 5$)



Unkorreliert oder unabhängig?

Entmischung gleichverteilter und normalverteilter Komponenten



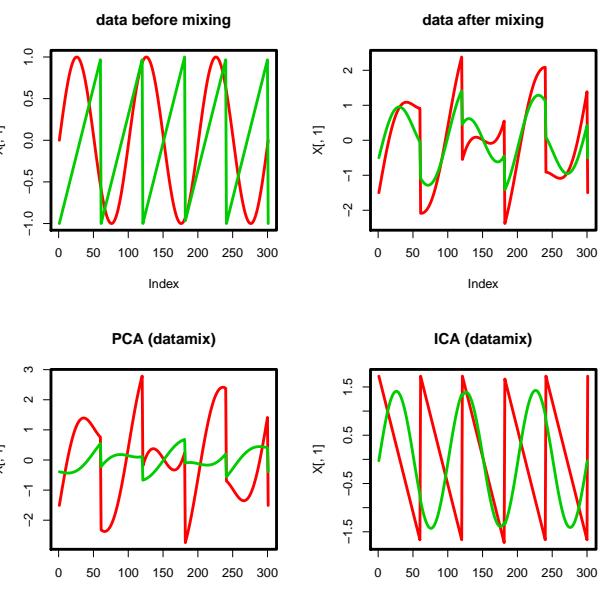
R-Code

```
layout(matrix (
  1:4, 2, 2, T))
require (ICS) # für 'ics'
par (pch=".")
n <- 10000
X <- cbind (
  normal=rnorm(n),
  uniform=runif(n,-3,+3))
A <- rbind (
  c(1,1.5),
  c(1,0.5))

plot (X,
  col="cyan",
  main="data before mixing")
X <- X %*% t(A)
plot (X,
  col="red",
  main="data after mixing")
plot (prcomp(X)$x,
  col="blue",
  main="PCA (datamix)")
plot (ics(X)$Scores,
  col="magenta",
  main="ICA (datamix)")
```

Zeitreihen mit nichtnormalen Werteverteilungen

Entmischung von Sinuswelle und Sägezahn auf Grundlage abweichender Kurtosiswerte



R-Code

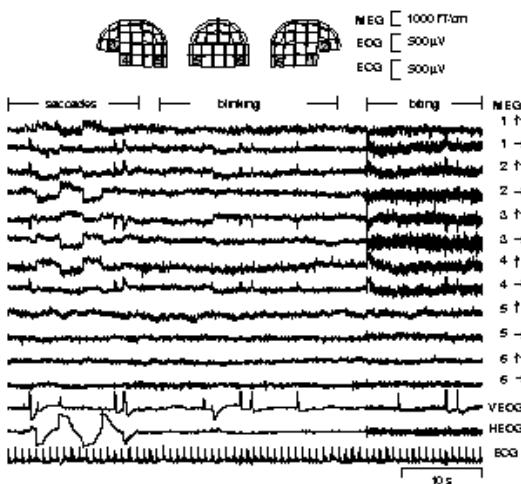
```
layout(matrix (1:4,2,2,T))
require (ICS) # für 'ics'
n <- 300
s <- 0:n/n
X <- cbind (
  x=sin(2*pi*3*s),
  y=s%/%0.2/0.1-1)
A <- rbind (2:3, 2:1) / 2

tsplot <- function (
  X, Color=c(2,3), ...) {
  plot (X[,1], type="l",
    ylim=range(X),
    col=Color[1], ...)
  lines (X[,2],
    col=Color[2])
}

tsplot (X,
  main="data before mixing")
X <- X %*% t(A)
tsplot (X,
  main="data after mixing")
tsplot (prcomp(X)$x,
  main="PCA (datamix)")
tsplot (ics(X)$Scores,
  main="ICA (datamix)")
```

Beispiel — Magnetoenzephalogramm (MEG)

Vigario et al. (1998)



MEG-Ableitung

61 Meßorte (Skalp)

Versuchsperson

- Blinzeln
- horizontale Sakkaden
- Zähneknirschen

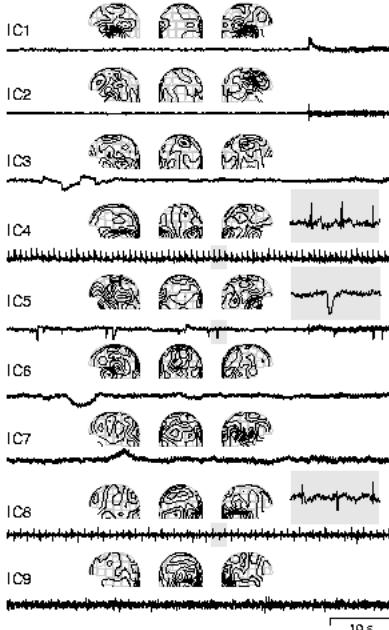
MEG-Kanäle

Auswahl 12 ∈ 122

- frontal
- temporal
- okzipital

Beispiel — Magnetoenzephalogramm (MEG)

Verarbeitungsschritte und Resultate



Vorbereitung

Zentrierung ($\mu = 0$)
PCA ($\Phi = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^T$)

Dimensionsreduktion

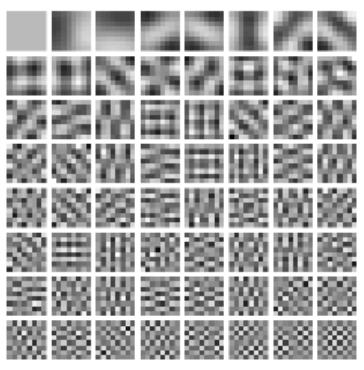
$\mathbb{R}^{122} \rightarrow \mathbb{R}^9$ via PCA
 $\mathbb{R}^9 \rightarrow \mathbb{R}^9$ via ICA

Komponentendeutung

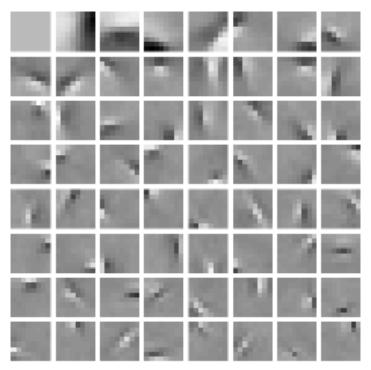
- v_1, v_2 Zähne
- v_3, v_5 Augen
- v_4 Herzzyklus
- v_8 Digitaluhr
- v_9 Sensorstörung

Beispiel — Bilddatenanalyse

Gleitende (8×8) -Grauwertbildblöcke zur Rauschunterdrückung (Hoyer 1999)



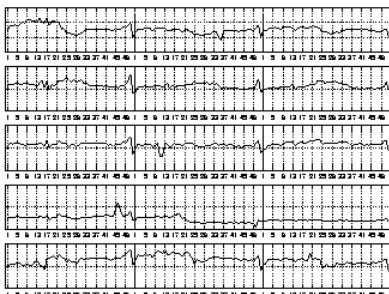
PCA: $\mathbb{R}^{8 \times 8} \rightarrow \mathbb{R}^{8 \times 8}$



ICA: $\mathbb{R}^{8 \times 8} \rightarrow \mathbb{R}^{8 \times 8}$

Beispiel — Betriebswirtschaftliche Anwendung

Cash-flow-Studien (Kiviluoto & Oja, 1998)



Datensammlung

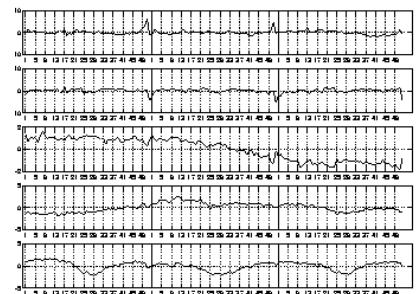
Cash Flow in 40 Geschäften über 140 Wochen hinweg (5 Beispielverläufe)

Verarbeitungspipeline

$\mathbb{R}^{40} \rightarrow \text{PCA} \rightarrow \mathbb{R}^5 \rightarrow \text{ICA} \rightarrow \mathbb{R}^5$

Unabhängige Komponenten

- v_1, v_2 Weihnachtsgeschäft · Feiertage
- v_5 Saisonale Effekte · glatte Mittelfristwirkungen · Sommerferienflaute
- v_3 Langfristverhalten · „Trend“
- v_4 konkurrenzdruckbezogene Geschäftsentwicklung (?)



Invariante Koordinatensysteme (ICS)

Tyler, Critchley, Dümbgen, Oja (2008)

Die PCA ist weder affin-invariant noch robust

- A $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{A}^\top \Rightarrow \tilde{\mathbf{S}} = \mathbf{A}\mathbf{S}\mathbf{A}^\top = \mathbf{A}\mathbf{U}\mathbf{D}\mathbf{U}^\top\mathbf{A}^\top \Rightarrow \tilde{\mathbf{V}} = \mathbf{X}\mathbf{A}^\top \cdot \mathbf{A}\mathbf{U}\mathbf{D}^{-1/2}$
- B PCA beruht auf „nervösen“ Statistiken $\mu = \tau(\mathbf{X})$ und $\mathbf{S} = \sigma(\mathbf{X})$

ICS für Streuungsstatistiken σ_1, σ_2

Mit den Eigenzerlegungen $\mathbf{S}_1 \hat{=} \mathbf{U}_1 \mathbf{D}_1 \mathbf{U}_1^\top$ und $\mathbf{B}_2 \hat{=} \mathbf{U}_2 \mathbf{D}_2 \mathbf{U}_2^\top$ gilt:

\mathbf{X}	$\xrightarrow{\mathbf{B}_1}$	\mathbf{X}'	$\xrightarrow{\mathbf{U}_2^\top}$	\mathbf{X}''	(zwei Stufen)
$\sigma_1 : \mathbf{S}_1$		$\mathbf{B}_1 \mathbf{S}_1 \mathbf{B}_1^\top = \mathbf{E}$		$\mathbf{U}_2^\top \mathbf{E} \mathbf{U}_2 = \mathbf{E}$	standardisiert
$\sigma_2 : \mathbf{S}_2$		$\mathbf{B}_1 \mathbf{S}_2 \mathbf{B}_1^\top =: \mathbf{B}_2$		$\mathbf{U}_2^\top \mathbf{B}_2 \mathbf{U}_2 = \mathbf{D}_2$	dekorreliert

Spezialfälle

	PCA	FDA	ICS	ICA
$\sigma_1 :$	\mathbf{S}	\mathbf{S}_W	Zentrum/Streuung	\mathbf{S}
$\sigma_2 :$	\mathbf{S}	\mathbf{S}_B	Schiefe/Wölbung	\mathbf{S}_4 (Wölbung)

Hauptachsentransformation

Mehrdimensionale Skalierung

Nichtnegative Matrixfaktorisierung

Independent Component Analysis

Faktorenanalyse

Merkmalkarte und autoassoziatives MLP

Explorative Datenanalyse & Grafik

Faktorenanalytische Zerlegung

$$\mathbb{X} = \mathbf{A}\mathbb{V} + \mathbb{E}$$

Daten = M Faktoren + Störung

- Datenvektor $\mathbb{X} \in \mathbb{R}^N$
- Störvektor $\mathbb{E} \in \mathbb{R}^N$
- Faktorenvektor $\mathbb{V} \in \mathbb{R}^M$
- Ladungsvektoren $\mathbf{a}_1, \dots, \mathbf{a}_M \in \mathbb{R}^N$
- Entkopplung von Faktoren und Störungen

Normalverteilte Daten

$$\mathbb{X} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \boldsymbol{\nu}, \mathbf{A}\mathbf{Q}\mathbf{A}^\top + \mathbf{R})$$

Wahlfreiheit

wg. $\mathbf{Q} = \mathbf{U}\mathbf{D}^2\mathbf{U}^\top$
Entweder \mathbf{Q} Einheitsmatrix oder
 \mathbf{Q} diagonal & alle $\|\mathbf{a}_m\| = 1$

Klassisches Faktorenanalysemodell

Datenkovarianz = Flachland + Berggrat

FA-Modell

$$\mathbb{X} = \boldsymbol{\mu} + \mathbb{E} + \mathbf{A}\mathbb{V} \quad \text{mit} \quad \begin{cases} \mathbb{E} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \\ \mathbb{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{E}) \\ \text{Cov}[\mathbb{V}, \mathbb{E}] = \mathbf{0} \end{cases}$$

$\mathbf{A} \in \mathbb{R}^{N \times M}$, \mathbf{R} Diagonalmatrix, \mathbf{E} Einheitsmatrix, Faktoren $\not\sim$ Störungen.

Lemma

Die Daten des klassischen FA-Modells sind gemäß $\mathcal{N}(\boldsymbol{\mu}, \mathbf{S})$ normalverteilt und es gilt:

$$\mathbf{S} = \mathbf{A}\mathbf{A}^\top + \mathbf{R}$$

Im Spezialfall $\mathbf{R} = \sigma^2 \cdot \mathbf{E}$ entsprechen die Ladungsvektoren $\mathbf{a}_1, \dots, \mathbf{a}_M$ den M ersten Hauptachsen der Datenkovarianzmatrix \mathbf{S} .

Beispiel

Dimensionen:
 $N = 3$ und $M = 1$

$$\begin{pmatrix} a_{11}^2 + r_1 & a_{11}a_{21} & a_{11}a_{31} \\ a_{11}a_{21} & a_{21}^2 + r_2 & a_{21}a_{31} \\ a_{11}a_{31} & a_{21}a_{31} & a_{31}^2 + r_3 \end{pmatrix}$$

$6 = 3 + 1 \cdot 3$ Parameter
 $6 = 3 + 2 + 1$ Gleichungen

Eigenschaften des FA-Modells

Lemma

1. **Skaleninvarianz:** Für den anisotrop skalierten Vektor $\mathbb{Y} = \mathbf{D}\mathbb{X}$ gilt

$$\text{Cov}[\mathbb{Y}] = \mathbf{DSD}^T = \mathbf{DAA}^T\mathbf{D}^T + \mathbf{DRD}^T = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T + \tilde{\mathbf{R}},$$

z.B. ist $\tilde{\mathbf{S}}$ die Korrelationsmatrix, wenn $\mathbf{D} = \text{diag}(\frac{1}{s_{11}}, \dots, \frac{1}{s_{NN}})$

2. **Kreuzkovarianz** zwischen Daten und Faktoren:

$$\text{Cov}[\mathbb{X}\mathbb{V}^T] = \mathcal{E}[(\mathbf{AV} + \mathbb{E}) \cdot \mathbb{V}^T] = \mathcal{E}[\mathbf{AVV}^T] + \mathcal{E}[\mathbb{EV}^T] = \mathbf{A}$$

3. **Freiheitsgrade:**

Die Matrizen \mathbf{A} und \mathbf{R} besitzen $N \cdot M + N$ frei wählbare Elemente. Diesen stehen $(N+1) \cdot N/2$ Bedingungsgleichungen gegenüber.

4. **Zerlegung der Datenvarianz:**

$$\text{Var}[\mathbb{X}_k] = S_{kk} = \sum_{m=1}^M a_{km}^2 + r_{kk} \quad (k = 1, \dots, N)$$

Hauptachsenverfahren

Algorithmen zur Faktorenanalyse (1)

1 EIGENDEKOMPOSITION

Berechne die M größten Eigenwerte und Eigenvektoren von \mathbf{S} :

$$\hat{\mathbf{A}}\hat{\mathbf{A}}^T \stackrel{\text{def}}{=} \sum_{m=1}^M \lambda_m \cdot \mathbf{u}_m \mathbf{u}_m^T$$

2 RESTVARIANZ

Die Störungsvarianzen ergeben sich aus der Differenzmatrix dieser Näherung:

$$\hat{r}_i \stackrel{\text{def}}{=} s_{ii} - \hat{h}_i^2, \quad \hat{h}_i^2 = \sum_{m=1}^M \hat{a}_{im}^2$$

(für alle $i = 1, \dots, N$)

Skalierung

Alle FA-Verfahren liefern erfahrungsgemäß die besten Resultate, wenn sie gleich auf die **Korrelationsmatrix** angewendet werden.

Die Hauptachsenzerlegung von \mathbf{S} ist deswegen ungünstig, weil hier unnötigerweise angestrebt wird, auch die Datenvarianzen s_{ii} zu approximieren.

Varianz = Kommunalität + Reststreuung

Definition

Sei $1 \leq n \leq N$; der zu den gemeinsamen Faktoren gehörende Anteil

$$h_n^2 \stackrel{\text{def}}{=} \sum_{m=1}^M a_{nm}^2 = \|\mathbf{a}_n\|^2, \quad \mathbf{A}^T = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N)$$

der Varianz heißt **Kommunalität von \mathbb{X}_n** .

Beispiel (Pathologische Faktorisierungen)

Die Varianzzerlegungen $S_{nn} = h_n^2 + r_{nn}$ können durchaus problematisch sein:

$$\begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix} \Rightarrow \mathbf{s}_{xx} = \begin{pmatrix} 1.255 \\ 0.717 \\ 0.558 \end{pmatrix} \cdot (1.255, 0.717, 0.558) + \begin{pmatrix} -0.575 & 0 & 0 \\ 0 & 0.486 & 0 \\ 0 & 0 & 0.689 \end{pmatrix}$$

Es ergibt sich zwar eine eindeutige Lösung; wegen der negativen Varianz $r_{11} = -0.575$ besitzt diese Konstellation jedoch keine wahrscheinlichkeitstheoretische Deutung!

Hauptfaktorenverfahren

Algorithmen zur Faktorenanalyse (2)

1 KOMMUNALITÄTEN

(Schätzung nach der Maximummethode)

$$\tilde{h}_i^2 = \max_{j \neq i} s_{ij} \quad (i = 1, \dots, N)$$

2 VARIANZAUSTRAG

(abgespeckte Korrelationsmatrix $\tilde{\mathbf{S}}$)

$$\tilde{s}_{ij} = \begin{cases} \tilde{h}_i^2 & i = j \\ s_{ij} & i \neq j \end{cases} \quad (i, j = 1, \dots, N)$$

3 EIGENDEKOMPOSITION

(die M führenden Eigenvektoren von $\tilde{\mathbf{S}}$)

$$\hat{\mathbf{A}}\hat{\mathbf{A}}^T \stackrel{\text{def}}{=} \sum_{m=1}^M \lambda_m \cdot \mathbf{u}_m \mathbf{u}_m^T$$

4 STÖRVARIANZEN

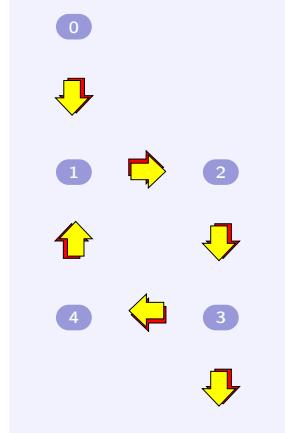
(aus den korrigierten Kommunalitäten)

$$\hat{r}_i \stackrel{\text{def}}{=} s_{ii} - \hat{h}_i^2, \quad \hat{h}_i^2 = \sum_{m=1}^M \hat{a}_{im}^2 \quad (i = 1, \dots, N)$$

Hauptfaktoreniteration

Algorithmen zur Faktorenanalyse (3)

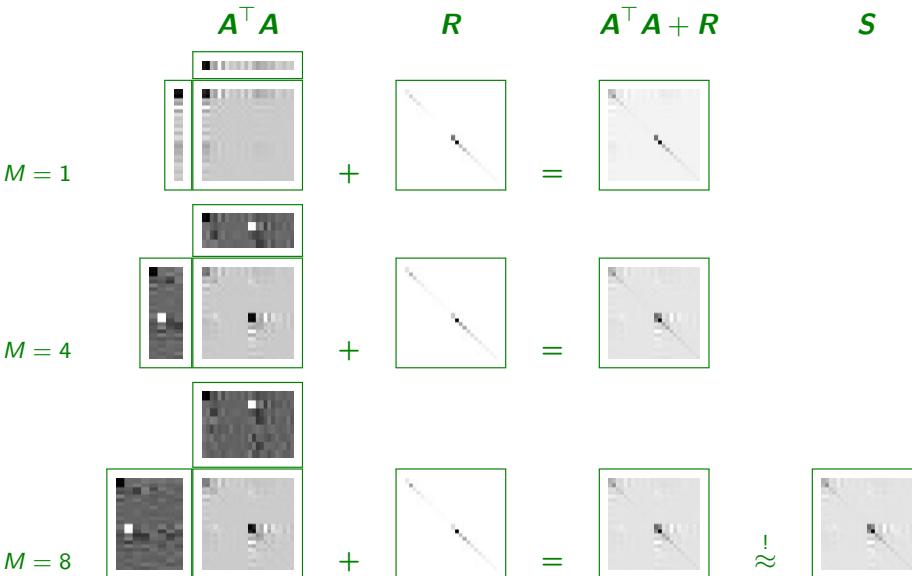
Ablaufschema



Zyklische Wiederholung

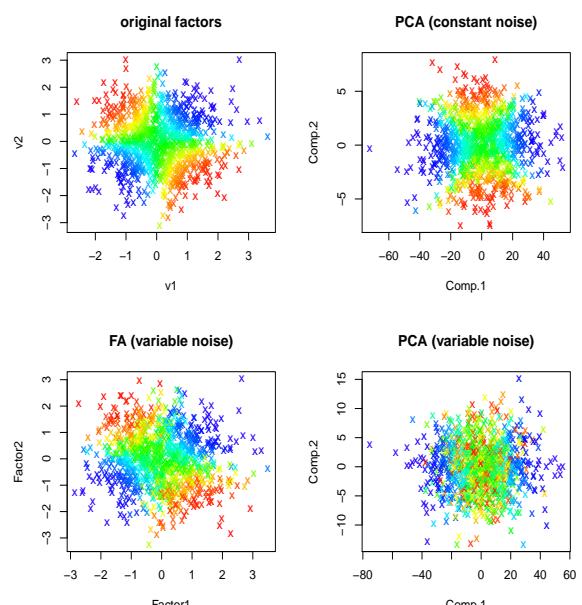
Ab dem zweiten Durchlauf dienen die \hat{h}_i^2 aus Schritt (4) als Vorgabe für die Werte \tilde{h}_i^2 aus Schritt (1).

Beispiel — Cepstrum-Sprachmerkmale (\mathbb{R}^{24})



Faktorenanalyse versus PCA

Zwei Faktoren v_i , mit Ladungsvektoren $a_i \in \mathbb{R}^5$ \Rightarrow nur FA packt heterogene Störung!



R-Code

```

layout(matrix(1:4,2,2),TRUE)
n <- 1000
v1 <- rnorm(n)
v2 <- rnorm(n)
VAt <-
  outer(v1, c(4.7,1.4,4.7)) +
  outer(v2, c(8,8,3,3,8))
cop <-
  rainbow(n, end=.7)[rank(v1*v2)]
plot(v1, v2,
  col=cop,
  main="original factors")
X <- VAt +
  rnorm(5*n, sd=1/2)
plot(princomp(X)$scores[,1:2],
  col=cop,
  main="PCA (constant noise)")
for(j in 1:5)
  X[,j] <- VAt[,j] +
    rnorm(n, sd=j^2/5)
plot(factanal(
  X, 2, scores="Bartlett"
 )$scores,
  col=cop,
  main="FA (variable noise)")
plot(princomp(X)$scores[,1:2],
  col=cop,
  main="PCA (variable noise)")
  
```

Hauptachsentransformation

Mehrdimensionale Skalierung

Nichtnegative Matrixfaktorisierung

Independent Component Analysis

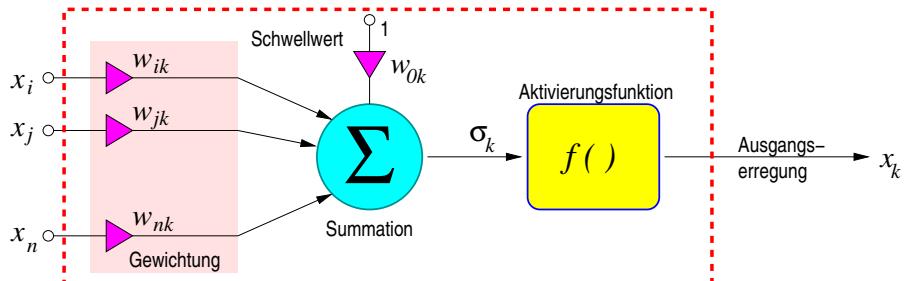
Faktorenanalyse

Merkmalkarte und autoassoziatives MLP

Explorative Datenanalyse & Grafik

Das Modellneuron

Grundbaustein des Paradigmas *massiv-paralleler Verarbeitung*



Parallel-Distributed Processing

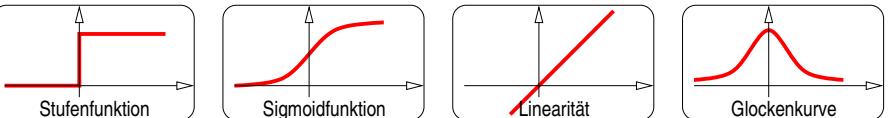
Ein **Künstliches Neuronales Netz** (KNN) ist ein Arrangement von **Modellneuronen** und ihren **Verbindungen**.

Dynamisches Verhalten

bestimmt durch

- Aktivierungsbegriff
- Kombination der Eingänge
- Aktivierungsfunktion
- Verschaltungstopologie

Neuronale Aktivierungsfunktionen



Lineare Aktivierung

$$x_k = a \cdot \sigma_k + b$$

lineare Netzwerke

Zielwertaktivierung

$$x_k = \exp \{ -C \cdot (\sigma_k - \theta_k)^2 \}$$

selbstorganisierende Karten

Sigmoidfunktion

$$x_k = \frac{1}{1 + \exp(-\sigma_k)}$$

Mehrschichtenperzeptron

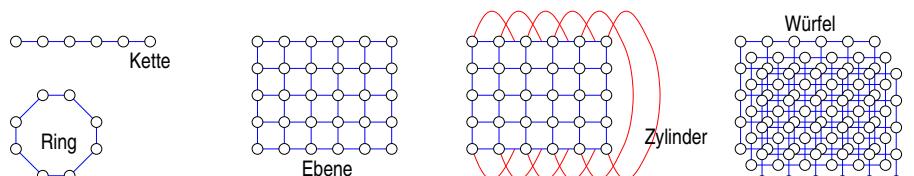
Sprungfunktion

$$x_k = \begin{cases} 1 & \sigma_k > \theta_k \\ 0 & \sigma_k \leq \theta_k \end{cases}$$

klassisches Perzeptron

Kohonen Mermalkarten

SOFM — 'self-organizing feature map' [?]



$$q^{\text{SOFM}} : \left\{ \begin{array}{c} \mathbb{R}^N \rightarrow \{\mathbf{o}_1, \dots, \mathbf{o}_L\} \subset \mathbb{R}^M \\ \mapsto \{\mathbf{w}_1, \dots, \mathbf{w}_L\} \subset \mathbb{R}^N \end{array} \right. , \quad M = 1, 2, 3$$

Definition

Ein Feld von L Knoten heißt **Selbstorganisierende Karte**, falls jeder Knoten ℓ durch einen **Referenzvektor** $\mathbf{w}_\ell \in \mathbb{R}^N$ sowie durch einen **Ortsvektor** $\mathbf{o}_\ell \in \mathbb{R}^M$ repräsentiert wird und die Ortsvektoren eine regelmäßige Punktmenge im \mathbb{R}^M bilden.

Kompetitives Lernen

Nachbarschaft in \mathbb{R}^N ↗ Nachbarschaft in \mathbb{R}^M

Neuronale Aktivität ('winner-takes-all')

$$u_0(\mathbf{x}) = \min_{\ell=1..L} u_\ell(\mathbf{x}) \quad \text{und für alle } \ell: \quad u_\ell(\mathbf{x}) = \|\mathbf{w}_\ell - \mathbf{x}\|^2$$

Lernen mit Nebenziele

Minimiere die Verzerrung $\sum_{\mathbf{x} \in \omega} u_0(\mathbf{x})$ unter Wahrung kleinstmöglicher Distanzabweichungen

$$\Delta_{k,\ell} = \|\mathbf{w}_k - \mathbf{w}_\ell\|^2 - \|\mathbf{o}_k - \mathbf{o}_\ell\|^2, \quad k, \ell \in \{1, \dots, L\}$$

zwischen Merkmal- und Ortsraum.

Gradientenabstiegsverfahren

(Algorithmus)

1 INITIALISIERUNG

Wähle zufällige Punkte $\mathbf{w}_1, \dots, \mathbf{w}_L \in \mathbb{R}^N$ aus.

2 ITERATIONSSCHRITT ($\forall t = 1, \dots, T$)

Berechne den Gewinnerknoten ℓ mit

$$\ell = \underset{1 \leq k \leq L}{\operatorname{argmin}} \|\mathbf{w}_k - \mathbf{x}_t\|^2$$

und aktualisiere alle (?) Prototypen:

$$\mathbf{w}_k \leftarrow \mathbf{w}_k + r_{k\ell} \cdot (\mathbf{x}_t - \mathbf{w}_k)$$

3 ABBRUCHKRITERIUM

Wiederhole Schritt 2 oder \rightsquigarrow ENDE.

(zum Diagramm)

Blasenfunktion

$$r_{ij} = \begin{cases} \eta & \|\mathbf{o}_i - \mathbf{o}_j\| < \rho \\ 0 & \text{sonst} \end{cases}$$

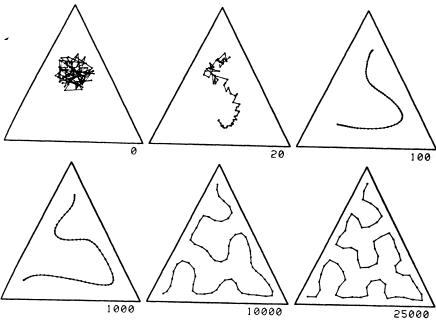
ρ Blasenradius, η
Lernrate

Gaußglocke

$$r_{ij} = \eta \cdot \exp \left(-\frac{\|\mathbf{o}_i - \mathbf{o}_j\|^2}{2\sigma^2} \right)$$

σ^2 Abklingrate, η
Lernrate

Beispielkartierung einfacher Figuren



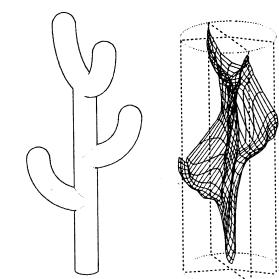
2D-Fläche \Rightarrow 1D-Kette

Referenzvektorpositionen nach
 $\begin{pmatrix} 0 & 20 & 100 \\ 1000 & 10000 & 25000 \end{pmatrix}$

Iterationsschritte

Quelldaten:

Uniform verteilte Punktwolke in Dreieckform



3D-Körper \Rightarrow 2D-Gitter

Quelldaten:

Punktwolke mit uniformer Verteilung auf kaktusförmigem Volumen

Zieldaten:

Referenzvektorengitter in Quellkoordinatensystem

Beispiel „Tiere unserer Heimat“

Kartierung der Spezies auf Grundlage nominaler Attribute

	dog	duck	goose	hen	horse	owl	zebra	wolf	tiger	cat	fox	dog	hen	owl	zebra	wolf	tiger	cat
is	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
small	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
medium	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
big	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
has	2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
	4 legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
	hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
	hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	mane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
likes	hunt	0	0	0	0	1	1	1	0	1	1	1	1	1	1	0	0	0
	run	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0
	fly	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Datenmatrix

16 Instanzen (Tierarten)

Attribute:

Größe (1x ordinal)

Extremitäten (2x nominal)

Zierden (4x nominal)

Hobbies (4x nominal)

Selbstorganisierende Karte

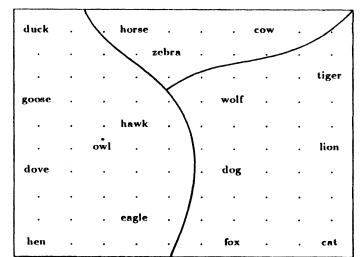
10 x 10 Gitterpunkte

2000 Objektpräsentationen in zufälliger

Folge

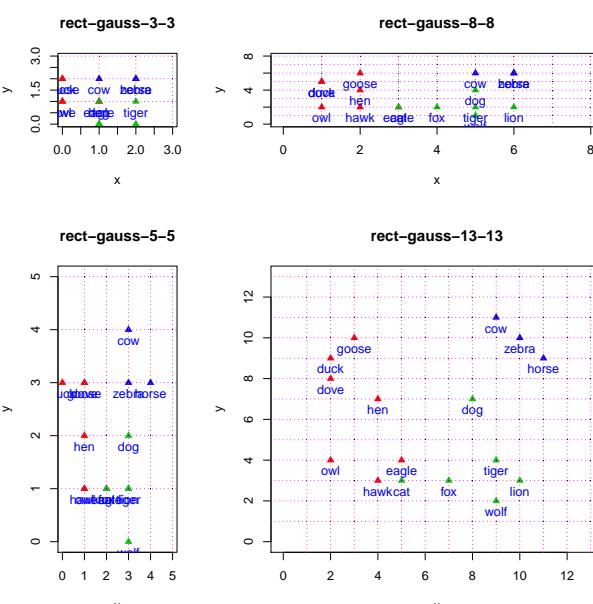
Gruppierung:

Vögel · Räuber · Huftiere



Planare (2D) Kohonenkarten

Rechteck/Hexagongitter · Gauß/Blasennachbarschaft



R-Code

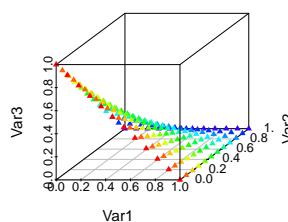
```

layout (matrix (1:4, 2),
height=1:2, width=1:2)
require (som)
for (k in c(3,5,8,13))
{
  y <- som (X,
  xdim=k, ydim=k,
  neigh="gauss",
  topol="rect")
  )$visual[1:2]
  plot (y,
  xlim=c(0,k),
  ylim=c(0,k),
  main=paste (
    "rect-gauss",
    k,
    sep="-"),
  col=rep (2:4,
  times=c(7,6,3)),
  pch=17)
  points (
  expand.grid (1:k, 1:k),
  pch=".")
  abline (v=1:k, h=1:k,
  lty=3, col="magenta")
  text (y, labels=rownames(X),
  pos=1, col="blue")
}
  
```

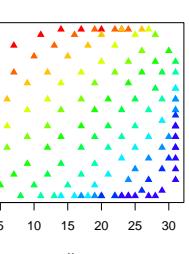
Planare (2D) Kohonenkarten

Synthetische Anwendungsbeispiele: Torsion & Spirale

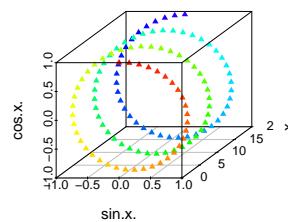
Torsionsdatensatz



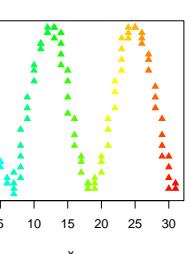
Kohonen's SOFM



Spiraldatensatz



Kohonen's SOFM



R-Code

```
layout (matrix (1:4,2,2,TRUE))
require (som)
k <- 32

n <- 12
x <- seq (0, 1, length=n)
xy <- as.matrix (
  expand.grid (x, x))
X <- data.frame (
  xy,
  Var3=(1-xy[,1])*(1-xy[,2]))
o.col <- rep (
  rainbow (n, end=0.7), each=n)
require (scatterplot3d)
scatterplot3d (
  X, pch=17, color=o.col,
  main="Torsionsdatensatz")
plot (som (X, k, k)$visual[1:2],
  col=o.col, pch=17,
  main="Kohonen's SOFM")

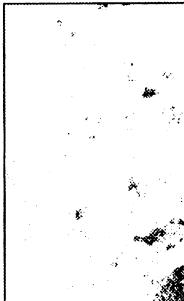
n <- 124
xmax <- 3*2*pi
x <- seq (0, xmax, length=n)
o.col <- rainbow (n, end=0.7)
X <- data.frame (
  sin(x), x, cos(x)
)
# ... und so weiter und so fort ...
```

WEBSOM

Kartierte Kurzfassungen (weltweiter) elektronischer Patentschriften

Patentdatensammlung

6 840 568 Kurzfassungen
(engl.)
733 179 \mapsto 43 222 Wortformen
1 002 240 Gitterpunkte

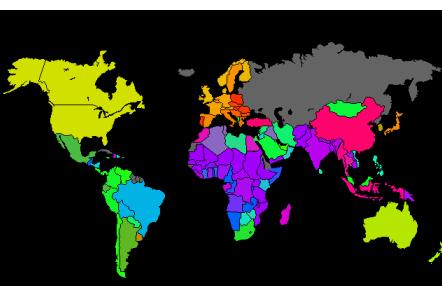
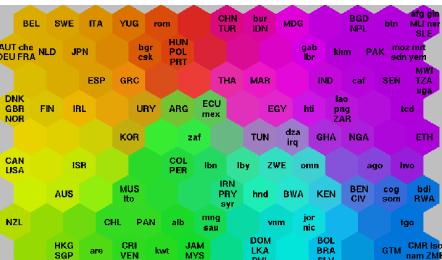


Datenmatrix

Worthäufigkeitsvektor $\text{IR}^{43222 \text{ ran}} \mapsto \text{IR}^{500}$
21 Sektionen (Patentklassifikationen)
Dreistufiges Bootstrapverfahren
(435 — 50 000 — alle Patente)

Kohonens „Welt-Hungerkarte“

Flächenrepräsentation einzelstaatlicher Wirtschafts- und Bevölkerungsdaten



Weltbankdaten

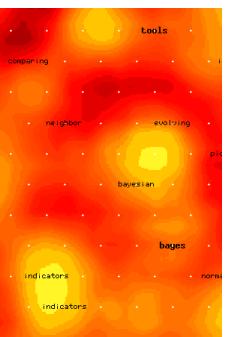
126 Staaten der Erde
39 Indikatoren (Armut)
48× weniger als 12 Attribute
(Kleinbuchstaben)

Selbstorganisierende Karte

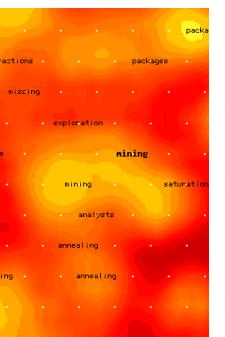
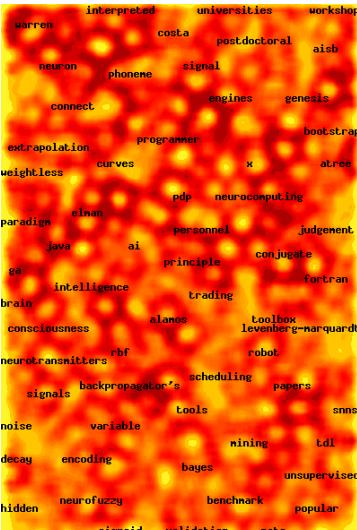
Hexagonales Gitter
450=18·25 Zellen
Häufungsgebiete \Rightarrow Farbtönen
(Rückfärbung in geographische Ebene)

WEBSOM

Kartierte Kurzfassungen von Machine-Learning Publikationen



„bayes“



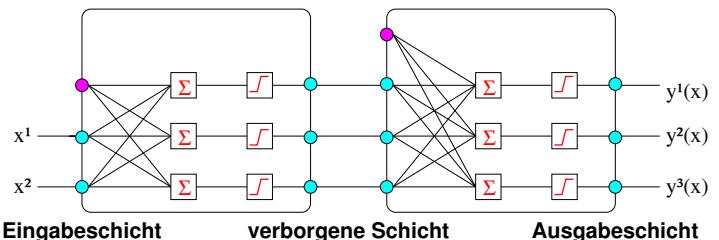
„mining“



Wurzelkarte

Das heteroassoziative Mehrschichtenperzeptron

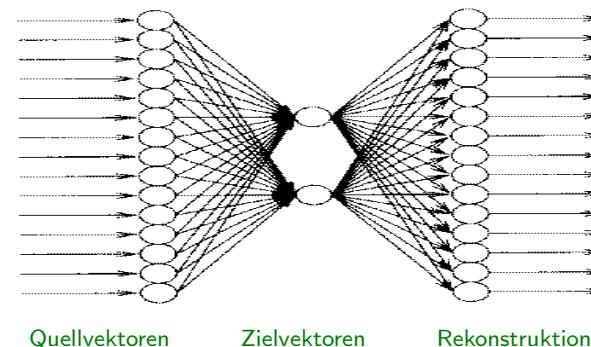
Beispiel: 2-3-3-Perzeptron (drei Schichten · zwei Stufen)



Konfiguration eines MLP als Klassifikator

1. N Eingabeneuronen und $(K - 1)$ Ausgabeneuronen $N = 2$ und $K = 4$ im Beispiel
2. Wieviele verborgene Schichten? eine verborgene Schicht im Beispiel
3. Wieviele Neuronen in verborgenen Schichten? 3 verborgene Neuronen im Beispiel
4. Maschinelles Lernen der Gewichtsmatrizen (3×3) und (4×3) im Beispiel

Das autoassoziative Mehrschichtenperzeptron

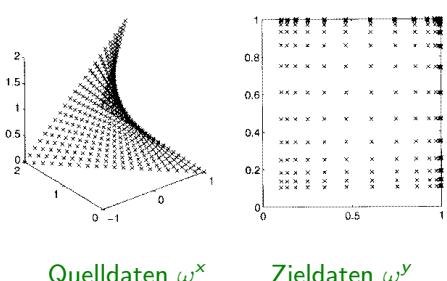


Konfiguration eines MLP zur nichtlinearen PCA

1. N Eingabeneuronen und N Ausgabeneuronen $N = 15$ im Beispiel
2. Genau eine verborgene Schicht!
3. $M \ll N$ Neuronen in der Mittelschicht $M = 2$ verborgene Neuronen im Beispiel
4. Maschinelles Lernen der Gewichtsmatrizen (16×2) und (3×15) im Beispiel

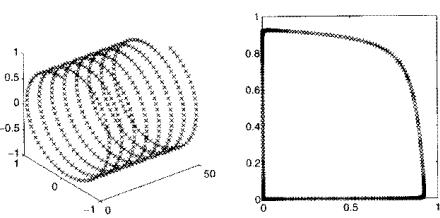
Synthetische Anwendungsbeispiele

Vom \mathbb{R}^3 in den \mathbb{R}^2 mit dem Auto-MLP · Th. Runkler (2000, S. 45)



Quelldaten ω^x

Zieldaten ω^y



3D-Torsionsfläche
 $\{(x_1, x_2, x_3)^\top \mid x_1 = (x_2 - 1)(x_3 - 1)\}$

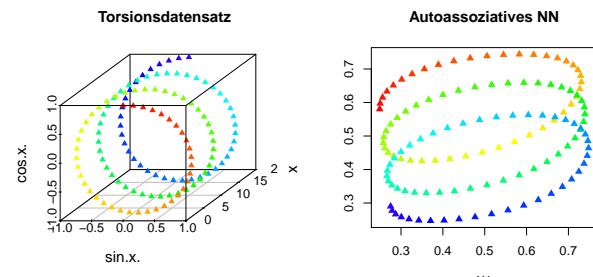
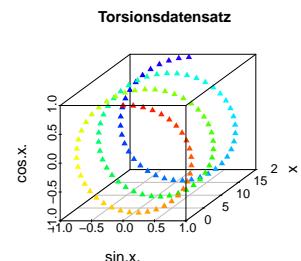
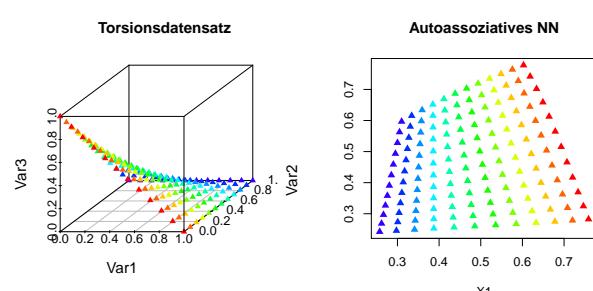
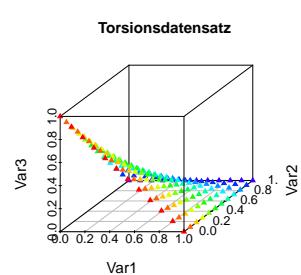
⊕ Gitterstruktur
 ⊖ Distanzartefakte

3D-Spirale
 $\{(x, \sin(x), \cos(x))^\top \mid x \in [0, 50]\}$

⊕ Welligkeit
 ⊕ Zusammenhang
 ⊖ Ringtopologie

Synthetische Anwendungsbeispiele

mit dem 'R'-Paket **neuralnet**



R-Code

```

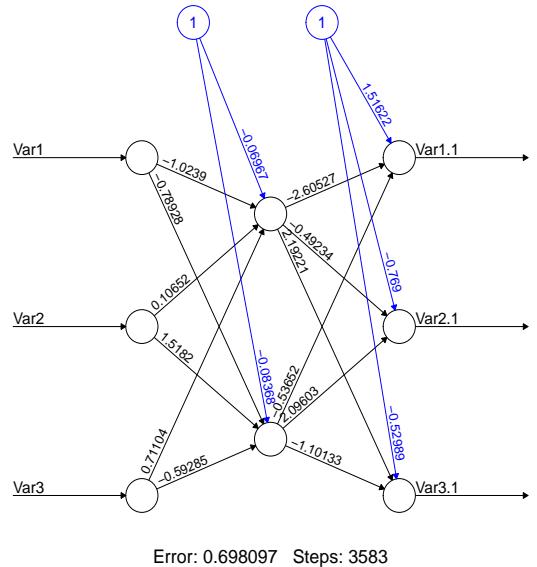
layout (matrix (1:4,2,2,TRUE))
n <- 12
x <- seq (0, 1, length=n)
xy <- as.matrix (
  expand.grid (x, x))
X <- data.frame (
  xy,
  Var3=(1-xy[,1])*(1-xy[,2]))
require (scatterplot3d)
scatterplot3d (
  X, pch=17, color=o.col,
  main="Torsionsdatensatz")
plot (predict (aann (X)),
  col=o.col, pch=17,
  main="Autoassoziatives NN")

n <- 124
xmax <- 3*2*pi
x <- seq (0, xmax, length=n)
X <- data.frame (
  sin(x), x, cos(x)
)
scatterplot3d (
  X, pch=17, color=o.col,
  main="Torsionsdatensatz")
plot (predict (aann (X)),
  col=o.col, pch=17,
  main="Autoassoziatives NN")

```

MLP (N - M - N) zur nichtlinearen Projektion

'R'-Paket **neuralnet**: Resilient Backpropagation & Weight Backtracking



R-Code

```
aann <- function (
  X, k=2, sd=1/3, plot=FALSE)
{
  X <- scale(X) * sd
  ivars <- colnames(X)
  ovars <- paste (
    ivars, "1", sep=".")
  form <- formula (paste (
    paste (ovars, collapse="+"),
    paste (ivars, collapse="+"),
    sep=""))
  require (neuralnet)
  o <- neuralnet (
    formula=form,
    data=data.frame (X,X),
    hidden=k)
  plot && plot (o, rep="best")
  structure (
    list (model=o), class="aann")
}
n <- 12
x <- seq (0, 1, length=n)
xy <- as.matrix (
  expand.grid (x, x))
X <- data.frame (xy,
  Var3=(1-xy[,1])*(1-xy[,2]))

aann (X, k=2, plot=TRUE)
```

Hauptachsentransformation

Mehrdimensionale Skalierung

Nichtnegative Matrixfaktorisierung

Independent Component Analysis

Faktorenanalyse

Merkmalkarte und autoassoziatives MLP

Explorative Datenanalyse & Grafik

VDE — Visuelle Datenerkundung

$$VDE = \sum_{i=1}^4 \left\{ \begin{array}{l} 1. \text{ Massendaten} \\ 2. \text{ visuelle Form} \\ 3. \text{ menschliche Wahrnehmung} \\ 4. \text{ M.-M.-Interaktion} \end{array} \right.$$

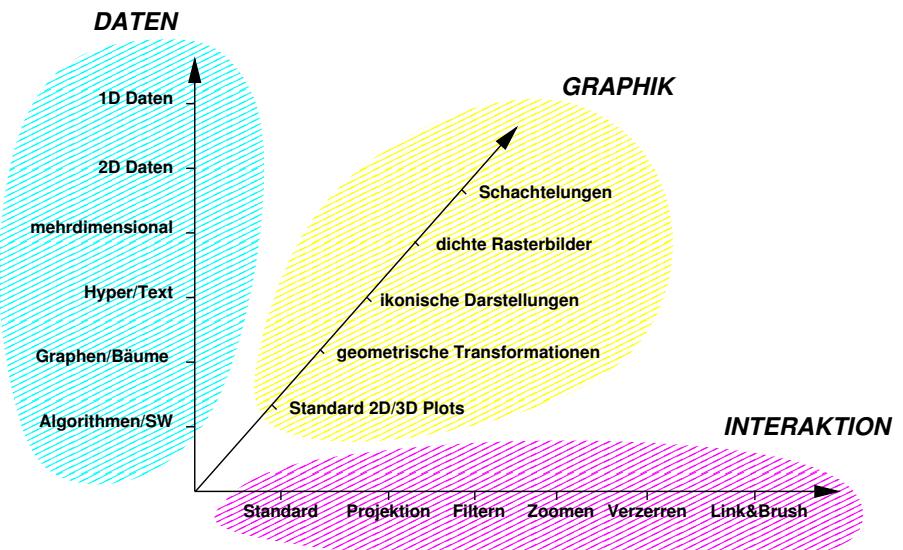
Ben Shneidermann¹⁹⁹⁶
„information seeking mantra“

1. Coarse Overview
2. Zoom & Filter
3. Details on demand

Taxonomie der VDE

- **Quelldatenstruktur**
Dimension? Kanäle?
Elementtyp? Relation?
- **Präsentationstechnik**
Koordinaten? Figuren?
Farbkarten? Schachteln?
- **Interaktionsform**
Perspektive? Sicht? Auswahl?
Verknüpfung?

Die drei Koordinaten der Datenvisualisierung



Graphische Darbietung

Standard 2D/3D

x/y-Plots · Balken · Kurven · Gebirge · Histogramme

Geometrische Transformationen

Scatterplot · Projektion · Prosektion · Parallelkoordinaten

Ikonen

Gesichter · Nadeln · Sterne · Männchen · Kreisel

Rasterbild

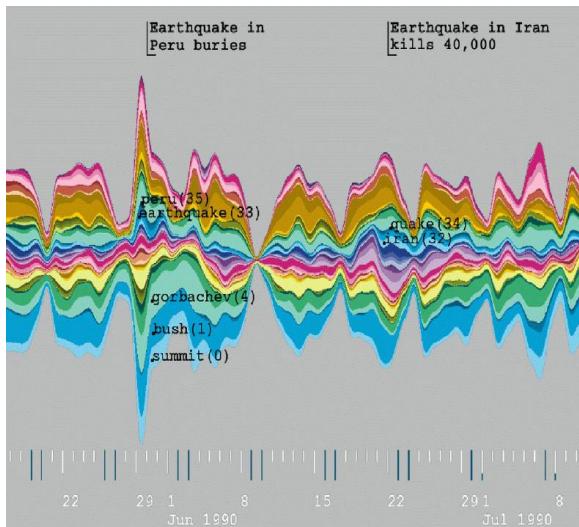
Attribute = Grau/Farbwert · Objektgeometrie (kartesisch/radial)

Schachtelung

Achsen ordnen & hierarchisch verschachteln

Mehrkanalige 1D-Daten

Thematische Veränderungen in Textdokumenten



ThemeRiver
Zeitverlauf der Topikproportionen

Datenquelle

Nachrichtenmeldungen
Associate Press
(Juni–Juli 1990)

Dokumentattribute

Worthäufigkeiten
PCA oder MDS

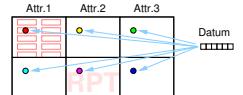
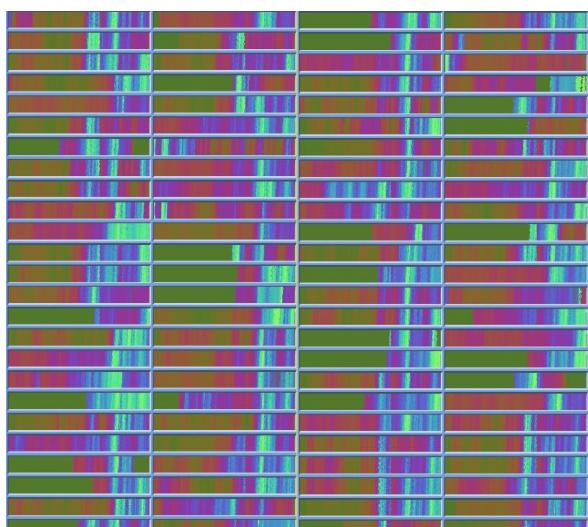
Themenschwerpunkte

Erdbeben Peru
Gipfel Bush/Gorbi

... ...

Dichte 1D-Daten

FAZ Frankfurter Aktienindex · 100 Verläufe · Jan'74–Apr'95



RPT
„recursive pattern technique“

Schachtelung

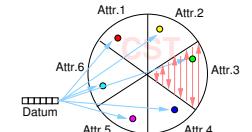
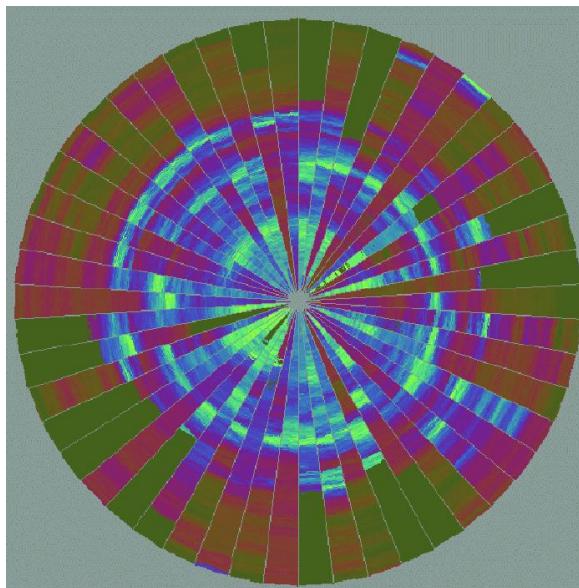
100 = 25 × 4

Regionen

Verläufe
spaltenweise
farbkodiert

Dichte 1D-Daten

FAZ Frankfurter Aktienindex · 50 Verläufe · Jan'74–Apr'95



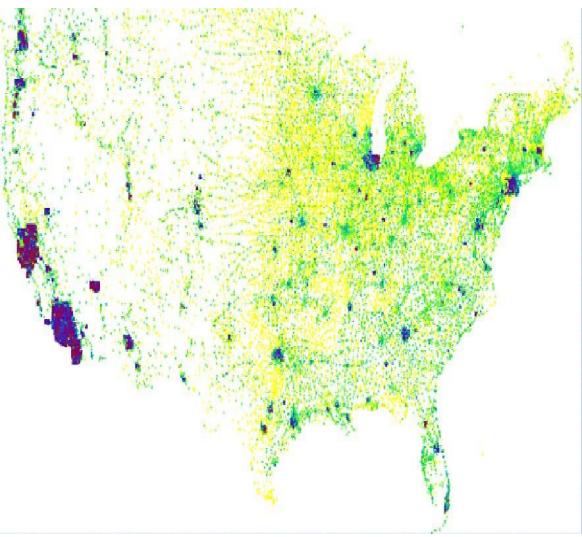
CST
„circle segment technique“

Schachtelung

50 Kreissegmente
Verläufe
spaltenweise
farbkodiert

Einkanalige 2D-Daten

Geographische Volumendaten für Telefonanrufe



GridFit-Grafik

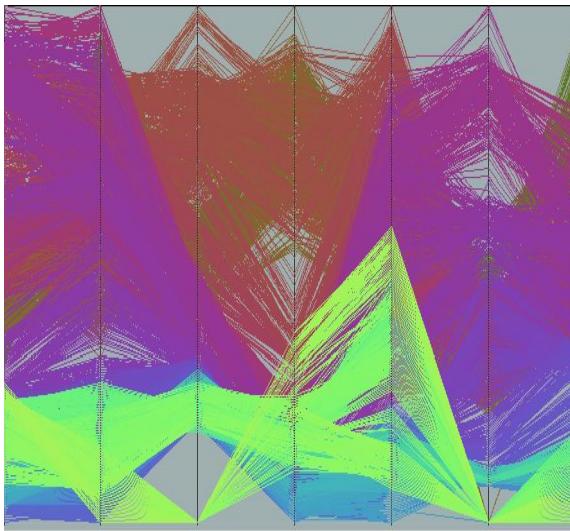
vermeidet
Überlappung der
Wertrepräsentanten
in Regionen hoher
Dichte

Datenquelle

Telefonaufkommen
US-amerikanischer
Bezirksstädte

Mehrdimensionale Daten

Tabellen relationaler DB · Merkmale von Sensordaten



PCT

,„parallel coordinate
technique“

Wäscheleinengrafik

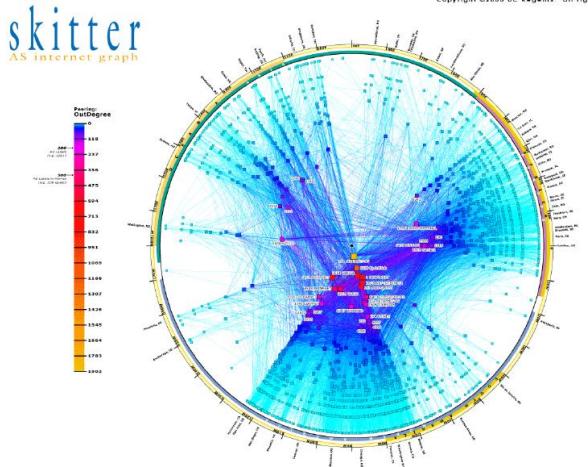
Polygonzug durch
parallel
aufgerichtete
Koordinatenachsen

Attributwerte = Schnitt-
höhe

Ähnlichkeitsfarbgebung

Graphenförmige Daten

Netzwerke · Bäume · Telefon/Verkehrsnetze · Taxonomien



Datenquelle

Globale Vernet-
zungsstruktur des
Internets
(Okt 2000)

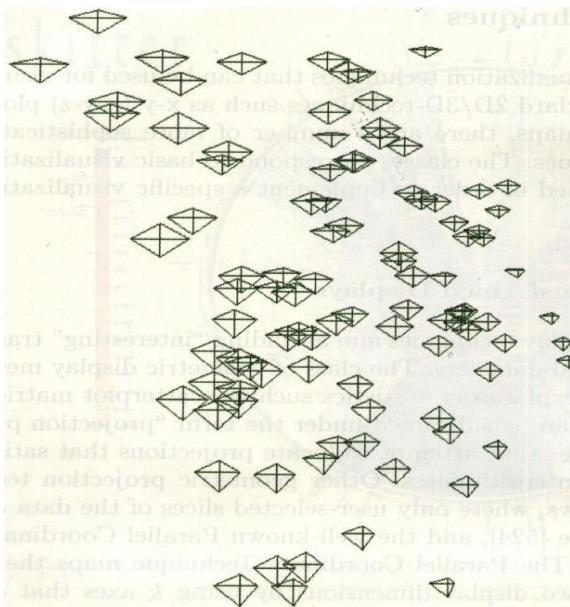
Skitter

geografische Länge
in Polarkoordinaten

Zentrierung und
Farbgebung nach
Verbindungsanzahl
(unten USA, rechts
EU, links Fernost)

Mehrdimensionale Daten

Tabellen relationaler DB · Merkmale von Sensordaten



SGD

,„star glyph display“

Datenquelle

Der legendäre
Irisdatensatz

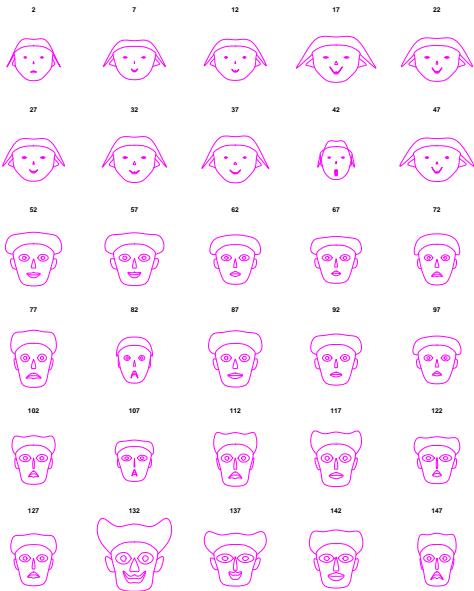
Ikonendarstellung

u_1 und u_2 :
Diamantposition im
 \mathbb{R}^2

u_3 bis u_6 :
Kantenlängen der
Windrose

Mehrdimensionale Daten

Physiognomische Darstellung numerischer Attribute ($N \leq 15$)



Parameter

1-height of face, 2-width of face, 3-shape of face,
4-height of mouth,
5-width of mouth, 6-curve of smile,
7-height of eyes,
8-width of eyes, 9-height of hair,
10-width of hair,
11-styling of hair,
12-height of nose,
13-width of nose,
14-width of ears,
15-height of ears.

Datenquelle

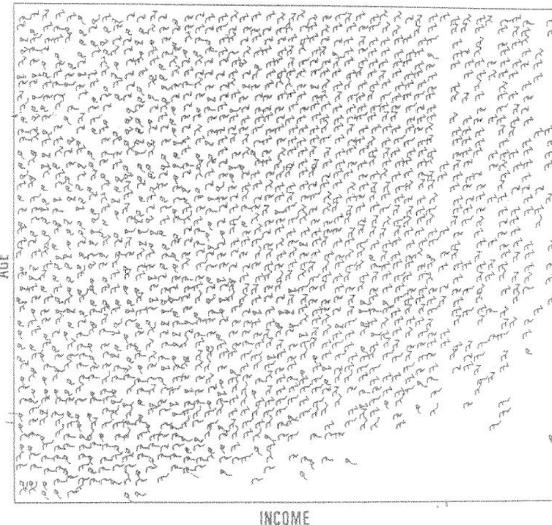
Der Irisdatensatz

Che~~R~~noff++

```
library (TeachingDemos)
#Qdev      width=12,height=14
par (col="magenta")
N=150; M=5; K=6;
idx <- (1:(K*M)-.5)*N/K/M
faces (iris[idx,-5], nr=K, nc=M)
```

Mehrdimensionale Daten

Kombination von Scatterplot & Ikonentechnik



Hustlergrafik
Strichmännchen/mädchen
(Univ. Lowell, MA)

Datenquelle
Mikrozensus USA

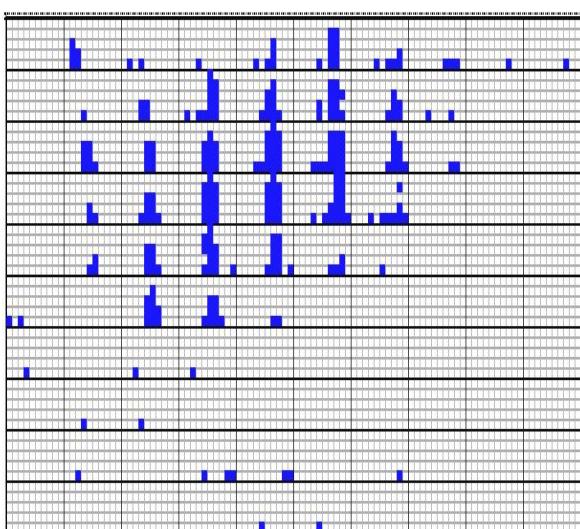
Repräsentation
 $x \hat{=} \text{Einkommen}$
 $y \hat{=} \text{Alter}$



Strichpersönchen:
Geschlecht,
Abstammung,
Schulbildung,
Hautfarbe, ...

Mehrdimensionale Daten

Hierarchisierung der Datenattribute



Aufschlußbohrdaten
Länge/Breite/Grad/Tiefe

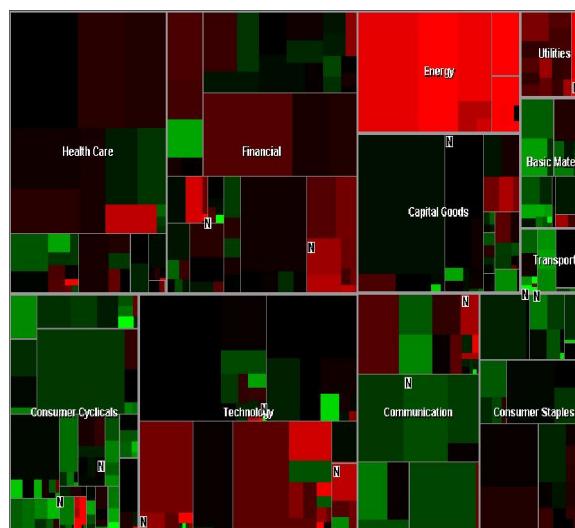
Stacking

Achsenpaare
schachtern

2D-Gitter \subset
2D-Gitter etc.

Mehrdimensionale Daten

Hierarchisierung der Datenobjekte



TreeMap

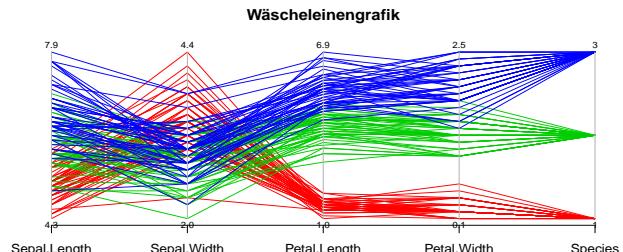
Sukzessive
Zerlegung des
Datensatzes
abwechselnd in
x-Richtung und
y-Richtung

Finanzdaten
nach
Marktsegmenten

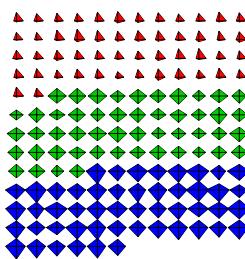
Größe $\hat{=}$
Kapitalisierung
Farbe $\hat{=}$ Aktienwert
absinkend \cdot ansteigend

Mehrdimensionale Daten in 'R'

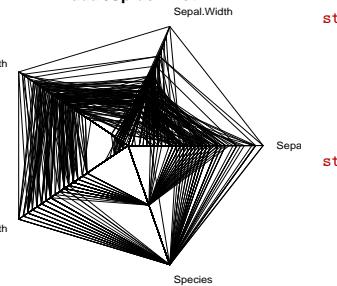
Parallele, sternförmige & konzentrische Koordinatenachsen



Star Plot



Radar/Spider Plot



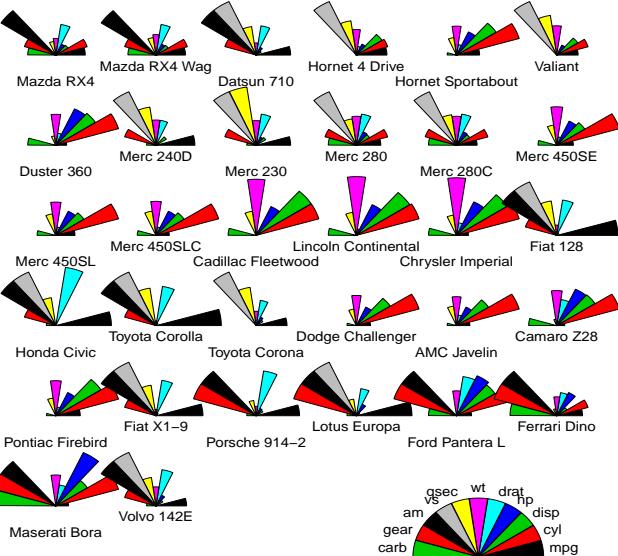
R-Code

```
layout (matrix (
  c(1,1,2,3), 2, 2, TRUE
))
f.col <- 1 +
  unclass (iris$Species)
Iris <- as.data.frame (
  lapply (
    iris, as.numeric))
require (MASS)

parcoord (
  Iris, col=f.col,
  var.label=TRUE,
  main="Wäschleinengrafik")
stars (
  scale (iris[-5],
  center=FALSE),
  col.stars=f.col,
  scale=FALSE,
  len=0.8,
  main="Star Plot")
stars (
  iris,
  locations=c(0,0),
  key.loc=c(0,0),
  main="Radar/Spider Plot")
```

Segmentale Sterngrafik

Attributwert $\hat{=}$ Kreissegmentfläche · Radius (Stern) vs. Winkel (Torte)



R-Code

```
stars (
  mtcars,
  key.loc=c(12,1),
  draw.segments=TRUE,
  len=1.5,
  full=FALSE,
  flip.labels=TRUE )
```

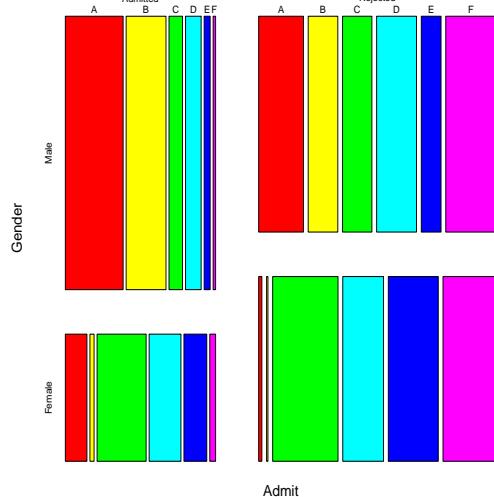
Motorcars

11 Attribute:

- X₁: Miles/gallon
- X₂: Cylinders
- X₃: Displacement
- X₄: Horsepower
- X₅: Rear axle ratio
- X₆: Weight
- X₇: 1/4 mile time
- X₈: V/S
- X₉: Transmission
- X₁₀: Gears (forward)
- X₁₁: Carburetors

Mosaikgrafik

Mehrdimensionale Histogrammdarstellung · Flächen $\hat{=}$ Anteile



R-Code

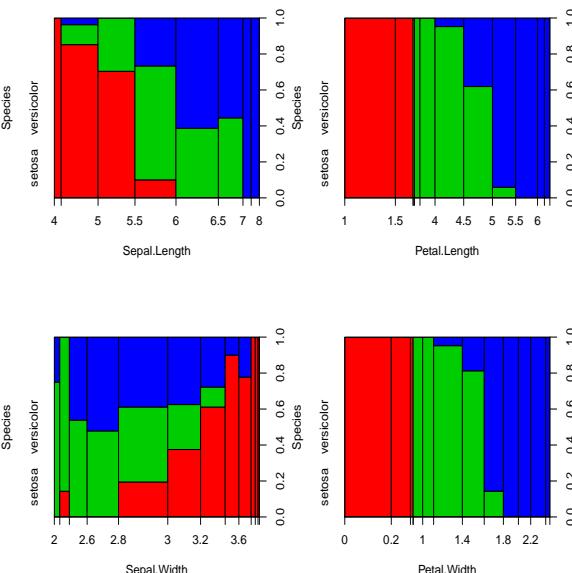
```
layout (matrix (1))
mosaicplot (
  UCBAmissions,
  sort=c(1,2,3),
  color=rainbow(6),
  main=NULL
)
```

UCB Admissions

- X₁: Zulassung (y/n)
- X₂: Geschlecht (m/f)
- X₃: Fakultät (A-F)

Spinogramm („Bandscheibengrafik“)

Histogrammdarstellung einzelattributbezogener Klassendiskriminanten



R-Code

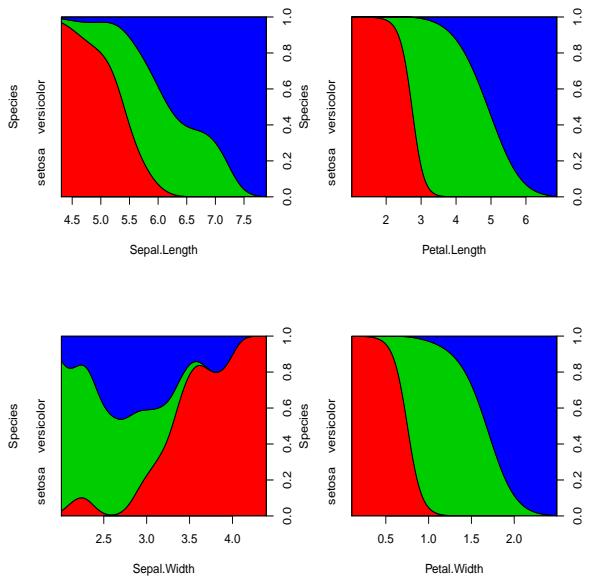
```
layout (matrix (1:4, 2))
for (att in names(iris)[-5])
  spineplot (
    formula (paste (
      "Species", "~", att
    )),
    iris,
    col=2:4
  )
```

P(y|x)-Balken

y nominal
x numerisch/diskret
Mosaikdarstellung der Vorhersagewahrscheinlichkeiten

Conditional Density Plot

Einzelattributbezogene empirische Klassendiskriminanten



R-Code

```
layout (matrix (1:4, 2))
for (att in names(iris)[-5])
  cdplot (
    formula (paste (
      "Species", "~", att
    )),
    iris,
    col=2:4
  )
```

$P(y|x)$ -Kurven

y nominal
x numerisch
kumulative
Darstellung der
Vorhersagewahr-
scheinlichkeiten

Mehrdimensionale Daten

Tabellen relationaler Datenbanken

	Avg	Career Avg	Team	Salary 87
Larry Herndon	0.24734983	0.27982076	Det.	225
Jesse Barfield	0.2989249	0.27950919	Tor.	231.5
Jeffrey Leonar	0.27859238	0.27950458	S.F.	900
Donne Hill	0.28318584	0.2795564	Dak.	275
Billy Sample	0.285	0.2718601	Atl.	na
Howard Johnson	0.24545455	0.25930268	N.Y.	297.5
Andres Thomas	0.28507174	0.951904	Atl.	75
Billy Hatcher	0.25755656	0.25911507	Hou.	110
Omar Moreno	0.2339433	0.2518029	Atl.	na
Darnell Coles	0.27955209	0.9515375	Det.	105

Datenquelle

Baseballspieler-DB
Tore \emptyset/Σ , Team,
Gehalt, ...

Table Lens

Ausschnittdarstellung
Vergrößerung
höhere Auflösung

Zooming

Balken \leadsto Text
Zeichenspots \leadsto
Text

Interaktive Graphik

1) Navigation · 2) Fokussierung · 3) Selektion

Dynamische Projektion

Perspektivenwechsel („Film“) · Grand-Tour-System

Filter

Achsen (*browse*) auswählen · Eigenschaften (*query*) benennen

Zoom

Ausschnitt größer · Ausschnitt in höherer Auflösung

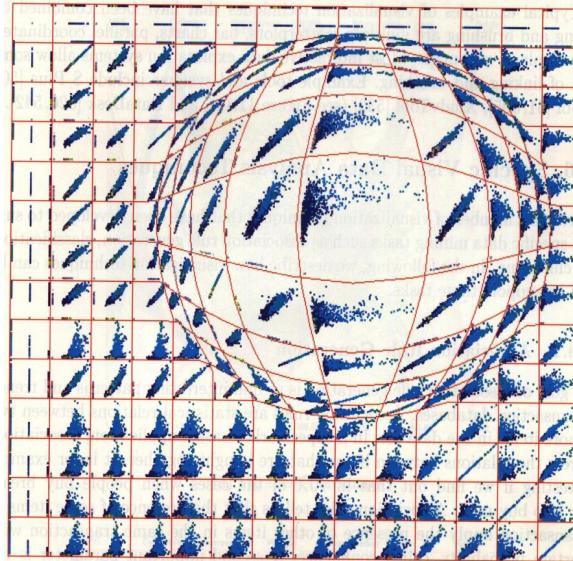
Verzerrung

konkurrierende Auflösungen:
Fischauge, Bifokalansicht, „perspective wall“

Brush & Link

dynamisch Verbindungen über Farben oder Kanten herstellen

Mehrdimensionale Daten



Datenquelle

Statlog-Datensatz
„vehicle“
 $x \in \mathbb{R}^{13} \subseteq \mathbb{R}^{19}$

Leselupeneffekt

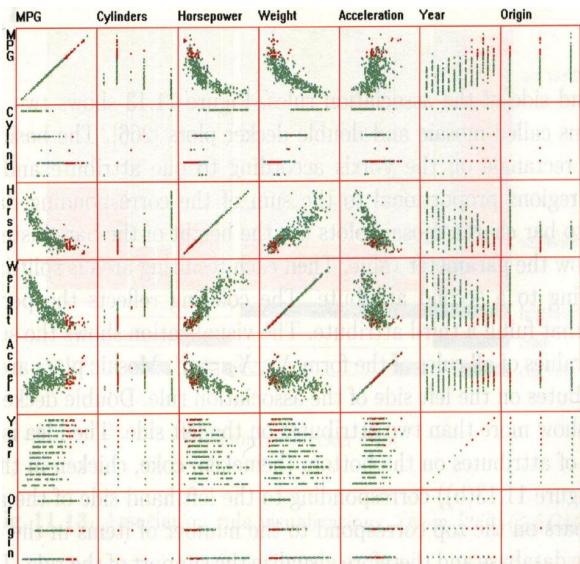
„fisheye lens“

Scatterplot

Topologische
Verzerrung der
Grafikebene

Verbundmarkierung

Verknüpfung zweier Darstellungstechniken durch Objektidentifikation

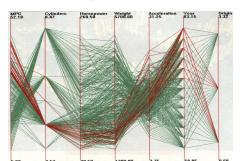


Brush & link

Interaktive
Farbmarkierung von
ROIs
Automatischer
Transfer auf
Zweitgrafik

Verknüpfungspartner

Wäscheleinengrafik
Scatterplot



Bewegtbildvisualisierung

library(tourr)

Pfad — Folge orthogonaler Projektionsmatrizen

- Zufallssequenz: alle PMs
- Zufallssequenz: koordinatenparallele PMs
- Alternierend: Start nahe Zufalls-PM
- Konvergierend: Interessantheitskriterium
- Drehbuch vorgegeben

grand_tour

little_tour

local_tour

guided_tour

planned/frozen_tour

display_dist

display_xy/image

display_depth/stereo

display_pcp/faces/stars

Schnappschuß — Visualisierungsmodus

- x-Achse Dichteplot
- (x, y)-Ebene Scatterplot/Pseudofarbraster
- (x, y, z)-Raum schattiert/Stereo
- \mathbb{R}^N -Raum Wäscheleinen/Glyphen

Zusammenfassung (3)

1. Data Mining ist ein **interaktiver Prozeß**; die Kommunikation zwischen Mensch und Maschine beruht auf **Datenvisualisierung**.
2. Die **graphische Darstellung** von Daten ist nur in **niederdimensionalen Räumen** praktikabel.
3. Die **PCA** projiziert die Datenvektoren auf die **Hauptachsen** — die führenden Eigenvektoren — ihrer Kovarianzmatrix.
4. Die **mehrdimensionale Skalierung** reduziert die Dimension unter größtmöglicher **Abstandstreue**.
5. Das **autoassoziative MLP** minimiert den Reproduktionsfehler; Kohonens **Merkmarkarten** pressen die Daten in eine **gitterförmige** Zielkonfiguration.
6. Die Zerlegung der Datenmatrix in **nichtnegative Faktoren** sichert eine **dünne** und **verteilte** Repräsentation.
7. Die Zerlegung in statistisch **unabhängige Komponenten** ermöglicht die **blinde Trennung** additiv überlagerter Signalquellen.
8. Das **Faktorenanalysemodell** strebt wie die PCA eine **niederrangige Approximation** der Datenkovarianz an, erlaubt jedoch attributweise **unterschiedliche Störvarianzen**.

MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2020

Prof. E.G. Schukat-Talamazzini

Stand: 7. Dezember 2020

Teil IV

Vorhersage und Kategorisierung

Vorhersage und statistische Abhangigkeit

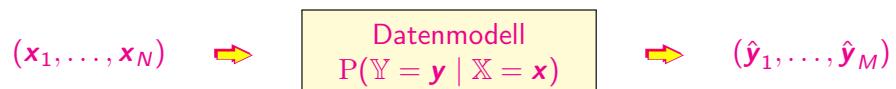
Charakterisierung der statistischen Unabhangigkeit

Zwei Variablenmengen $\mathbb{X} = (\mathbb{X}_1, \dots, \mathbb{X}_N)$ und $\mathbb{Y} = (\mathbb{Y}_1, \dots, \mathbb{Y}_M)$ heien statistisch unabhangig voneinander gdw. gilt:

$$(\forall x)(\forall y) \quad P(\mathbb{X} = x, \mathbb{Y} = y) = P(\mathbb{X} = x) \cdot P(\mathbb{Y} = y)$$

Fur Tupel x mit $P(x) \neq 0$ ist das aquivalent zu:

$$P(\mathbb{Y} = y \mid \mathbb{X} = x) = P(\mathbb{Y} = y)$$



Fakt

Im Fall statistischer Abhangigkeit besteht eine Chance, die Werte der **endogenen** Variablen \mathbb{Y}_m aus den Werten der **exogenen** Variablen \mathbb{X}_n zu „erraten“.

Statistische Pradiktion von Einzelvariablen

Quellvariable $(\mathbb{X}_1, \dots, \mathbb{X}_N) \Rightarrow$ Zielvariable $\mathbb{Y}_1 =: \mathbb{Y}$

Maschinelles Lernen eines Vorhersagemodells

Entscheidungsfunktion: $f : \mathcal{X}_1 \times \dots \times \mathcal{X}_N \rightarrow \mathcal{Y}$

Kostenfunktion („loss“): $\mathcal{L}(x, y, \hat{y})$ mit $\hat{y} = f(x)$

Risiko (zu minimieren): $\mathfrak{R}(f) := \mathcal{E}_{P(x,y)}[\mathcal{L}(\mathbb{X}, \mathbb{Y}, f(\mathbb{Y}))]$

\mathbb{Y} nominal

$$\mathcal{L}(x, y, \hat{y}) = c_{y\hat{y}}$$

Kostenmatrix C mit
 $c_{\kappa\kappa} \leq c_{\kappa\lambda}$

\mathbb{Y} ordinal

$$\mathcal{L}(x, y, \hat{y}) = c_{y\hat{y}}$$

Diskrepanzmatrix C mit
 $c_{k\ell} \leq c_{k'\ell'}$ fur
 $k' \leq k \leq \ell \leq \ell'$

\mathbb{Y} kardinal

$$\mathcal{L}(x, y, \hat{y}) = d(y, \hat{y})$$

metrische Distanzmae
 $d(y, \hat{y}) = |y - \hat{y}|^p$, $p \geq 0$

Spezialfall

(Fehlerrate)

$$c_{\kappa\lambda} = \begin{cases} 0 & \kappa = \lambda \\ 1 & \kappa \neq \lambda \end{cases}$$

Spezialfall

(Linearskala)

$$c_{k\ell} = |z_k - z_\ell|$$

Spezialfall

(Quadratmittel)

$$d(y, \hat{y}) = (y - \hat{y})^2$$

Optimale Prädiktion in den Spezialkonfigurationen

$$\mathfrak{R}(f) = \mathcal{E}[\mathcal{L}(\mathbb{X}, \mathbb{Y}, f(\mathbb{Y}))] = \int \sum_y P(x, y) \cdot c_{y, f(x)} dx$$

Klassifikation (Bayesregel)

\mathbb{Y} ist **nominal**

$$\hat{y}(\mathbf{x}) = \operatorname{argmax}_{\kappa \in \Omega_y} P(\mathbb{Y} = \kappa | \mathbf{x})$$

Ordinal Klassifikation

\mathbb{Y} ist **ordinal**

$$\hat{y}(\mathbf{x}) = \operatorname{median}_{\ell \in \Omega_y} P(\mathbb{Y} = \ell | \mathbf{x})$$

Quadratmittel-Regression

\mathbb{Y} ist **kardinal**

$$\hat{y}(\mathbf{x}) = \mathcal{E}_{\mathbb{Y}|\mathbf{x}}[\mathbb{Y}] = \int_{\mathbb{R}} P(y|\mathbf{x}) \cdot y dy$$

Modus

Median

Mean

Numerisch

- NV-Klassifikator
- Polynomklassifikator
- Multilayer-Perzeptron
- Supportvektormaschine

Diskret

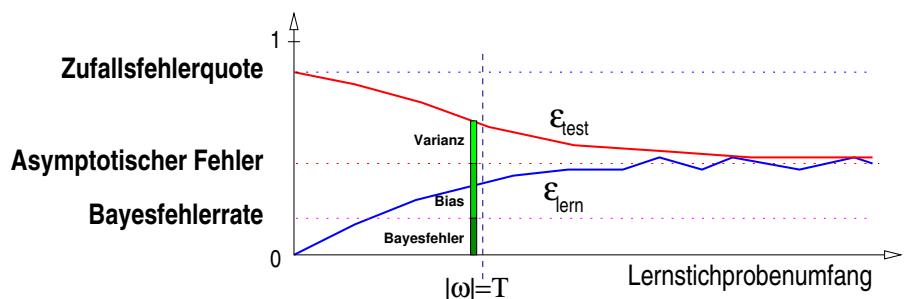
- Versionenraumverfahren
- Kanonische+naive Bayesregel
- Markovnetze

Numerisch & diskret

- Entscheidungsbäume
- Loglinearmodelle
- Bayesnetze
(Konversion)

Fehlerrate, Überanpassung & Unteranpassung

Was wir schon in der Vorlesung „Mustererkennung“ über das Lernen gelernt haben



Lernstichprobe des Klassifikationsverfahrens

Fehlerrate auf den Lerndaten

Fehlerrate auf den Testdaten (\approx Fehlerwahrscheinlichkeit)

$$\omega = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$$

$$\epsilon_{\text{lern}}$$

$$\epsilon_{\text{test}}$$

Bayesfehler — weniger geht nicht
Zufallsfehler — mehr muss nicht
Grenzfehler — Daten! Daten!!

Bias
Datenmodell

Varianz
Lernprobe

Versionenräume

Naive Bayesregel

Multivariate lineare Regression

Logistische Regression

Ordinale Regression und Präferenzmodelle

Statistische Entscheidungsbäume

Zusammenfassung

Aussagenlogisches Lernen

Begriffe lernen · Klassifikation · Gruppierung

Segel-Szenarium

\mathcal{X}_1 sky	\mathcal{X}_2 air	\mathcal{X}_3 humidity	\mathcal{X}_4 wind	\mathcal{X}_5 water	\mathcal{X}_6 forecast
{sunny}	{warm}	{normal}	{strong}	{warm}	{same}
rainy	cool	high	weak	cool	change
cloudy		low			

Single Representation Trick

z.B. Hypothesen als unvollständige Attributwertspezifikationen

- **Objekte** $\hat{=}$ Attributbelegungen

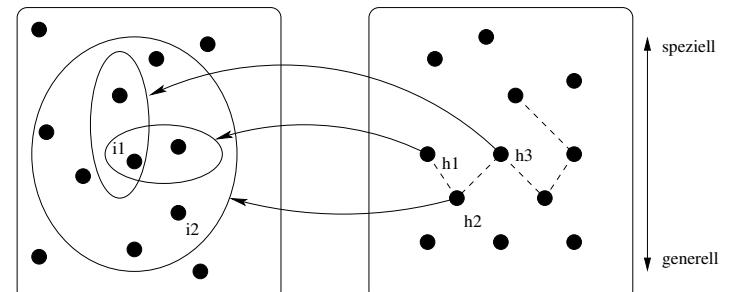
$$(sunny, \text{warm}, \text{normal}, \text{strong}, \text{warm}, \text{same}) \in \Omega$$

- **Hypothesen** $\hat{=}$ partielle Attributbelegungen

$$(sunny, ?, ?, \text{strong}, ?, ?) \in \mathcal{H}$$

Hypothesen und Objektmengen

$$\Omega(h) \hat{=} \{x \in \Omega \mid h \models x\}$$



Segel-Szenarium

$$\begin{aligned} i_1 &: (\text{sunny}, \text{warm}, \text{high}, \text{strong}, \text{cool}, \text{same}) \\ i_2 &: (\text{sunny}, \text{warm}, \text{high}, \text{light}, \text{warm}, \text{same}) \end{aligned}$$

$$\begin{aligned} h_1 &: (\text{sunny}, ?, ?, \text{strong}, ?, ?, ?) \\ h_2 &: (\text{sunny}, ?, ?, ?, ?, ?, ?) \\ h_3 &: (\text{sunny}, ?, ?, ?, ?, \text{cool}, ?) \end{aligned}$$

$$\begin{aligned} h_2 &\supseteq h_1, h_3 \\ &\text{bzw.} \\ h_1, h_3 &\Rightarrow h_2 \end{aligned}$$

Hypothesenraum

Konjunktionen positiver Literale (KPL)

Definition

Es sei $\Omega = \mathcal{X}_1 \times \dots \times \mathcal{X}_N$ ein Objektraum. Dann heißen die Elemente aus

$$\mathcal{H} = (\mathcal{X}_1 \cup \{?\}) \times \dots \times (\mathcal{X}_N \cup \{?\})$$

KPL-Hypothesen über Ω . Die Menge \mathcal{H} heißt **KPL-Hypothesenraum** über Ω .

Ein Beispielobjekt $x \in \Omega$ genügt der Hypothese $h \in \mathcal{H}$ (x erfüllt h bzw. $h \models x$) genau dann, wenn gilt:

$$\forall i = 1, \dots, N : (h_i = ?) \vee (h_i = x_i)$$

Bemerkung

Vollständige KPL $\hat{=}$ Objekte

Leere KPL $\hat{=}$ Konzept $\mathcal{C} = \Omega$

Definiere $h_\emptyset \hat{=}$ Leerkonzept

$\mathcal{C} = \emptyset$

Segel-Szenarium

Objektraum: $|\Omega| = 144$

KPL-Hypothesenraum: $|\mathcal{H}| = 1296$

Konzeptraum:

$$|\mathfrak{P}\Omega| = 2^{144} \approx 1000^{14.4} \approx 10^{43}$$

Der Verband aller KPL-Hypothesen

Definition

Für jede Hypothese $h \in \mathcal{H}$ sei $\Omega(h) \stackrel{\text{def}}{=} \{x \in \Omega \mid h \models x\}$ (Extension) definiert. Die Menge \mathcal{H} erbt von $\mathfrak{P}\Omega$ die **Inklusionsrelation** (h ist „allgemeiner“ oder „genereller“ als h'):

$$h \supseteq h' \Leftrightarrow \forall x \in \Omega : (h' \models x \Rightarrow h \models x)$$

Der Raum aller DNF-Hypothesen (disjunktive Normalform) ist die Boolesche Algebra $(\mathfrak{P}\Omega, \subseteq)$.

Lemma

Der Raum (\mathcal{H}, \subseteq) bildet eine Halbordnung.

$$\left\{ \begin{array}{l} \text{reflexiv} \\ \text{transitiv} \\ \text{antisymmetrisch} \end{array} \right\}$$

Die KPL-Hypothesen sind abgeschlossen gegenüber Durchschnittsbildung.

Die KPL-Hypothesen sind nicht abgeschlossen gegenüber der Mengenvereinigung, es existiert das Supremum je zweier Hypothesen:

$$(h \vee h')_n \stackrel{\text{def}}{=} \begin{cases} v & (\exists v \in \mathcal{X}_n) h_n = v = h'_n \\ ? & h_n \neq h'_n \\ ? & h_n = ? = h'_n \end{cases}$$

Sukzessiver Generalisierungsalgorithmus

Definition

Eine Hypothese $h \in \mathcal{H}$ heißt **konsistent** mit den Lerndaten (ω^+, ω^-) genau dann wenn gilt:

$$\begin{aligned} x \in \omega^+ &\Rightarrow h \models x \\ x \in \omega^- &\Rightarrow h \not\models x \end{aligned}$$

Bemerkungen

1. Konsistenz falls $\omega^+ \subseteq \Omega_h \subseteq \Omega \setminus \omega^-$
2. Jedes h ist konsistent mit (\emptyset, \emptyset) .
3. Kein h ist konsistent wenn $\omega^+ \cap \omega^- \neq \emptyset$.
4. Auch für disjunkte (ω^+, ω^-) enthält \mathcal{H} nicht notwendig eine konsistente Hypothese!

(Algorithmus)

1 INITIALISIERUNG

Setze $h \leftarrow h_\emptyset$.

2 GENERALISIERUNG

Setze für alle $x \in \omega^+$:

$$h \leftarrow h \vee x$$

(„speziellste Erweiterung“ von h um x)

3 TERMINIERUNG

Das Ergebnis ist h .Keine $x \in \omega^-$ verwendet.Resultat genügt allen $x \in \omega^+$.
 h ist minimal mit dieser Eigenschaft.

Wenn konsistente Hypothese existiert, wird sie gefunden.

Nur für KPL (Supremum!) realisierbar.

Für $\mathcal{H} = \wp \Omega$ ist SGA trivial.

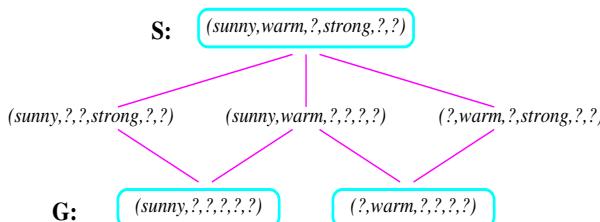
Der Versionenraum

Definition

Die Menge der mit den Lernbeispielen konsistenten Hypothesen

$$\{h \in \mathcal{H} \mid h \text{ konsistent mit } (\omega^+, \omega^-)\}$$

heißt **Versionenraum** von (ω^+, ω^-) bezüglich \mathcal{H} und wird mit $\mathfrak{V}(\mathcal{H}, \omega^+, \omega^-)$ (oder \mathfrak{V}) bezeichnet.



Beispiel

Versionenraum mit 6 Hypothesen

1x minimal
2x maximal
3x weder/noch

Minimale und maximale VR-Elemente

$$\mathfrak{V}_S \stackrel{\text{def}}{=} \{h \in \mathfrak{V} \mid \forall h' \in \mathfrak{V}: h' \subseteq h \Rightarrow h' = h\}$$

$$\mathfrak{V}_G \stackrel{\text{def}}{=} \{h \in \mathfrak{V} \mid \forall h' \in \mathfrak{V}: h \subseteq h' \Rightarrow h' = h\}$$

Kandidateneliminationsalgorithmus

Suche operiert auf („Kandidaten“-) Mengen von Hypothesen

(Algorithmus)

1 INITIALISIERUNG

Setze $H \leftarrow \mathcal{H}$

2 GENERALISIERUNG / SPEZIALISIERUNG

Eliminiere für alle $x \in \omega^+ \cup \omega^-$

- a Fall $x \in \omega^+$: alle $h \in H$ mit $h \not\models x$
- b Fall $x \in \omega^-$: alle $h \in H$ mit $h \models x$

3 TERMINIERUNG

Das Ergebnis ist h , falls $H = \{h\}$ ist.

Am Ende enthält die Kandidatenmenge genau die konsistenten Hypothesen aus \mathcal{H} .

Es gibt keine, eine oder mehrere Lösungen.

Das Verfahren ist aus Aufwandsgründen impraktikabel!

Versionenräume als Halbordnungsintervalle

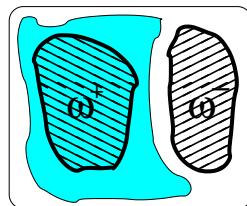
Beispiel

Im vollständigen Hypothesenraum $\mathcal{H} = \wp \Omega$ sind die minimalen und die maximalen VR-Elemente eindeutig:

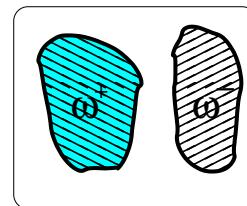
$$\mathfrak{V}_S = \{\omega^+\} \quad \text{und} \quad \mathfrak{V}_G = \{\Omega \setminus \omega^-\}$$

Versionenräume besitzen die Gestalt einer **Intervalldarstellung**:

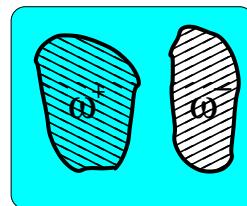
$$\mathfrak{V}(\mathcal{H}, \omega^+, \omega^-) = \{h \in \mathcal{H} \mid \omega^+ \subseteq h \subseteq \Omega \setminus \omega^-\} = [\mathfrak{V}_S, \mathfrak{V}_G]_{\mathcal{H}}$$



eine VR-Hypothese



die kleinste VR-Hypothese



die größte VR-Hypothese

Der Versionenraum-Darstellungssatz

Die Intervalldarstellung gilt in allen beliebigen Hypothesenräumen

Definition

Sei $\mathcal{H} \subseteq \mathfrak{P}\Omega$; ein **einfaches HO-Intervall** in \mathcal{H} hat die Form:

$$[h_u, h_o]_{\mathcal{H}} \stackrel{\text{def}}{=} \{h \in \mathcal{H} \mid h_u \subseteq h \subseteq h_o\}$$

Ein **verallgemeinertes HO-Intervall** in \mathcal{H} hat die Form:

$$[\mathcal{H}_u, \mathcal{H}_o]_{\mathcal{H}} = \{h \in \mathcal{H} \mid \exists h_u \in \mathcal{H}_u, \exists h_o \in \mathcal{H}_o : h_u \subseteq h \subseteq h_o\}$$

Satz

Für den Versionenraum \mathfrak{V} der Beispieldaten ω^+ und ω^- bezüglich \mathcal{H} gilt eine Intervalldarstellung:

$$\mathfrak{V}(\mathcal{H}, \omega^+, \omega^-) = [\mathfrak{V}_S, \mathfrak{V}_G]_{\mathcal{H}}$$

Dabei sind \mathfrak{V}_S und \mathfrak{V}_G die Mengen der \subseteq -minimalen (\subseteq -maximalen) Elemente des Versionenraums \mathfrak{V} .

$$[\mathcal{H}_u, \mathcal{H}_o]_{\mathcal{H}} = \bigcup_{h_u \in \mathcal{H}_u} \bigcup_{h_o \in \mathcal{H}_o} [h_u, h_o]_{\mathcal{H}}$$

Beweis.

Inklusionsrichtung \subseteq :

Sei $h \in \mathfrak{V}$. Sei $G(h) := \{h' \in \mathfrak{V} \mid h' \supseteq h\}$.

Wegen $h \in G(h)$ ist $G(h) \neq \emptyset$.

- Sei h_G ein maximales Element aus $G(h)$.
Dann ist $h_G \in \mathfrak{V}_G$ und $h_G \supseteq h$.

(Die Existenz eines $h_S \in \mathfrak{V}_S$ zeigt man/frau analog.)

Inklusionsrichtung \supseteq :

Sei $h \in \mathcal{H}$ mit $h \subseteq h_G \in \mathfrak{V}_G$ und $h \supseteq h_S \in \mathfrak{V}_S$.

Zu zeigen: h ist konsistent mit (ω^+, ω^-) , d.h. $h \in \mathfrak{V}$.

- Sei $x \in \omega^+$.
Wegen $h_S \in \mathfrak{V}_S \subseteq \mathfrak{V}$ gilt $h_S \models x$.
Wegen $h \supseteq h_S$ gilt auch $h \models x$.
- Sei $x \in \omega^-$.
Wegen $h_G \in \mathfrak{V}_G \subseteq \mathfrak{V}$ gilt $h_G \not\models x$.
Wegen $h \subseteq h_G$ gilt auch $h \not\models x$.

□

Versionenraum-Kandidateneliminationsalgorithmus

(Algorithmus)

1 INITIALISIERUNG

Setze $G \leftarrow \{\Omega\}$ und $S \leftarrow \{\emptyset\}$.

2+ POSITIVE BEISPIELE

Für alle $x \in \omega^+$:

- Entferne alle $h \in G$ mit $h \not\models x$
- Für alle $h \in S$:

Generalisiere h zu h' mit $h' \models x$

Behalte $h' \in S$, falls h' spezieller als G

- Entferne alle nichtminimalen $h \in S$

2- NEGATIVE BEISPIELE

Für alle $x \in \omega^-$:

- Entferne alle $h \in S$ mit $h \models x$
- Für alle $h \in G$:

Spezialisiere h zu h' mit $h' \not\models x$

Behalte $h' \in G$, falls h' allgemeiner als S

- Entferne alle nichtmaximalen $h \in G$

3 TERMINIERUNG

Das Ergebnis ist h , falls $G = \{h\} = S$ ist.

(summarizing)

Bemerkungen

- Grundidee: alle Versionenräume werden als „Intervalle“ $[S, G]$ abgespeichert, und auch die Hypothesenelimination geschieht auf S, G und nicht auf \mathfrak{V} .
- Es gilt natürlich $\mathcal{H} = [\emptyset, \Omega]_{\mathcal{H}}$.
- Wenn es geeignete Hypothesen mit $\Omega(h_\emptyset) = \emptyset$ und $\Omega(h_\Omega) = \Omega$ gibt, kann entsprechend initialisiert werden.
- Hypothesen $h \in G$, die einem Positivbeispiel $x \in \omega^+$ nicht genügen, dürfen ohne weiteres eliminiert werden, da jegliche Spezialisierung von h ebenfalls an x scheitern würde. Dasselbe gilt für $h \in S$, $x \in \omega^-$ mit $h \models x$.
- Gilt jedoch für $x \in \omega^+$ und ein $h \in S$ die Aussage $h \not\models x$, so darf h wegen der Gefährdung des Teilraums $[h, G]$ nicht einfach gelöscht werden!
- Von allen Generalisierungen h' von h mit $h' \models x$ interessieren natürlich nur diejenigen mit $[h', G] \neq \emptyset$, und die auch minimal sind in S mit dieser Eigenschaft.
- Am Ende sind alle Hypothesen aus S und aus G und auch aus $[S, G]$ konsistent mit den Beispieldaten, und $[S, G]$ ist auch diesbezüglich vollständig.

Beispiel (Segeln im KPL-Hypothesenraum)

Versionenraum nach VRE-Algorithmus

	sky	air	humidity	wind	water	forecast
$h_1 \in S$	sunny	warm	?	strong	?	?
h_2	sunny	?	?	strong	?	?
h_3	sunny	warm	?	?	?	?
h_4	?	warm	?	strong	?	?
$h_5 \in G$	sunny	?	?	?	?	?
$h_6 \in G$?	warm	?	?	?	?

Unbeobachtete („neue“) Objekte

Vorhersage des Konzepts „go_sailing“:

	sky	air	hum	wind	water	fore	S	h_1	h_2	h_3	h_4	G	h_5	h_6
x_1	sunny	warm	norm	strong	cool	change	1	1	1	1	1	1	1	1
x_2	rainy	cold	norm	weak	warm	same	0	0	0	0	0	0	0	0
x_3	sunny	warm	norm	weak	warm	same	0	0	1	0	1	1	1	1

Die induktive Hülle

Definition

Wir bezeichnen die Menge

$$\overline{\omega^+} \stackrel{\text{def}}{=} \{x \in \Omega \mid h \in \mathfrak{V}(\omega^+, \omega^-) \Rightarrow h \models x\}$$

als **induktive Hülle** der Positivbeispiele ω^+ und die Menge

$$\overline{\omega^-} \stackrel{\text{def}}{=} \{x \in \Omega \mid h \in \mathfrak{V}(\omega^+, \omega^-) \Rightarrow h \not\models x\}$$

als **induktive Hülle** der Negativbeispiele ω^- . Die Elemente aus

$$\omega^? \stackrel{\text{def}}{=} \Omega \setminus (\overline{\omega^+} \cup \overline{\omega^-})$$

heißen **ambige Objekte** von Ω bezüglich \mathcal{H} , ω^+ und ω^- .

Lemma

Die Operatoren $\omega^+ \mapsto \overline{\omega^+}$ und $\omega^- \mapsto \overline{\omega^-}$ sind tatsächlich Hüllenoperatoren:

1. $\omega_1^+ \subseteq \omega_2^+ \Rightarrow \overline{\omega_1^+} \subseteq \overline{\omega_2^+}$ (Monotonie)
2. $\omega^+ \subseteq \overline{\omega^+}$ und $\omega^- \subseteq \overline{\omega^-}$ (Inklusion)
3. $\overline{\omega^+} = \overline{\overline{\omega^+}}$ und $\overline{\omega^-} = \overline{\overline{\omega^-}}$ (Involution)

Parlamentarischer Alltag im Versionenraum

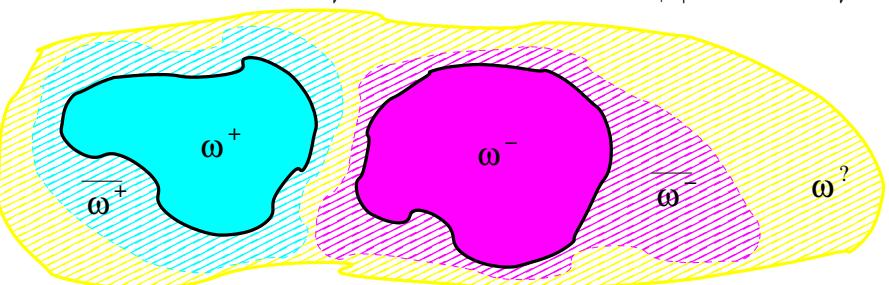
Positiver Konsens

Für alle $h \in \mathfrak{V}$ gilt
 $h \models x$

Negativer Konsens

Für alle $h \in \mathfrak{V}$ gilt
 $h \not\models x$

Ambiges Votum

Ex. $h_+, h_- \in \mathfrak{V}$ mit
 $h_+ \models x$ und $h_- \not\models x$ 

Bemerkungen

1. Alle $h \in \mathfrak{V}$ sind konsistent $\Rightarrow \left\{ \begin{array}{l} \text{alle } x \in \omega^+ \text{ werden einstimmig akzeptiert} \\ \text{alle } x \in \omega^- \text{ werden einstimmig abgewiesen} \end{array} \right\}$
2. Für $\mathfrak{V} = \emptyset$ folgt $\overline{\omega^+} \cap \overline{\omega^-} = \emptyset$.
3. Ist $x \in \Omega$ ambig, so ex. Hypothesen $h^+ \in \mathfrak{V}_G$, $h^- \in \mathfrak{V}_S$ mit $\left\{ \begin{array}{l} h^+ \models x \\ h^- \not\models x \end{array} \right\}$.

Der induktive Bias

Großartiger Lernerfolg durch mangelhafte Ausdrucksfähigkeit

Induktives Schließen

steht und fällt mit dem **Ausdrucksdefizit** des Hypothesenraums \mathcal{H} .

ω^+	sunny	warm	normal	strong	warm	same
ω^+	rainy	warm	normal	weak	warm	same
\mathfrak{V}_S	?	warm	normal	?	warm	same
$\omega^?$	sunny	warm	normal	weak	warm	same

KPL: $x_i = \xi$ oder ? $x_i = \xi$ oder $x_i \neq \xi$ oder ? $x_i \in \mathcal{X}^+$

(dual zu oben)

 $\mathcal{H} = \mathfrak{P}\Omega$

Aussagenlogisch orientierte Hypothesenräume

- Konjunktion positiver Literale
- Konjunktion positiver und negativer Literale $x_i = \xi$ oder $x_i \neq \xi$ oder ?
- Konjunktion disjunktiver Komplexe
- Disjunktion positiver (und negativer) Literale
- Disjunktion von Konjunktionen positiver Literale

Lernen einelementiger Versionenräume

Zur Auswahl neuer Lernbeispiele

- Erweiterung von ω^+ um Beispiele aus $\overline{\omega^+}$ ist überflüssig.
Erweiterung von ω^- um Beispiele aus $\overline{\omega^-}$ ist überflüssig.
- Erweiterung von ω^+ um Beispiele aus $\overline{\omega^-}$ bewirkt Inkonsistenz.
Erweiterung von ω^- um Beispiele aus $\overline{\omega^+}$ bewirkt Inkonsistenz.
- Nur die Erweiterung von (ω^+, ω^-) um ambige Beispiele $x \in \omega^?$ ist zugleich **konsistent** und **produktiv**!

Exploratives Lernen

Sukzessives Akquirieren produktiver neuer Beispiele, bis

1. der Versionenraum \mathfrak{V} nur noch ein h enthält oder
2. der Versionenraum \mathfrak{V} leergelaufen ist.

Ambiguität und Rückweisung

Votierungstechniken für die Entscheidungsphase

Faules Lernen

Fallbasiertes Schließen

$$x \mapsto \begin{cases} \Omega^+ & x \in \omega^+ \\ \Omega^- & x \in \omega^- \\ \Omega^? & \text{sonst} \end{cases}$$

(keine Verallgemeinerung)

Fleißiges Lernen

einer Hypothese ($\mathfrak{V} = \{h^*\}$)

$$x \mapsto \begin{cases} \Omega^+ & h^* \models x \\ \Omega^- & h^* \not\models x \end{cases}$$

Orakel

Occam's Razor (MDL, BIC, AIC)

Wahrscheinlichkeiten ...

Einstimmigkeit

$$x \mapsto \begin{cases} \Omega^+ & h \in \mathfrak{V}_S \Rightarrow h \models x \\ \Omega^- & h \in \mathfrak{V}_G \Rightarrow h \not\models x \\ \Omega^? & \text{sonst} \end{cases}$$

Generalkonsens

$$x \mapsto \begin{cases} \Omega^+ & h \in \mathfrak{V}_G \Rightarrow h \models x \\ \Omega^- & \text{sonst} \end{cases}$$

Mehrheitsvotum

$$x \mapsto \begin{cases} \Omega^+ & |\{h \in \mathfrak{V} \mid h \models x\}| > |\mathfrak{V}|/2 \\ \Omega^- & |\{h \in \mathfrak{V} \mid h \models x\}| < |\mathfrak{V}|/2 \\ \Omega^? & |\{h \in \mathfrak{V} \mid h \models x\}| = |\mathfrak{V}|/2 \end{cases}$$

Gibbs-Sampling

Auswürfeln von $h^* \in \mathfrak{V}$ und

$$x \mapsto \begin{cases} \Omega^+ & h^* \models x \\ \Omega^- & h^* \not\models x \end{cases}$$

Lernen mit Orakel

INITIALISIERUNG

Setze $G \leftarrow \{\Omega\}$ und $S \leftarrow \{\emptyset\}$ und $\omega^? \leftarrow \Omega$.

EXPLORATIONSSCHRITT

Solange $\omega^? \neq \emptyset$ gilt:

- Wähle ein Beispiel $x \in \omega^?$ aus
- Befrage das Orakel nach $x \in \mathcal{C}$
- Modifiziere den Versionenraum vermöge

$$\mathfrak{V} \leftarrow \mathfrak{V}(\omega^+ \cup \{x\}, \omega^-)$$

im Fall einer positiven Antwort und vermöge

$$\mathfrak{V} \leftarrow \mathfrak{V}(\omega^+, \omega^- \cup \{x\})$$

im Fall einer negativen Antwort des Orakels.

- Aktualisiere die Menge $\omega^?$ der ambigen Objekte

TERMINIERUNG

Das Ergebnis ist h , falls $G = \{h\} = S$ gilt.

ILP — Induktive logische Programmierung

Hypothesenraumbias wird explizit durch eine logische Theorie \mathcal{B} vorgegeben

Gegeben

Hypothesen $h \in \mathcal{H}$ sind **prädikatenlogische** Formeln

Objekte $x \in \Omega$ als Singleton-Hypothesen h_x

Positive und negative Lerndatensätze ω^+, ω^-

p.l. Formelmenge \mathcal{B} als expliziter Bias („Sachbereichstheorie“)

Gesucht

Eine Hypothese $h \in \mathcal{H}$ mit den Eigenschaften

1. Vollständigkeit

$$\mathcal{B}, h \models \omega^+$$

2. Korrektheit

für alle $x \in \omega^-$ gilt $\mathcal{B}, h \not\models x$

3. Konsistenz

$$\mathcal{B}, h, \omega^+, \omega^- \not\models \text{false}$$

die zudem ein Gütekriterium $\begin{cases} \text{speziell} \\ \text{generell} \\ \text{interessant} \\ \text{kurz} \end{cases}$ optimiert.

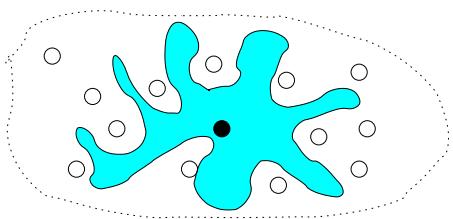
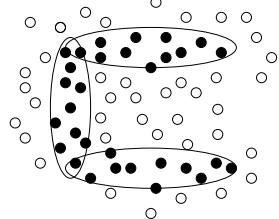
⇒ nicht entscheidbar in der Prädikatenlogik erster Stufe

Michalskis Stern

Abgrenzung eines Positivbeispiels gegen alle Negativbeispiele

Problem

Die Menge ω^+ ist schwer gegen ω^- abgrenzbar.
 ω^+ zerfällt jedoch in einfacher strukturierte Teilmengen.



Definition

Es seien $\omega^+ \subset \Omega$, $\omega^- \subset \Omega$ und $x \in \omega^+$ ein Positivbeispiel. Die Hypothesenmenge

$$\mathcal{S}(x|\omega^-) \stackrel{\text{def}}{=} \mathfrak{V}_G(\mathcal{H}, \{x\}, \omega^-)$$

heißt **Stern** von x gegen ω^- .

Achtung!

Der Stern ist keine Hypothese, sondern ein Intervall.

Sterne und ihre Vereinigung

Lemma

Für jede Hypothese $h \in \mathcal{S}(x|\omega^-)$ gilt:

1. h wird von x erfüllt. $h \models x$
2. h wird von keinem $y \in \omega^-$ erfüllt. $(\forall y \in \omega^-) h \not\models y$
3. h ist maximal mit diesen Eigenschaften, d.h., es gilt:

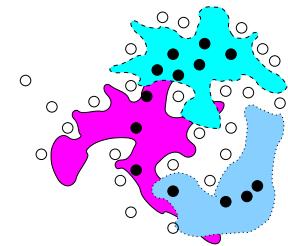
$$h' \supset h \rightarrow h' \not\models x \text{ oder ex. } y \in \omega^- : h' \models y$$

Lemma

Aus **nichtleeren** Sternen lassen sich konsistente Disjunktionen konstruieren, d.h. die Vereinigungsmenge

$$h^* = \bigcup_{x \in \omega^+} \mathcal{S}(x|\omega^-)$$

ist konsistent mit (ω^+, ω^-) .



\oplus Disjunktion \cdot \ominus Hypothese

Sternerzeugungsalgorithmus

(Algorithmus)

- 1 Wählte zufällig ein $x^* \in \omega^+$.
- 2 Erzeuge den Stern $\mathcal{S}(x^*|\omega^-) = \mathfrak{V}_G(\mathcal{H}, \{x^*\}, \omega^-)$.
- 3 Wähle eine Vorzugshypothese $h^* \in \mathcal{S}(x^*|\omega^-)$ mit maximaler Präferenz $\gamma(h^*)$.
- 4 Wenn $h^* \models x$ für alle $x \in \omega^+$, so \rightsquigarrow 6.
- 5 Tilge alle $x \in \omega^+$ mit $h^* \models x$ und \rightsquigarrow 1.
- 6 Bilde die logische Disjunktion

$$h_{\text{dis}} \stackrel{\text{def}}{=} h_1 \vee \dots \vee h_r$$

aller bislang erzeugten Hypothesen.

(summitnogIA)

Gegeben

\mathcal{H} , ω^+ , ω^- und eine **Präferenzfunktion** $\gamma: \mathcal{H} \mapsto \mathbb{IR}$

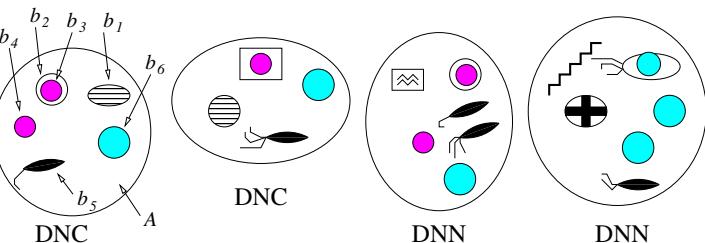
Gesucht

Eine disjunktive Beschreibung $h_1 \vee \dots \vee h_r$, $h_i \in \mathcal{H}$, die konsistent mit (ω^+, ω^-) ist.

Beispiel — Konzeptualisierung von Krebszellen

Aufgabenstellung

Unterscheide Krebszellen (DNC) von gesunden Zellen (DNN) auf Grundlage numerischer, kategorialer und struktureller Zellmerkmale.



Mensch-Maschine-Mensch-Zyklus

- (1) Definiere relevante Deskriptoren · etikettiere (ω^+, ω^-)
- (2) Lerne induktiv passende Hypothesen für $\mathcal{C} = \mathcal{C}_{\text{DNC}}$.
- (3) Evaluiere, analysiere und modifiziere das Szenarium.

Beispiel — Konzeptualisierung von Krebszellen

Globale (zellbezogene) Merkmale

1. $circ \in \{1, 2, \dots, 10\}$ (Anzahl der Zellsegmente)
2. $pplasm \in \{A, B, C, D\}$ (Protoplasmatyp der Zelle)

Lokale (segmentbezogene) Merkmale

- $shape(i) \in \{\text{triangle}, \text{circle}, \text{ellipse}, \text{heptagon}, \text{square}, \text{boat}, \text{spring}\}$ (bzw. eine Baumstruktur dieser Formklassen)
- $texture(i) \in \{\text{blank}, \text{shaded}, \text{black}, \text{grey}, \text{stripes}, \text{crossed}, \text{wavy}\}$
- $weight(i) \in \{1, 2, 3, 4, 5\}$
- $orient(i) \in \{N, NE, E, SE, S, SW, W, NW\}$
- $contains(c, b_1, b_2, \dots) \in \{T, F\}$
- $hastails(c, b_1, b_2, \dots) \in \{T, F\}$

Beispiel — Konzeptualisierung von Krebszellen

Objektbeschreibung der ersten DNC-Zelle

$$\begin{aligned} & contains(c, b_1, \dots, b_6) \wedge circ(c) = 8 \wedge pplasm(c) = A \\ & \wedge shape(b_1) = \text{ellipse} \wedge texture(b_1) = \text{stripes} \wedge weight(b_1) = 4 \\ & \wedge orient(b_1) = NW \wedge shape(b_2) = \text{circle} \wedge contains(b_2, b_3) \\ & \wedge texture(b_2) = \text{blank} \wedge weight(b_2) = 3 \wedge \dots \\ & \wedge shape(b_6) = \text{circle} \wedge texture(b_6) = \text{shaded} \wedge weight(b_6) = 5 \end{aligned}$$

DNC-Charakterisierung durch prädikatenlogische Formel

$$\begin{aligned} & \exists_1 b \ (weight(b) = 5) \\ & \exists_1 b \ (shape(b) = \text{circle} \wedge texture(b) = \text{shaded} \wedge weight(b) \geq 3) \\ & \exists b_1 \exists b_2 \ (contains(b_1, b_2) \wedge shape(b_1) = \text{circle} \wedge shape(b_2) = \text{circle}) \\ & \dots \wedge \dots \wedge \dots \end{aligned}$$

Konzeptuelle Klassifikation

Gegeben

Klassenspezifische Lernstichproben

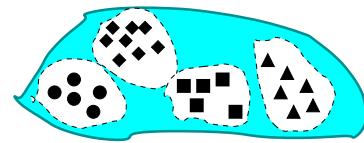
$$\omega_\kappa \subseteq C_\kappa \subseteq \Omega, \quad \kappa = 1, \dots, K$$

für die Konzepte $C_1, \dots, C_K \in \mathcal{C}$ mit
 $C_\kappa \cap C_\lambda = \emptyset$ für $\kappa \neq \lambda$.

Gesucht

Ein konsistentes System $h_1, \dots, h_K \in \mathcal{H}$, d.h. für alle $1 \leq \kappa \leq K$ gilt:

$$\left\{ \begin{array}{ll} h_\kappa \models x & x \in \omega_\kappa \\ h_\kappa \not\models x & x \in \omega_\lambda, \lambda \neq \kappa \end{array} \right. \quad \text{und „quodlibet“ sonst}$$



Versionenraum-Methode

Berechne für jede Objektklasse κ einen **diskriminativen VR**

$$\mathfrak{V}_\kappa = \mathfrak{V}(\mathcal{H}, \omega_\kappa, \bigcup_{\lambda \neq \kappa} \omega_\lambda)$$

Stern-Methode

Berechne für jedes κ eine **Sternidisjunktion**

$$h_\kappa^* = \bigcup_{x \in \omega_\kappa} \mathcal{S}(x | \omega \setminus \omega_\kappa)$$

Votierung beim K -Klassen-Problem

$$\beta(x) = (\beta_1, \dots, \beta_K) \in \{1, 0, ?\}^K$$

Stern-Methode

Es gibt keine Fehlanzeigen.
 Aber es gibt u.U. Konflikte.
 Und es gibt u.U. Leerrunden.

Versionenraum-Methode

Es gibt Konflikte & Leerrunden.
 Es gibt auch Fehlanzeigen:
 $\left\{ \begin{array}{l} \text{eine FA statt PRO} \\ \text{weniger als } K - 1 \text{ CONs} \end{array} \right\}$

Definition

Sei $\mathcal{A} \subseteq \Omega$. Das Hypothesensystem (h_1, \dots, h_K) heißt **konzeptuelle Partition** von \mathcal{A} , wenn es für jedes $x \in \mathcal{A}$ einen Klassenindex κ gibt mit

$$\forall \lambda = 1, \dots, K : \quad (h_\lambda \models x \Leftrightarrow \lambda = \kappa)$$

Bemerkungen

1. Ein konsistentes System (h_1, \dots, h_K) ist konzeptuelle Partition seiner Lerndaten $\bigcup_\kappa \omega_\kappa$.
2. Läßt sich jedes konsistente System zu einer konzeptuellen Partition von Ω erweitern?

Versionenräume

Naive Bayesregel

Multivariate lineare Regression

Logistische Regression

Ordinale Regression und Präferenzmodelle

Statistische Entscheidungsbäume

Zusammenfassung

Satz

Ist für den Objektraum \mathcal{X} und das Klasseninventar $\mathcal{K} = \{1, \dots, K\}$ die wahre Verbundverteilung $P(\kappa, \mathbf{x})$ bekannt, so liefert die **Bayesentscheidungsregel** (MAP-Regel)

$$\begin{aligned}\delta(\mathbf{x}) &= \operatorname{argmax}_{\kappa \in \mathcal{K}} P(\kappa | \mathbf{x}) \\ &= \operatorname{argmax}_{\kappa \in \mathcal{K}} \frac{P(\kappa) \cdot P(\mathbf{x} | \kappa)}{P(\mathbf{x})}\end{aligned}$$

die minimale erwartete Klassifikationsfehlerrate.

Marginal 1

$$P(\mathbf{x}) = \sum_{\kappa=1}^K P(\kappa, \mathbf{x})$$

Marginal 2

$$P(\kappa) = \sum_{\mathbf{x} \in \mathcal{X}} P(\kappa, \mathbf{x})$$

Bedingte Vtl.

$$P(\mathbf{x} | \kappa) = \frac{P(\kappa, \mathbf{x})}{P(\kappa)}$$

Posterior Vtl.

$$P(\kappa | \mathbf{x}) = \frac{P(\kappa, \mathbf{x})}{P(\mathbf{x})}$$

Versionenräume	Bayesregel	Regression	Logitmodell	Präferenzen	CART	Σ
Die Bayesregel für diskrete Attribute						

Die Bayesregel für diskrete Attribute

Kanonische multivariat-diskrete Verteilung („Hypertabelle“)

Lemma

Der Objektraum $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_N$ enthalte ausschließlich **diskrete** Attribute mit Wertebereichen \mathcal{X}_n der Größe $L_n = |\mathcal{X}_n|$.

1. Die gemeinsame Verteilung $P(\kappa, \mathbf{x})$ ist durch die $K \cdot L_1 \cdot \dots \cdot L_N$ Einträge

$$p_{\kappa, x_1, \dots, x_N} = P(\kappa, \mathbf{x}), \quad \kappa \in \mathcal{K}, \mathbf{x} \in \mathcal{X}$$

eines $(1 + N)$ -dimensionalen Hyperwürfels $\mathbf{P} \in [0, 1]^{K \times L_1 \times \dots \times L_N}$ charakterisiert.

2. Für einen etikettierten Lerndatensatz $\{(\kappa_t, \mathbf{x}_t) \mid t = 1..T\}$ mit den absoluten Häufigkeiten $T_{\kappa, \mathbf{x}}$, $(\kappa, \mathbf{x}) \in \mathcal{K} \times \mathcal{X}$, lauten die Maximum-Likelihood-Parameter

$$\hat{b}_{\kappa, x_1, \dots, x_N} = T_{\kappa, x_1, \dots, x_N} / T.$$

3. Die Bayesentscheidungsregel lautet $\delta(\mathbf{x}) = \operatorname{argmax}_{\kappa \in \mathcal{K}} \hat{b}_{\kappa, x_1, \dots, x_N}$.

Versionenräume	Bayesregel	Regression	Logitmodell	Präferenzen	CART	Σ
Beispiel — kanonische Bayesregel ... mit ML-geschätzten Verteilungsparametern						

Lerndatensammlung „Tennis“

ω	outlook	temp	humid	wind	Tennis?
o ₁	sunny	hot	high	weak	no
o ₂	sunny	hot	high	strong	no
o ₃	overcast	hot	high	weak	yes
o ₄	rain	mild	high	weak	yes
o ₅	rain	cool	normal	weak	yes
o ₆	rain	cool	normal	strong	no
o ₇	overcast	cool	normal	strong	yes
o ₈	sunny	mild	high	weak	no
o ₉	sunny	cool	normal	weak	yes
o ₁₀	rain	mild	normal	weak	yes
o ₁₁	sunny	mild	normal	strong	yes
o ₁₂	overcast	mild	high	strong	yes
o ₁₃	overcast	hot	normal	weak	yes
o ₁₄	rain	mild	high	strong	no
o _{neu}	sunny	cool	high	strong	?

Parameter
 $2 \cdot 3^2 \cdot 2^2 = 72$
Einträge

14 Einsen
58 Nullen

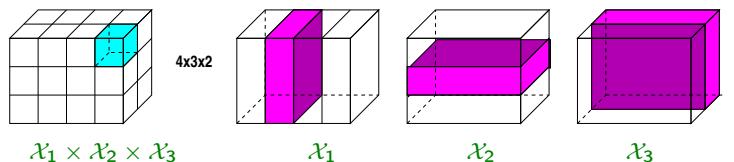
Neuzugang
Nulleintrag bei
(yes, o_{neu}) und
(no, o_{neu}).
Nennerausdruck
 $\hat{P}(o_{neu}) = 0$

Dann gilt für die a posteriori Wahrscheinlichkeit:

$$P(\text{no} \mid (\text{sunny}, \text{cool}, \text{high}, \text{strong})^\top) = \text{undef.}$$

Die naive Bayesregel

Klassenbedingte statistische Unabhängigkeit zwischen allen Objektattributen



NBK-Entscheidungsregel

$$\delta(\mathbf{x}) = \operatorname{argmax}_{\kappa \in \mathcal{K}} P(\kappa, \mathbf{x}) = \operatorname{argmax}_{\kappa \in \mathcal{K}} \left\{ P(\kappa) \cdot \prod_{n=1}^N P(x_n | \kappa) \right\}$$

(maximale faktorierte Verbundwahrscheinlichkeit)

Modellparameter und ihre ML-Schätzwerte

$$\hat{a}_\kappa = \frac{T_\kappa}{T}, \quad \hat{b}_{\xi|\kappa,n} = \frac{T_{\kappa,n,\xi}}{T_\kappa}, \quad T_\kappa = \sum_{\xi \in \mathcal{X}_1} T_{\kappa,1,\xi}, \quad \begin{cases} \kappa = 1..K \\ n = 1..N \\ \xi = 1..L_n \end{cases}$$

Das sind $K \cdot \sum_n L_n$ Parameter statt $K \cdot \prod_n L_n$ Parameter!

Beispiel — naive Bayesregel

... mit ML-geschätzten Verteilungsparametern

Attribut „outlook“

	sunny	over	rain	Σ
yes	2	4	3	9
no	3	0	2	5
Σ	5	4	5	14

Attribut „humidity“

	high	normal	Σ
yes	3	6	9
no	4	1	5
Σ	7	7	14

Attribut „temp“

	hot	mild	cool	Σ
yes	2	4	3	9
no	2	2	1	5
Σ	4	6	4	14

	weak	strong	Σ
yes	6	3	9
no	2	3	5
Σ	8	6	14

Parametertabelle und Neuklassifikation (6 + 6 + 4 + 4 = 20 Einträge)

$$\begin{aligned} P(\text{no, sunny, cool, high, strong}) &= \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} = \frac{180}{8750} = 0.02057 \\ P(\text{yes, sunny, cool, high, strong}) &= \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} = \frac{486}{91854} = 0.005291 \\ P(\text{no} | (\text{sunny, cool, high, strong})^\top) &= \frac{0.02057}{0.02057 + 0.005291} = 0.7954 \end{aligned}$$

Versionenräume
Bayesregel
Regression
Logitmodell
Präferenzen
CART
 Σ

NTF — Nichtnegative Tensorfaktorisierung
Mischung naiver Verbundverteilungen von $N \in \{2, 3\}$ nominalen Attributen

Matrix ($\Omega_1 \times \Omega_2$)

Verteilungsparameter

$$P(i,j) =: x_{ij}$$

Naive Faktorisierung

$$P(i,j) = p_1(i) \cdot p_2(j)$$

Mischungsmodell

$$P(i,j) = \sum_{m=1}^M \underbrace{\pi_m \cdot p_1^{(m)}(i)}_{v_{im}} \cdot \underbrace{p_2^{(m)}(j)}_{a_{jm}}$$

Reduktion

$$L_1 \cdot L_2 \rightarrow M \cdot (1 + L_1 + L_2)$$

Würfel ($\Omega_1 \times \Omega_2 \times \Omega_3$)

Verteilungsparameter

$$P(i,j,k) =: x_{ijk}$$

Naive Faktorisierung

$$P(i,j,k) = p_1(i) \cdot p_2(j) \cdot p_3(k)$$

Mischungsmodell

$$P(i,j,k) = \sum_{m=1}^M \pi_m \cdot p_1^{(m)}(i) \cdot p_2^{(m)}(j) \cdot p_3^{(m)}(k)$$

Reduktion

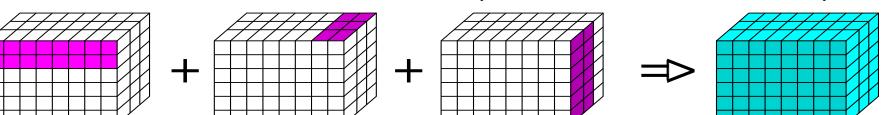
$$L_1 \cdot L_2 \cdot L_3 \rightarrow M \cdot (1 + L_1 + L_2 + L_3)$$

Wahrscheinlichkeitshyperwürfel ($\Omega_1 \times \Omega_2 \times \dots \times \Omega_N$)

Naive Mischung

$$P(x_1, \dots, x_N) = \sum_{m=1}^M \pi_m \cdot \prod_{n=1}^N p_n^{(m)}(x_n)$$

Parameter lernen nach EM-Prinzip (expectation-maximization)



Reduktion

$$L_1 \cdot \dots \cdot L_N \rightarrow M \cdot (1 + L_1 + \dots + L_N)$$

$$\prod \rightsquigarrow M \cdot \sum$$

EM-Algorithmus für das NTF-Modell

- 1 Initialisierung
- 2 E-Schritt
- 3 M-Schritt
- 4 Abbruch

A posteriori Wahrscheinlichkeiten der Komponentenauswahl

Für jedes Lerndatensatzobjekt x_1, \dots, x_T berechne

$$\gamma_t(m) \stackrel{\text{def}}{=} P(\mathbb{M} = m \mid x_t, \theta^{\text{alt}}) = \pi_m^{\text{alt}} \cdot \prod_{n=1}^N \theta_{m,n,x_{tn}}^{\text{alt}} / P^{\text{alt}}(x_t)$$

Neuschätzung durch a posteriori Erwartungswerte

$$\hat{\pi}_m = \sum_{t=1}^T \gamma_t(m) / T \quad \text{und} \quad \hat{\theta}_{m,n,\xi} = \sum_{t=1}^T \gamma_t(m) \cdot \mathbf{1}_{x_{tn}=\xi} \Bigg/ \sum_{t=1}^T \gamma_t(m)$$

Startparameter

zufällig · wiederholt · lokale Optima

Rechenaufwand

$O(I_{\max} \cdot T \cdot M \cdot (N + \sum_n L_n))$

Bayesregel für gemischte Attributskalen

$$z = (x, y) \in \underbrace{\mathbb{R} \times \dots \times \mathbb{R}}_{\Omega' = \mathbb{R}^{N'}} \times \underbrace{\mathcal{X}_1 \times \dots \times \mathcal{X}_{N''}}_{\Omega''}$$

Normalzerlegung

$$P(z) = P(y) \cdot P(x|y) = P(y_1, \dots, y_{N''}) \cdot \mathcal{N}(x | \mu_y, \mathbf{S}_y)$$

- $L^\times = \prod_{n=1}^{N''} L_n$ kanonische W'keitsparameter
- $N' + \binom{N'}{2}$ Dichteparameter je NV-Dichte
- ⇒ insgesamt $O(K \cdot L^\times \cdot N'^2)$ Parameter
- Unabhängigkeitsannahme für Ω'' bringt wenig Vorteile.
- Unabhängigkeit in Ω' reduziert auf $O(K \cdot L^\times \cdot N')$ Parameter.

Versionenräume

Naive Bayesregel

Multivariate lineare Regression

Logistische Regression

Ordinale Regression und Präferenzmodelle

Statistische Entscheidungsbäume

Zusammenfassung

Diskriminative Klassifikatoren

$$\kappa(x) = \operatorname{argmax}_{\lambda} h_{\lambda}(x)$$

Definition

Es sei C_1, \dots, C_K ein K -Klassen-Problem über $\Omega = \mathbb{R}^N$. Die Elemente von $\mathbf{h} = (h_1, \dots, h_K)^\top$ mit

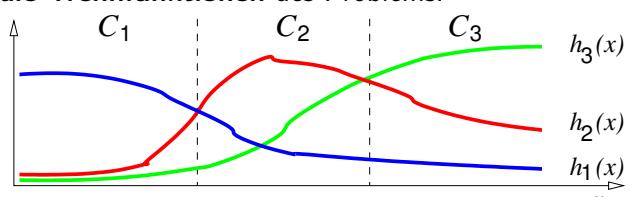
$$h_{\kappa} : \mathbb{R}^N \rightarrow \mathbb{R}, \quad \kappa = 1, \dots, K$$

heißen **Trennfunktionen** der Klassen $\kappa = 1, \dots, K$.

Die Abbildungen $\mathbf{d} = (d_1, \dots, d_K)^\top$ mit

$$d_{\kappa} : x \mapsto \begin{cases} 1 & x \in C_{\kappa} \\ 0 & x \notin C_{\kappa} \end{cases}$$

heißen **ideale Trennfunktionen** des Problems.



Quadratmittelklassifikator

Willkürlicher Zielausdruck — willkürliche Straffunktional

Definition

Es sei C_1, \dots, C_K ein K -Klassen-Problem über $\Omega = \mathbb{R}^N$ und \mathcal{H} eine Menge von Trennfunktionen. Die Trennfunktion $\mathbf{h} \in \mathcal{H}$ mit minimalem erwarteten quadratischen Fehler

$$\varepsilon(\mathbf{h}) \stackrel{\text{def}}{=} \mathbb{E}[\|\mathbf{h}(\mathbb{X}) - \mathbf{d}(\mathbb{X})\|^2]$$

heißt **Quadratmitteldiskriminante**, der zugehörige Klassifikator heißt **Quadratmittelklassifikator**.

Sind ferner die Lerndaten $\omega_1, \dots, \omega_K$ gegeben, so heißt der Klassifikator mit minimalem Fehler

$$\varepsilon(\mathbf{h}, \{\omega_\kappa\}) \stackrel{\text{def}}{=} \sum_{\kappa=1}^K \sum_{\mathbf{x} \in \omega_\kappa} \|\mathbf{h}(\mathbf{x}) - \mathbf{e}^{(\kappa)}\|^2$$

empirischer QMK. Dabei bezeichne $\mathbf{e}^{(\kappa)}$ den κ -ten Einheitsvektor.

Lernen als skalare Regressionsaufgabe

Zerlegung in Zweiklassenprobleme

Für jedes κ ergibt sich das QM-Approximationsproblem

$$h_\kappa \approx d_\kappa : \mathbf{x} \mapsto \begin{cases} 1 & \mathbf{x} \in \omega^+, \omega^+ = \omega_\kappa \\ 0 & \mathbf{x} \in \omega^-, \omega^- = \bigcup_{\lambda \neq \kappa} \omega_\lambda \end{cases}$$

Skalares Regressionsproblem

Für die Daten $\{(x_t, y_t) \in \mathbb{R}^N \times \mathbb{R} \mid t = 1, \dots, T\}$ finde die Regressionsfunktion $h \in \mathcal{H}$ mit minimalem Fehler

$$\varepsilon(h) \stackrel{\text{def}}{=} \sum_{t=1}^T (y_t - h(x_t))^2$$

$$\left\{ \begin{array}{l} \text{quadratisches Fehlermaß} \\ \{0, 1\}\text{-Zielgröße} \end{array} \right\} \rightarrow \left\{ \begin{array}{l} \text{entkoppelte Diskriminantfunktionen} \\ \text{kein Problem mit } K > 2 \end{array} \right\}$$

Multivariate lineare Regression

Linearer Ansatz

$$h(\mathbf{x}) = \sum_n a_n \cdot x_n = \mathbf{a}^\top \mathbf{x}, \quad \mathbf{x}, \mathbf{a} \in \mathbb{R}^N$$

Affiner Ansatz

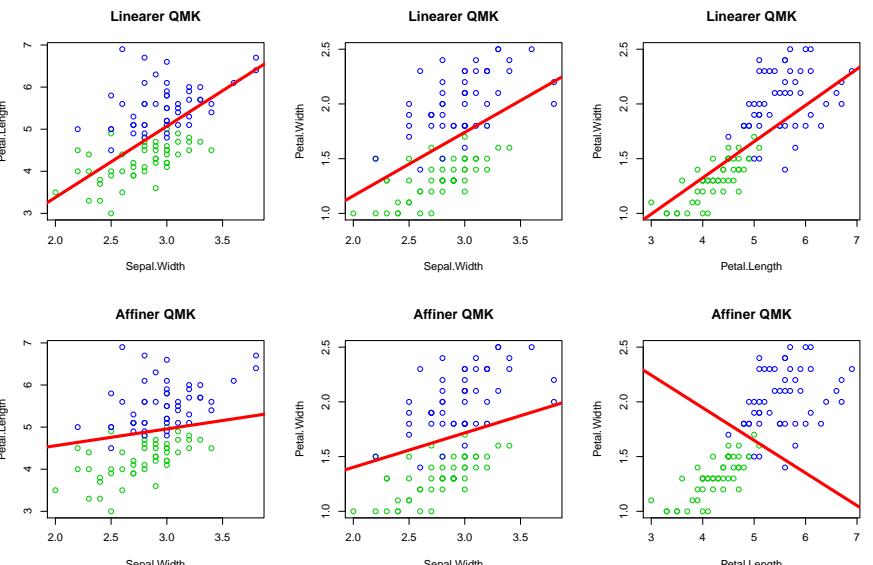
$$h(\mathbf{x}) = a_0 + \sum_n a_n \cdot x_n = \mathbf{a}^\top \mathbf{x}', \quad \left\{ \begin{array}{l} \mathbf{a} \in \mathbb{R}^{N+1} \\ \mathbf{x}' \stackrel{\text{def}}{=} (1, \mathbf{x}^\top)^\top \end{array} \right.$$

Was heißt hier eigentlich „Regressionsproblem“?

- Datenmodell $\mathbb{Y} = h(\mathbb{X}_1, \dots, \mathbb{X}_N) + \mathbb{E}$ mit Störterm $\mathbb{E} \sim \mathcal{N}(0, \sigma^2)$
- Datenprobe $\{(\mathbf{x}_t, y_t)\}_1^T$ bzw. gemeinsame Datenverteilung $f_{\mathbb{X}, \mathbb{Y}}(\cdot, \cdot)$
- Posterior-Erwartungswerte $\hat{h}(\mathbf{x}) = \mathbb{E}_{\mathbb{Y} \mid \mathbf{x}}[\mathbb{Y}]$, also $\hat{h}(\mathbf{x}) = \int f(y|\mathbf{x}) \cdot y \, dy$

Linearer vs. affiner Quadratmittelklassifikator

Beispiel: Iris-Datensatz, 2D-Träger für einige (x_i, x_j) -Kombinationen



Linearer versus affiner Ansatz

Lineare Funktionen allein beschreiben wegen $h(\mathbf{0}) = 0$ ausschließlich Hyperebenen, die durch den Koordinatenursprung verlaufen und sind als Regressionsmodell unzureichend. Affine Funktionen verfügen zusätzlich über den y -Schnittpunkt a_0 (*intercept*); affine Regression kann aber leicht auf lineare Regression zurückgeführt werden. Wir verwenden wieder die Notation \mathbf{X}, \mathbf{y} für Datenmatrix und Zielwertevektor und betrachten den Abweichungsvektor $\mathbf{y} - a_0 \mathbf{1} - \mathbf{X}\mathbf{a}$ des affinen Modells sowie den resultierenden quadratischen Fehler:

$$\begin{aligned}\varepsilon(a_0, \mathbf{a}) &= \|\mathbf{y} - a_0 \mathbf{1} - \mathbf{X}\mathbf{a}\|^2 \\ &= (\mathbf{y} - a_0 \mathbf{1} - \mathbf{X}\mathbf{a})^\top \cdot (\mathbf{y} - a_0 \mathbf{1} - \mathbf{X}\mathbf{a}) \\ &= (\mathbf{y}^\top - a_0 \mathbf{1}^\top - \mathbf{a}^\top \mathbf{X}^\top) \cdot (\mathbf{y} - a_0 \mathbf{1} - \mathbf{X}\mathbf{a}) \\ &= \underbrace{\|\mathbf{y}\|^2}_{c} + \underbrace{a_0^2 T - 2a_0 T \mu_y}_{\varepsilon(a_0)} + \underbrace{\mathbf{a}^\top \mathbf{X}^\top \mathbf{X}\mathbf{a} - 2\mathbf{a}^\top \mathbf{X}^\top \mathbf{y}}_{\varepsilon(\mathbf{a})} + \underbrace{2a_0 T \mu_x^\top \mathbf{a}}_0\end{aligned}$$

Wenn wir o.B.d.A. mittelwertfreie Vektordaten annehmen ($\mu_x = \mathbf{0}$), so verschwindet der Kopplungsterm und wir dürfen a_0 und \mathbf{a} separat optimieren.

Für a_0 ergibt sich nach Nullsetzen der Ableitung

$$\partial \varepsilon(a_0) / \partial a_0 = 2Ta_0 - 2T\mu_y$$

der Minimalwert $a_0 = \mu_y$. Der Fehler $\varepsilon(\mathbf{a})$ und die Konstante $\|\mathbf{y}\|^2$ ergeben zusammen den Minimierungsausdruck der linearen Regressionsaufgabe ...

Beweis.

Zur Lösung des Quadratmittelproblems setzen wir die partiellen Ableitungen der Koeffizienten a_1, \dots, a_N gleich Null — wir verwenden die Gradientenvektorschreibweise:

$$\begin{aligned}\nabla_{\mathbf{a}} \varepsilon(\mathbf{a}) &= \nabla_{\mathbf{a}} \left\{ \|\mathbf{y}\|^2 + \mathbf{a}^\top \mathbf{X}^\top \mathbf{X}\mathbf{a} - 2\mathbf{y}^\top \mathbf{X}\mathbf{a} \right\} \\ &= \mathbf{0} + 2\mathbf{X}^\top \mathbf{X}\mathbf{a} - 2\mathbf{X}^\top \mathbf{y} \\ &= 2 \cdot (\mathbf{X}^\top \mathbf{X}\mathbf{a} - \mathbf{X}^\top \mathbf{y}) \\ &\stackrel{!}{=} \mathbf{0}\end{aligned}$$

□

Bemerkung

Das LGS $\mathbf{X}^\top \mathbf{X}\mathbf{a} = \mathbf{X}^\top \mathbf{y}$ heißt System der *Gaußschen Normalengleichungen*. Wir schreiben auch kürzer $\mathbf{R}\mathbf{a} = \mathbf{m}$; dabei ist \mathbf{R} wieder die unzentrierte, unnormierte Kovarianzmatrix der Vektordaten.

Multivariate lineare Regression

Lösen des Systems der Gaußschen Normalengleichungen

Satz

Es seien die Regressionsdaten $(\mathbf{x}_t, y_t) \in \mathbb{R}^N \times \mathbb{R}$, $t = 1, \dots, T$ in der Matrixnotation

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)^\top, \quad \mathbf{y} = (y_1, \dots, y_T)^\top$$

dargestellt, und es sei $h : \mathbf{x} \mapsto \mathbf{a}^\top \mathbf{x}$ linear. Dann lautet der quadratische Regressionsfehler

$$\varepsilon(h) = \varepsilon(\mathbf{a}) = \|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2$$

und wird durch jede Lösung \mathbf{a} der **Gaußschen Normalengleichungen**

$$\mathbf{X}^\top \mathbf{X}\mathbf{a} = \mathbf{X}^\top \mathbf{y}$$

minimiert.

Ausgleichsrechnung und Lineare Gleichungssysteme

Was Sie schon immer über lineare Algebra wissen wollten, aber nie zu fragen wagten

$$\mathbf{X} \cdot \mathbf{a} \stackrel{!}{=} \mathbf{y} \quad \text{mit dem Fehlervektor } \mathbf{e} := \mathbf{X}\mathbf{a} - \mathbf{y}$$

LGS eindeutig

Matrix \mathbf{X} ist quadratisch und vollrangig.

$f : z \mapsto \mathbf{X}z$ ist injektiv
 $\mathbf{X}^\top \mathbf{X}$ ist regulär

$\mathbf{a} = \mathbf{X}^{-1}\mathbf{y}$ ist die eindeutige Lösung mit Fehler $\mathbf{e} = \mathbf{0}$.

Bemerkung

Das Gaußsche Normalengleichungssystem ist entweder eindeutig lösbar oder besitzt unendlich viele Lösungen.

... überbestimmt

Matrix \mathbf{X} hat den vollen Spaltenrang.

$f : z \mapsto \mathbf{X}z$ ist surjektiv
 $\mathbf{X}^\top \mathbf{X}$ ist regulär

$\mathbf{a} = (\mathbf{X}^\top \mathbf{X})^{-1} \cdot \mathbf{X}^\top \mathbf{y}$ ist eine Lösung mit minimalem Gesamtfehler $\|\mathbf{e}\|$.

... unterbestimmt

Matrix \mathbf{X} hat den vollen Zeilenrang.

$f : z \mapsto \mathbf{X}z$ ist surjektiv
 $\mathbf{X}^\top \mathbf{X}$ ist regulär

$\mathbf{a} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \cdot \mathbf{y}$ ist eine Lösung mit Fehler $\mathbf{e} = \mathbf{0}$ und minimaler Länge $\|\mathbf{a}\|$.

System der Gaußschen Normalengleichungen

Linearer Quadratmittelklassifikator ($K = 2$)

$$\mathbf{R} \cdot \mathbf{a} = \mathbf{m}$$

$$\mathbf{R} = \frac{1}{T} \cdot \mathbf{X}^\top \mathbf{X} = \mathbf{S} + \boldsymbol{\mu} \boldsymbol{\mu}^\top$$

$$\mathbf{m} = \frac{1}{T} \cdot \mathbf{X}^\top \mathbf{y} = \frac{1}{T} \cdot \sum_{\omega^+} \mathbf{x}_t = p^+ \cdot \boldsymbol{\mu}^+$$

$p^+ = T^+/T$, $\boldsymbol{\mu}^+$ Positivstatistiken; \mathbf{R} Momentenmatrix der Gesamtprobe.

Linearer Quadratmittelklassifikator ($K > 2$)

$$\mathbf{R} \cdot \mathbf{a}_1 = \mathbf{m}_1$$

$$\vdots = \vdots, \quad \mathbf{m}_\kappa = \frac{1}{T} \cdot \sum_{\omega_\kappa} \mathbf{x}_t = p_\kappa \cdot \boldsymbol{\mu}_\kappa$$

$$\mathbf{R} \cdot \mathbf{a}_K = \mathbf{m}_K$$

Kompaktschreibweise: $\mathbf{R} \cdot \mathbf{A} = \mathbf{M}$ mit $\mathbf{M} = (p_1 \boldsymbol{\mu}_1, \dots, p_K \boldsymbol{\mu}_K)$.

Beweis.

Für eine beliebige Rechteckmatrix mit der SV-Zerlegung $\mathbf{X} = \mathbf{V} \mathbf{D} \mathbf{U}^\top$ heißt die Matrix

$$\mathbf{X}^+ = \mathbf{U} \mathbf{D}^+ \mathbf{V}^\top$$

die **Moore-Penrose-Inverse** oder **Pseudoinverse**. Die Pseudoinverse \mathbf{D}^+ einer Diagonalmatrix \mathbf{D} wiederum enthält auf ihrer Diagonalen die Pseudo-Reziproken:

$$d_n^+ = \begin{cases} 1/d_n & d_n \neq 0 \\ 0 & d_n = 0 \end{cases}, \quad n = 1, \dots, N$$

Diese Pseudoinverse gehorcht der **Moore-Penrose-Gleichung**, denn es gilt:

$$\begin{aligned} \mathbf{X}^\top \mathbf{X} \mathbf{X}^+ &= \mathbf{U} \mathbf{D} \mathbf{V}^\top \cdot \mathbf{V} \mathbf{D} \mathbf{U}^\top \cdot \mathbf{U} \mathbf{D}^+ \mathbf{V}^\top \\ &= \mathbf{U} \mathbf{D}^2 \mathbf{D}^+ \mathbf{V}^\top = \mathbf{U} \mathbf{D} \mathbf{V}^\top = \mathbf{X}^\top \end{aligned}$$

Folglich löst $\mathbf{a}^+ = \mathbf{X}^+ \mathbf{y}$ auch die Gaußschen Normalengleichungen:

$$\mathbf{T} \cdot \mathbf{R} \mathbf{a}^+ = \mathbf{X}^\top \mathbf{X} \cdot \mathbf{X}^+ \mathbf{y} = \mathbf{X}^\top \cdot \mathbf{y} = \mathbf{T} \cdot \mathbf{m}$$

Der Beweis der Minimaleigenschaft erfordert einen Lagrange-Ansatz:

$$\frac{1}{2} \cdot \|\mathbf{a}\|^2 + \lambda \cdot \|\mathbf{X}^\top \mathbf{X} \mathbf{a} - \mathbf{X}^\top \mathbf{y}\|^2 \stackrel{!}{\rightarrow} \text{MIN}$$

□

Minimalnormlösung des GNG-Systems

Lemma

Es sei $\mathbf{R} \cdot \mathbf{a} = \mathbf{m}$ bzw. $\mathbf{X}^\top \mathbf{X} \cdot \mathbf{a} = \mathbf{X}^\top \mathbf{y}$ das GNG-System einer linearen Regressionsaufgabe.

1. Die Matrix \mathbf{R} ist symmetrisch und positiv-semidefinit.
2. Das Gleichungssystem hat stets mindestens eine Lösung.
3. Ist \mathbf{R} invertierbar, so existiert eine eindeutige Lösung:

$$\mathbf{a}^* = \mathbf{R}^{-1} \cdot \mathbf{m}$$

4. Ist \mathbf{X}^+ die Pseudoinverse der Datenmatrix, so löst

$$\mathbf{a}^+ = \mathbf{X}^+ \cdot \mathbf{y}$$

das Gleichungssystem und besitzt unter allen Lösungen die minimale Norm $\|\mathbf{a}^+\|$.

Die Berechnung der Minimalnormlösung ist **nicht praktikabel!**

Gratregularisierung

Lemma

Der regularisierte quadratische Regressionsfehler

$$\varepsilon_\lambda(\mathbf{a}) = \|\mathbf{y} - \mathbf{X} \mathbf{a}\|^2 + \lambda \cdot \|\mathbf{a}\|^2$$

($\lambda > 0$) wird durch die (eindeutige) Lösung

$$\mathbf{a}_\lambda^* = (\mathbf{R}_\lambda)^{-1} \cdot \mathbf{m}, \quad \mathbf{R}_\lambda \stackrel{\text{def}}{=} \mathbf{R} + \lambda \mathbf{E}$$

minimiert.

Beweis.

Der regularisierte Quadratmittelfehler besitzt den Gradientenvektor

$$\nabla_{\mathbf{a}} \varepsilon_\lambda(\mathbf{a}) = \nabla_{\mathbf{a}} \varepsilon(\mathbf{a}) + \lambda \cdot \nabla_{\mathbf{a}} \|\mathbf{a}\|^2 = 2 \cdot (\mathbf{R} \mathbf{a} + \lambda \mathbf{a} - \mathbf{m}) = 2 \cdot (\mathbf{R}_\lambda \mathbf{a} - \mathbf{m})$$

Die Gratregularisierungsmatrix ist für $\lambda \neq 0$ stets invertierbar, denn wegen

$$\mathbf{R}_\lambda = \mathbf{R} + \lambda \mathbf{E} = \mathbf{U} \mathbf{D}^2 \mathbf{U}^\top + \lambda \mathbf{U} \mathbf{E} \mathbf{U}^\top = \mathbf{U} \cdot (\mathbf{D}^2 + \lambda \mathbf{E}) \cdot \mathbf{U}^\top = \mathbf{U} \cdot (\mathbf{D}^2)_\lambda \cdot \mathbf{U}^\top$$

besitzen alle Eigenwerte von \mathbf{R}_λ die Form $d_n^2 + \lambda > 0$.

□

Gewichtete & nichtquadratische Regression

Historische Wurzeln des IRLS: „Iteratively Reweighted Least Squares“

Quadratischer Fehler	$\sum_t (x_t^\top \mathbf{a} - y_t)^2 = \ X\mathbf{a} - \mathbf{y}\ _2^2$
Allgemeiner L_p -Fehler	$\sum_t x_t^\top \mathbf{a} - y_t ^p = \ X\mathbf{a} - \mathbf{y}\ _p^p$
Gewichteter Fehler	$\sum_t w_t^2 \cdot (x_t^\top \mathbf{a} - y_t)^2 = \ W \cdot (X\mathbf{a} - \mathbf{y})\ _2^2$ $W = \text{diag}(w_1, \dots, w_T)$

Gewichtete Ausgleichsrechnung

Wegen $\|W \cdot (X\mathbf{a} - \mathbf{y})\|_2^2 = \|WX\mathbf{a} - Wy\|_2^2 = \|\tilde{X}\mathbf{a} - \tilde{\mathbf{y}}\|_2^2$ lautet der Lösungskoeffizientenvektor $\mathbf{a} = (X^\top W^\top WX)^{-1} \cdot X^\top W^\top Wy$

Ausgleichsrechnung in der L_p -Fehlernorm (Betrag/Minimum)

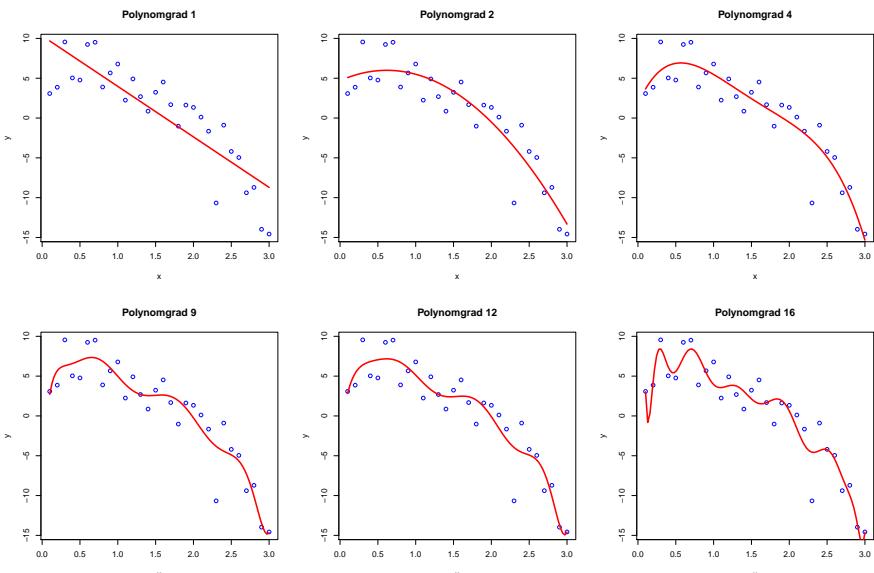
Die Fehlerminimierung kann wegen

$$\|\mathbf{e}\|_p^p = \sum_t |e_t|^p = \sum_t |e_t|^{p-2} |e_t|^2 = \sum_t w_t^2 e_t^2$$

auf IRLS mit Gewichten $w_t = |e_t|^{(p-2)/2}$ zurückgeführt werden.

Überanpassungseffekt bei Ausgleichspolynomen

Weiß verrauschte Daten zur Kurve $y = 7 + 2x - 3x^2$



Lineare versus nichtlineare QMK

Angriffspunkt: 1. Quellvariable 2. Berechnungsweg 3. Zielvariable

Termexpansion

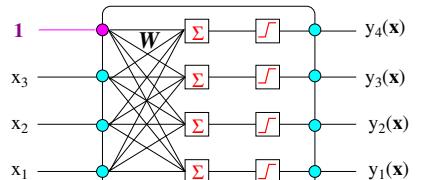
GNNS für alle Koeffizienten

- Linear & affin $O(N^1)$
- Quadratisch $O(N^2)$
- Kubisch $O(N^3)$
- Polynomansatz $O((N+p) \choose p)$

Neuronale Berechnungsmodelle

Error Backpropagation

- Mehrschichtenperzeptron
- Radiale Basisfunktionen
- Time-Delay Neural Network



Nominale Attribute

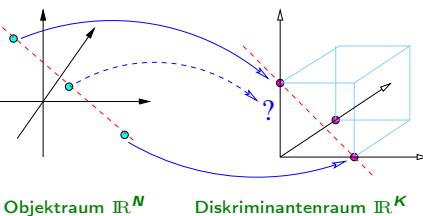
Kontrastmatrizen ($L_n - 1$)

- ohne Interaktionsterme $O(\ell)$, $\ell = \sum_n L_n$
- einfache Interaktionsterme $O(\ell^2)$

Gelenkfunktion $\phi(y) = x^\top a$
Generalized Linear Model

Maskierungseffekt

Lineare Quadratmitteldiskriminanten in Mehrklassensituationen



Quadratmittel

Minimiere den quadratischen Vorhersagefehler

$$\sum_t (y_t - a^\top x_t)^2$$

- ⊖ negative Werte!
- ⊖ nicht normiert!

Kollineare Klassenzentren

$\mu_\kappa = \beta_\kappa a + b$
↔ kollineare Diskriminatenvektoren
 $h(\mu_\kappa) = \beta_\kappa \tilde{a} + \tilde{b}$
(ideal = κ -te Einheitsvektoren)

Probit

Maximiere Wahrsch'keitssumme

$$\prod_t P(y_t | x_t)$$

$$\sum_t P(y_t | x_t)$$

Logit

Maximiere Datenwahrsch'keit

mit Posterior-Wahrscheinlichkeiten der logistischen Form $P(\Omega_1 | x) \propto e^{a^\top x}$

Lineare logistische Regression

Versionenräume

Naive Bayesregel

Multivariate lineare Regression

Logistische Regression

Ordinale Regression und Präferenzmodelle

Statistische Entscheidungsbäume

Zusammenfassung

Zweiklassenmodell

Lineares Vorhersagemodell für die **log-odds**

$$\log \frac{P(\Omega_1 | \mathbf{x})}{P(\Omega_0 | \mathbf{x})} \stackrel{!}{=} h(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$$

Konsistente W'keiten

Alle $P(\Omega_\lambda | \mathbf{x}) \in [0, 1]$.

Alle Odds $\in [0, +\infty]$.

Log-odds $\in [-\infty, +\infty]$.

Umkehrformeln

Alle Klassen $\lambda \neq K$

$$p_\lambda(\mathbf{x}) = \frac{\exp(\mathbf{a}_\lambda^\top \mathbf{x})}{1 + \sum_{\kappa \neq K} \exp(\mathbf{a}_\kappa^\top \mathbf{x})}$$

Referenzklasse $\lambda = K$

$$p_\lambda(\mathbf{x}) = \frac{1}{1 + \sum_{\kappa \neq K} \exp(\mathbf{a}_\kappa^\top \mathbf{x})}$$

bzw. $p_1(\mathbf{x}) =$
 $1 - p_0(\mathbf{x}) = \frac{e^{\mathbf{a}^\top \mathbf{x}}}{(1+e^{\mathbf{a}^\top \mathbf{x}})}.$

Mehrklassenmodell

$K - 1$ Modelle für logarithmierte Kontrastwahrscheinlichkeiten

$$\log \frac{P(\Omega_\lambda | \mathbf{x})}{P(\Omega_K | \mathbf{x})} \stackrel{!}{=} h_\lambda(\mathbf{x}) = \mathbf{a}_\lambda^\top \mathbf{x}$$

für alle $1 \leq \lambda < K$.

Maximum-Likelihood-Schätzung

Vereinfachter Fall: $K = 2$

Lemma

Für das binäre logistische Modell $p_1(\mathbf{x}) \propto \exp(\mathbf{a}^\top \mathbf{x})$ mit den Lerndaten $(\mathbf{x}_t, y_t) \in \mathbb{R}^N \times \{1, 0\}$, $t = 1, \dots, T$ gilt:

1. Die ML-Zielgröße besitzt die Darstellung

$$\ell(\mathbf{a}) = \log \prod_{t=1}^T p_{y_t}(\mathbf{x}_t) = \sum_{t=1}^T \left\{ y_t \cdot \mathbf{a}^\top \mathbf{x}_t - \log \left[1 + \exp(\mathbf{a}^\top \mathbf{x}_t) \right] \right\} .$$

2. Für ihren Gradientenvektor der partiellen Ableitungen gilt:

$$\nabla_{\mathbf{a}} = \frac{\partial \ell(\mathbf{a})}{\partial \mathbf{a}} = \sum_{t=1}^T \mathbf{x}_t \cdot (y_t - p_1(\mathbf{x}_t)) = \mathbf{X}^\top \cdot (\mathbf{y} - \mathbf{p})$$

3. Für ihre Hessematrix der gemischten partiellen Ableitungen gilt

$$\mathbf{H}_{\mathbf{a}} = \frac{\partial^2 \ell(\mathbf{a})}{\partial \mathbf{a} \partial \mathbf{a}^\top} = - \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \cdot p_1(\mathbf{x}_t) \cdot (1 - p_1(\mathbf{x}_t)) = -\mathbf{X}^\top \mathbf{W} \mathbf{X} ,$$

wobei $\mathbf{W} = \text{diag}(w_1, \dots, w_T)$ und $w_t = p_1(\mathbf{x}_t) \cdot (1 - p_1(\mathbf{x}_t))$ bezeichnet.

Der IRLS-Algorithmus

„Iteratively Reweighted Least Squares“

(Algorithmus)

1 INITIALISIERUNG $\mathbf{a} \leftarrow \mathbf{0}$

2 NEWTON-RAPHSON-SCHRITT

$$\mathbf{a} \leftarrow \mathbf{a} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \cdot \mathbf{X}^\top \cdot (\mathbf{y} - \mathbf{p})$$

Die diagonale Skalierungsmatrix $\mathbf{W} \in \mathbb{R}^{T \times T}$ hat Einträge $w_{tt} = p_t \cdot (1 - p_t)$, $p_t = \hat{p}_1(\mathbf{x}_t)$.

3 TERMINIERUNG

Prüfe Abbruchbedingung; gehe \rightsquigarrow 2 oder ENDE.

(zum Abschluss)

Newton-Raphson-Optimierungsschritt

Gradientenaufstieg mit quadratisch berechneter Schrittweite

$$\mathbf{a} \leftarrow \mathbf{a} - \mathbf{H}_{\mathbf{a}}^{-1} \cdot \nabla_{\mathbf{a}}$$

IRLS und Regularisierung

Was heißt eigentlich „wiederholte Neugewichtung“?

Newtonsschritt = gewichtete lineare Regression

$$\begin{aligned}\mathbf{a}^* &\leftarrow \mathbf{a} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \cdot \mathbf{X}^\top \cdot (\mathbf{y} - \mathbf{p}) \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \cdot \mathbf{X}^\top \mathbf{W}^{1/2} \cdot \mathbf{W}^{1/2} \cdot \underbrace{(\mathbf{X}\mathbf{a} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p}))}_z\end{aligned}$$

\mathbf{a}^* ist Lösung der gewichteten Regressionsaufgabe $\|\mathbf{W}^{1/2} \cdot (z - \mathbf{X}\mathbf{a})\|^2 \rightarrow \text{MIN}$

WLS-Regularisierung in jedem Newtonsschritt

Löse gewichtetes GNG-System $\mathbf{R}_W \mathbf{a} = \mathbf{m}_W$ mit $\mathbf{R}_W = \mathbf{X}^\top \mathbf{W} \mathbf{X}$ und $\mathbf{m}_W = \mathbf{X}^\top \mathbf{W} \mathbf{z}$ mittels regularisierter Koeffizientenmatrix:

$$\mathbf{R}_{W,\lambda} \stackrel{\text{def}}{=} (\mathbf{R}_W)_\lambda = \mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \cdot \mathbf{E}$$

Das Probit-Modell ($K = 2$)

Logistisches Wahrscheinlichkeitsmodell
mit einer additiven Zielfunktion

$$p_y(\mathbf{x}) = \frac{e^{\mathbf{y} \cdot \mathbf{a}^\top \mathbf{x}}}{1 + e^{\mathbf{a}^\top \mathbf{x}}}, \quad \ell(\mathbf{a}) = \sum_{t=1}^T p_{y_t}(\mathbf{x}_t) \rightarrow \text{MAX}$$

Gradientenvektor

$$\nabla_{\mathbf{a}} \ell(\mathbf{a}) = \sum_{t=1}^T p_{y_t}(\mathbf{x}_t) \cdot \{y_t - p_1(\mathbf{x}_t)\} \cdot \mathbf{x}_t = \mathbf{X}^\top \mathbf{Q}(\mathbf{y} - \mathbf{p})$$

mit $\mathbf{Q} = \text{diag}(\{p_{y_t}(\mathbf{x}_t)\}_t)$

Hessematrix

$$\mathbf{H}_{\mathbf{a}} = \sum_{t=1}^T p_{y_t}(\mathbf{x}_t) \cdot \{(y_t - p_1(\mathbf{x}_t))^2 + p_1^2(\mathbf{x}_t) - p_1(\mathbf{x}_t)\} \cdot \mathbf{x}_t \mathbf{x}_t^\top = -\mathbf{X}^\top \mathbf{W} \mathbf{P} \mathbf{X}$$

mit $\mathbf{W} = \text{diag}(\{p_1(\mathbf{x}_t) \cdot p_0(\mathbf{x}_t)\}_t)$ und $\mathbf{P} = \text{diag}(\{p_{y_t}(\mathbf{x}_t) - p_{1-y_t}(\mathbf{x}_t)\}_t)$

Maximum-Likelihood-Schätzung

Allgemeiner Fall: $K \geq 2$

Lemma

Für das logistische Modell $p_\lambda(\mathbf{x}) \propto \exp(\mathbf{a}_\lambda^\top \mathbf{x})$ mit den Lerndaten $(\mathbf{x}_t, g_t) \in \mathbb{R}^N \times \{1, \dots, K\}$, $t = 1, \dots, T$ gilt:

1. Die ML-Zielgröße besitzt die Darstellung

$$\ell(\mathbf{A}) = \log \prod_{t=1}^T p_{g_t}(\mathbf{x}_t) = \sum_{t=1}^T \left\{ \mathbf{a}_t^\top \mathbf{x}_t - \log \sum_{\nu} e^{\mathbf{a}_\nu^\top \mathbf{x}_t} \right\}, \quad \mathbf{a}_K = \mathbf{0} \in \mathbb{R}^N.$$

2. Für die $K \cdot N$ partiellen Ableitungen ihrer Gradientenmatrix gilt:

$$\frac{\partial \ell(\mathbf{A})}{\partial \mathbf{a}_{\lambda,i}} = \sum_{t=1}^T \mathbf{x}_{t,i} \cdot (y_{t,\lambda} - p_\lambda(\mathbf{x}_t))$$

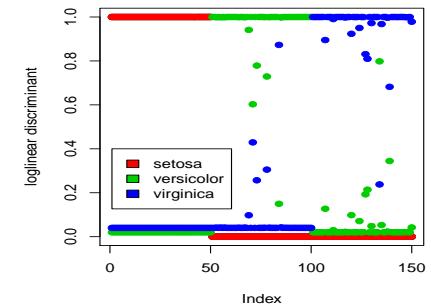
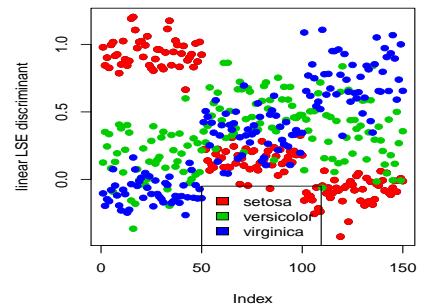
3. Für die $K^2 \cdot N^2$ gemischten partiellen Ableitungen ihres Hessetensors gilt:

$$\frac{\partial \ell^2(\mathbf{A})}{\partial \mathbf{a}_{\lambda,i} \cdot \partial \mathbf{a}_{\kappa,j}} = \sum_{t=1}^T \mathbf{x}_{t,i} \mathbf{x}_{t,j} \cdot (p_\lambda(\mathbf{x}_t) \cdot p_\kappa(\mathbf{x}_t) - \delta_{\lambda,\kappa} \cdot p_\lambda(\mathbf{x}_t))$$

Es bezeichne $y_{t,\lambda} = \delta_{g_t,\lambda}$ die Klassenindikatorfunktion der Lerndaten.

Reklassifikationsexperiment — Irisblüten-Datensatz

3 Klassen · 4 numerische Attribute · 50+50+50 Objekte



Quadratmittelmodell

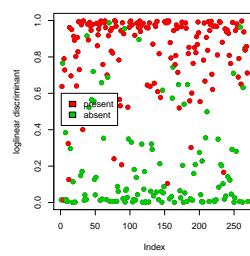
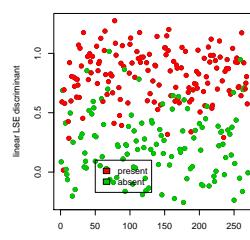
3 affine Prädiktoren $\mathbf{a}_\lambda \in \mathbb{R}^5$
Starke Schwankung um 1 und 0
Vertauschungen {‘versicolor’, ‘virginica’}

Loglinearmodell

2 affine Prädiktoren $\mathbf{a}_\lambda \in \mathbb{R}^5$
Fast alle Wahrsch'keiten bei {0, 1}
Fast perfekte Klassenidentifikation

Reklassifikationsexperiment — Herzkrankheiten-Datensatz

2 Klassen · 13 diskrete & numerische Attribute · 270 Objekte



Auszug Datenfriedhof

70.0	1.0	4.0	130.0	322.0	0.0	2.0	109.0	0.0	2.4	2.0	3.0	3.0	2
67.0	0.0	3.0	115.0	564.0	0.0	2.0	160.0	0.0	1.6	2.0	0.0	7.0	1
57.0	1.0	2.0	124.0	261.0	0.0	0.0	141.0	0.0	0.3	1.0	0.0	7.0	2
64.0	1.0	4.0	128.0	263.0	0.0	0.0	105.0	1.0	0.2	2.0	1.0	7.0	1
74.0	0.0	2.0	120.0	269.0	0.0	2.0	121.0	1.0	0.2	1.0	1.0	3.0	1
65.0	1.0	4.0	120.0	177.0	0.0	0.0	140.0	0.0	0.4	1.0	0.0	7.0	1
56.0	1.0	3.0	130.0	256.0	1.0	2.0	142.0	1.0	0.6	2.0	1.0	6.0	2
...

Attribute, Skalen, Werte

1. age (IR)	-0.02511018
2. sex {male, female}	1.89901910
3. chest pain {A, B, C, D}	1.741, 0.784, 2.748
4. blood pressure (IR)	0.03110868
5. serum cholesterol (IR)	0.00655756
6. fasting blood sugar {T, F}	-0.37604461
...	...
13 thal {normal, fixed, defect}	-0.318, 1.468
intercept	-7.68704469

Darstellungssatz für QM-Lösungen

Endlichdimensionaler Spezialfall des Satzes von Kimeldorf & Wahba (1971)

Satz

Die regularisierten (unregularisierten) und gewichteten (ungewichteten) Quadratmittelaufgaben mit den Normalengleichungen

$$\mathbf{R} \cdot \mathbf{a} = \mathbf{m}$$

(LS)

$$\mathbf{R}_\lambda \cdot \mathbf{a} = \mathbf{m}$$

(RLS)

$$\mathbf{R}_w \cdot \mathbf{a} = \mathbf{m}_w$$

(WLS)

$$\mathbf{R}_{w,\lambda} \cdot \mathbf{a} = \mathbf{m}_w$$

(RWLS)

besitzen jeweils mindestens eine Lösung, die sogar als Linearkombination der Datenvektoren $\mathbf{x}_1, \dots, \mathbf{x}_T$ darstellbar ist, d.h. es gilt:

$$\mathbf{a}^* \in \text{Lin}(\mathbf{X})$$

Bezeichnungen

für die nicht normierten und unzentrierten Momente:

$$\mathbf{m} = \mathbf{X}^\top \mathbf{y}$$

$$\mathbf{R} = \mathbf{X}^\top \mathbf{X}$$

$$\mathbf{R}_\lambda = \mathbf{R} + \lambda \mathbf{E}$$

$$\mathbf{m}_w = \mathbf{X}^\top \mathbf{Wz}$$

$$\mathbf{R}_w = \mathbf{X}^\top \mathbf{WX}$$

$$\mathbf{R}_{w,\lambda} = \mathbf{R}_w + \lambda \mathbf{E}$$

Dualisierung der Regressionsaufgabe

Der Schlüssel zum Kerneltrick für Quadratmittel- & logistische Prädiktoren

Definition

Für eine Datenmatrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)^\top$ des \mathbb{R}^N bezeichne $\text{Lin}(\mathbf{X})$ die **lineare Hülle** der Vektoren und $\text{Lin}(\mathbf{X}^\perp)$ ihren **Orthogonalraum**.

Lineare Hülle

Die Menge aller Linearkombinationen der Matrixzeilen \mathbf{x}_t ; sie bildet den kleinsten Untervektorraum von \mathbb{R}^N , der alle \mathbf{x}_t enthält.

$$\text{Lin}(\mathbf{X}) = \{\mathbf{X}^\top \mathbf{a} \mid \mathbf{a} \in \mathbb{R}^T\}$$

Orthogonalraum

Die Menge aller Vektoren, die auf allen Matrixzeilen \mathbf{x}_t senkrecht stehen; sie bildet den größten Untervektorraum von \mathbb{R}^N , der keines der \mathbf{x}_t enthält.

$$\text{Lin}(\mathbf{X}^\perp) = \{\mathbf{z} \in \mathbb{R}^N \mid \mathbf{Xz} = \mathbf{0}\}$$

Lemma

Lineare Hülle und Orthogonalraum spannen stets den Gesamtraum auf:

$$\mathbb{R}^N = \text{Lin}(\mathbf{X}) \oplus \text{Lin}(\mathbf{X}^\perp)$$

Beweis.

- REPRÄSENTATION FÜR LS-LÖSUNG

Ist $\mathbf{a} = \mathbf{a}_0 + \mathbf{a}_\perp$ mit $\mathbf{a}_0 \in \text{Lin}(\mathbf{X})$ und $\mathbf{a}_\perp \in \text{Lin}(\mathbf{X}^\perp)$ eine Lösung der GNG $\mathbf{Ra} = \mathbf{m}$, so gilt:

$$\mathbf{m} = \mathbf{Ra} = \mathbf{X}^\top \mathbf{Xa}_0 + \mathbf{X}^\top \mathbf{Xa}_\perp = \mathbf{Ra}_0$$

Wir können folglich auch eine Lösung in $\text{Lin}(\mathbf{X})$ finden.

- REPRÄSENTATION FÜR RLS-LÖSUNG

Ist $\mathbf{a} = \mathbf{a}_0 + \mathbf{a}_\perp$ eine Lösung der GNG $\mathbf{R}_\lambda \mathbf{a} = \mathbf{m}$, so gilt:

$$\mathbf{m} = \mathbf{R}_\lambda \mathbf{a} = \mathbf{X}^\top \mathbf{Xa} + \lambda \mathbf{a}_0 + \lambda \mathbf{a}_\perp$$

Da sowohl $\mathbf{m} = \mathbf{X}^\top \mathbf{y}$ als auch $\mathbf{X}^\top \mathbf{Xa}$ und $\lambda \mathbf{a}_0$ offensichtlich aus $\text{Lin}(\mathbf{X})$ sind, ist das auch für den verbleibenden Ausdruck $\lambda \mathbf{a}_\perp$ der Fall. Wegen $\lambda > 0$ folgt $\mathbf{a}_\perp = \mathbf{0}$, also ist $\mathbf{a} = \mathbf{a}_0$ zwingend aus der linearen Hülle von \mathbf{X} .

- REPRÄSENTATION FÜR WLS-LÖSUNG

Im IRLS-Schritt sei $\mathbf{a} = \mathbf{a}_0 + \mathbf{a}_\perp$ eine Lösung der GNG $\mathbf{R}_w \mathbf{a} = \mathbf{m}_w$. Dann gilt:

$$\mathbf{m}_w = \mathbf{R}_w \mathbf{a} = \mathbf{X}^\top \mathbf{WX} \cdot \mathbf{a}_0 + \mathbf{X}^\top \mathbf{WX} \cdot \mathbf{a}_\perp = \mathbf{R}_w \mathbf{a}_0$$

- REPRÄSENTATION FÜR RWLS-LÖSUNG

Für die Lösung $\mathbf{a} = \mathbf{a}_0 + \mathbf{a}_\perp$ im regularisierten IRLS-Schritt gilt wie bei RLS:

$$\mathbf{m}_w = \mathbf{R}_{w,\lambda} \mathbf{a} = (\mathbf{X}^\top \mathbf{WX} + \lambda \mathbf{E}) \cdot (\mathbf{a}_0 + \mathbf{a}_\perp) = \mathbf{X}^\top \mathbf{WXa} + \lambda \mathbf{a}_0 + \lambda \mathbf{a}_\perp$$

LS⁺ — die dualisierte Quadratmittelaufgabe

Speicher- und Rechenaufwand $O(T^2)$ und $O(T^3)$

Duale Lösungsdarstellung

als Linearkombination der Objektvektoren:

$$\mathbf{a} = \mathbf{X}^\top \mathbf{b} = \sum_{t=1}^T b_t \cdot \mathbf{x}_t, \quad \mathbf{b} \in \mathbb{R}^T$$

Duale Regressionsfehlerformel

in Abhängigkeit vom Vektor \mathbf{b} der Lösungskoeffizienten:

$$\varepsilon(\mathbf{b}) = \|\mathbf{y} - \mathbf{X} \cdot \mathbf{X}^\top \mathbf{b}\|^2 \rightarrow \text{MIN}$$

Duale Gauß'sche Normalengleichungen

Lineares Gleichungssystem (Dimension $T \times T$) mit Gram'scher Matrix:

$$\mathbf{G}^2 \cdot \mathbf{b} = \mathbf{G} \cdot \mathbf{y}, \quad \mathbf{G} = \mathbf{X} \cdot \mathbf{X}^\top$$

Beweis.

- UNREGULARISIERTE LÖSUNG:
Der Gradientenvektor der Zielgröße

$$\varepsilon(\mathbf{b}) = \|\mathbf{y} - \mathbf{G}\mathbf{b}\|^2 = \mathbf{y}^\top \mathbf{y} - 2 \cdot \mathbf{b}^\top \mathbf{G}\mathbf{y} + \mathbf{b}^\top \mathbf{G}^2 \mathbf{b}$$

lautet

$$\nabla_{\mathbf{b}} \varepsilon(\mathbf{b}) = \mathbf{0} - 2 \cdot \mathbf{G}\mathbf{y} + 2 \cdot \mathbf{G}^2 \mathbf{b}.$$

Nullsetzen ergibt die GNG. Unter der Annahme einer regulären Gramschen Matrix ergibt sich die Lösung durch Multiplikation beider Gleichungsseiten mit \mathbf{G}^{-2} .

- REGULARISIERTE LÖSUNG I:
Wir regularisieren im Vektorraum \mathbb{R}^N ; der Fehlerterm besitzt den Gradientenvektor

$$\nabla_{\mathbf{b}} \varepsilon_\lambda(\mathbf{b}) = -2\mathbf{G}\mathbf{y} + 2\mathbf{G}^2 \mathbf{b} + 2\lambda \cdot \mathbf{G}\mathbf{b} = -2\mathbf{G} \cdot (\mathbf{y} - (\mathbf{G} + \lambda\mathbf{E}) \cdot \mathbf{b})$$

Da \mathbf{G}_λ regulär ist für $\lambda > 0$ liefert $\mathbf{b} = \mathbf{G}_\lambda^{-1} \mathbf{y}$ eine Lösung.

- REGULARISIERTE LÖSUNG II:
Wir regularisieren im Vektorraum \mathbb{R}^T ; der Fehlerterm besitzt den Gradientenvektor

$$\nabla_{\mathbf{b}} \varepsilon_\lambda(\mathbf{b}) = -2\mathbf{G}\mathbf{y} + 2\mathbf{G}^2 \mathbf{b} + 2\lambda \cdot \mathbf{b} = -2 \cdot (\mathbf{G}\mathbf{y} - (\mathbf{G}^2 + \lambda\mathbf{E}) \cdot \mathbf{b})$$

Da auch $(\mathbf{G}^2)_\lambda$ regulär ist für $\lambda > 0$ liefert $\mathbf{b} = (\mathbf{G}^2)_\lambda^{-1} \cdot \mathbf{G}\mathbf{y}$ eine Lösung.



Regularisierung dualisierter QM-Aufgaben

Ungewichteter und gewichteter Fall

Lemma

Die Lösungen der dualisierten **LS-Aufgabe** lauten je nach Regularisierungstechnik:

$$\begin{aligned} \mathbf{b}^* &= \mathbf{G}^{-1} \cdot \mathbf{y} & \varepsilon(\mathbf{b}) &= \|\mathbf{y} - \mathbf{G}\mathbf{b}\|^2 \\ \mathbf{b}^* &= (\mathbf{G} + \lambda\mathbf{E})^{-1} \cdot \mathbf{y} & \varepsilon_\lambda(\mathbf{b}) &= \|\mathbf{y} - \mathbf{G}\mathbf{b}\|^2 + \lambda \cdot \|\mathbf{X}^\top \mathbf{b}\|^2 \\ \mathbf{b}^* &= (\mathbf{G}^2 + \lambda\mathbf{E})^{-1} \cdot \mathbf{G}\mathbf{y} & \varepsilon'_\lambda(\mathbf{b}) &= \|\mathbf{y} - \mathbf{G}\mathbf{b}\|^2 + \lambda \cdot \|\mathbf{b}\|^2 \end{aligned}$$

Lemma

Die Lösungen der dualisierten **WLS-Aufgabe** lauten je nach Regularisierungstechnik:

$$\begin{aligned} \mathbf{b}^* &= \mathbf{G}^{-1} \cdot \mathbf{z} & \varepsilon_W(\mathbf{b}) &= \|\mathbf{z} - \mathbf{G}\mathbf{b}\|_W^2 \\ \mathbf{b}^* &= (\mathbf{W}\mathbf{G} + \lambda\mathbf{E})^{-1} \cdot \mathbf{W}\mathbf{z} & \varepsilon_{W,\lambda}(\mathbf{b}) &= \|\mathbf{z} - \mathbf{G}\mathbf{b}\|_W^2 + \lambda \cdot \|\mathbf{X}^\top \mathbf{b}\|^2 \\ \mathbf{b}^* &= (\mathbf{G}\mathbf{W}\mathbf{G} + \lambda\mathbf{E})^{-1} \cdot \mathbf{G}\mathbf{W}\mathbf{z} & \varepsilon'_{W,\lambda}(\mathbf{b}) &= \|\mathbf{z} - \mathbf{G}\mathbf{b}\|_W^2 + \lambda \cdot \|\mathbf{b}\|^2 \end{aligned}$$

Beweis.

Das zweite Lemma dient der schrittweisen Berechnung und Regularisierung im IRLS-Algorithmus für loglineare Modelle.

An Stelle des Fehlerfunktional $\|\mathbf{y} - \mathbf{G}\mathbf{b}\|^2$ wird

$$\|\mathbf{z} - \mathbf{G}\mathbf{b}\|_W^2 \stackrel{\text{def}}{=} (\mathbf{z} - \mathbf{G}\mathbf{b})^\top \cdot \mathbf{W} \cdot (\mathbf{z} - \mathbf{G}\mathbf{b})$$

minimiert. Wir unterscheiden wieder zwischen der Regularisierung im Raum \mathbb{R}^N und im Raum \mathbb{R}^T .

- Ist \mathbf{G} invertierbar, so hängt die Lösung $\mathbf{b}^* = \mathbf{G}^{-1} \mathbf{b}$ nicht von der (diagonalen) Gewichtsmatrix \mathbf{W} ab, denn $\mathbf{z} \approx \mathbf{G}\mathbf{b}$ wird ja mit exakter Gleichheit erfüllt:

$$\nabla \varepsilon_W(\mathbf{b}) = -2\mathbf{G}\mathbf{W}\mathbf{z} + 2\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{b} = -2\mathbf{G}\mathbf{W} \cdot (\mathbf{z} - \mathbf{G}\mathbf{b})$$

- Bei Regularisierung im Raum \mathbb{R}^N ergibt sich:

$$\nabla \varepsilon_{W,\lambda}(\mathbf{b}) = -2\mathbf{G}\mathbf{W}\mathbf{z} + 2\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{b} + 2\lambda\mathbf{G}\mathbf{b} = -2\mathbf{G} \cdot (\mathbf{W}\mathbf{z} - (\mathbf{W}\mathbf{G})_\lambda \cdot \mathbf{b})$$

- Bei Regularisierung im Raum \mathbb{R}^T ergibt sich:

$$\nabla \varepsilon'_{W,\lambda}(\mathbf{b}) = -2\mathbf{G}\mathbf{W}\mathbf{z} + 2\mathbf{G}\mathbf{W}\mathbf{G}\mathbf{b} + 2\lambda\mathbf{b} = -2 \cdot (\mathbf{G}\mathbf{W}\mathbf{z} - (\mathbf{G}\mathbf{W}\mathbf{G})_\lambda \cdot \mathbf{b})$$



Kombinatorische Regression

Aufgabenstellung

Klassifikation von Texten

$v \in \Omega = \mathcal{V}^*$ über Wortschatz

\mathcal{V}

Termexpansion

Binärattribute:

Wort- m -Tupel oder

Wort- m -Subsets

$$\phi : \Omega \rightarrow \{0, 1\}^{\mathcal{V}^m}$$

$$\text{mit } \phi_u(x) = \begin{cases} 1 & u \in x \\ 0 & u \notin x \end{cases}$$

Loglinearmodell

der Dimension L^m bzw. $\binom{L}{m}$

Duales Loglinearmodell

Gramsche T^2 -Matrix mit Einträgen

$$K(x_s, x_t) = \langle \phi(x_s), \phi(x_t) \rangle$$

Kombinat. Kernoperator

$$\begin{aligned} K(x, y) &= \sum_{u \in \mathcal{V}^m} \phi_u(x) \cdot \phi_u(y) \\ &= \begin{cases} |V_x^m \cap V_y^m| & \text{Tupel} \\ \binom{|V_x^m \cap V_y^m|}{m} & \text{Subsets} \end{cases} \end{aligned}$$

Die Zählaufgaben $|V_x^m \cap V_y^m|$ sind sehr effizient zu bewältigen.

Versionenräume

Naive Bayesregel

Multivariate lineare Regression

Logistische Regression

Ordinal Regression und Präferenzmodelle

Statistische Entscheidungsbäume

Zusammenfassung

Ordinal Regression

Reelle Quellattribute $\mathbb{X}_1, \dots, \mathbb{X}_N \Rightarrow$ geordnetes Zielattribut $\mathbb{Y} \in \{1, \dots, L\}$

Nominales Attribut

A posteriori Verteilung

$$p_\ell(x) \stackrel{\text{def}}{=} P(\mathbb{Y} = \ell \mid \mathbb{X} = x)$$

Normierungsbedingung:

$$\sum_\ell p_\ell(x) = 1$$

Nomiale Beispiele

RedGreenBlue-Skala:

$$p = (\frac{1}{2}, \frac{1}{6}, \frac{1}{3})$$

Unfaire Würfel:

$$p = (0, 0, 0, \frac{1}{4}, \frac{1}{4}, \frac{1}{2})$$

Müssen ordinale Verteilungen zwangsläufig „unimodal“ sein?

Ordinales Attribut

Kumulative a post. Verteilung

$$q_\ell(x) \stackrel{\text{def}}{=} P(\mathbb{Y} \leq \ell \mid \mathbb{X} = x)$$

Skalenbindung:

$$x \rightsquigarrow z(x) \in J_\ell \subset \mathbb{R}$$

Ordinal Beispiele

HighMediumLow-Skala:

$$p = (\frac{1}{2}, \frac{1}{6}, \frac{1}{3})$$

Zensuren-Skala:

$$p = (\frac{3}{7}, \frac{1}{7}, 0, 0, \frac{3}{7}, 0)$$

Postulat der verborgenen dichten Qualitätsskala

„Cumulative link model“ — Agresti 2002

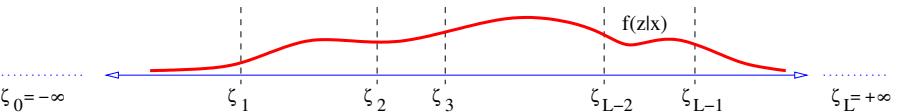
Kumulatives Gelenkfunktionsmodell

Latente Variable \mathbb{Z} auf der Skala $-\infty = \zeta_0 < \zeta_1 < \dots < \zeta_L = +\infty$ mit

$$\mathbb{Y} = \ell \Leftrightarrow \mathbb{Z} \in (\zeta_{\ell-1}, \zeta_\ell] \quad \text{und} \quad \mathbb{Z} \sim f(\mu = h(x), \sigma^2 = 1)$$

$h(\cdot)$ Gelenkfunktion, $f(\cdot)$ Verteilungsgesetz.

$$\left\{ \begin{array}{l} \text{lognormal} \\ \text{normal} \\ \text{Extremwert} \\ \text{Cauchy} \end{array} \right\}$$



Bemerkung

Es gilt $q_\ell(x) = P(\mathbb{Y} \leq \ell \mid \mathbb{X} = x) = P(\mathbb{Z} \leq \zeta_\ell \mid \mathbb{X} = x) = F(\zeta_\ell - h(x))$.

Proportional Odds Linear Regression

Lineares Binomialmodell für die Gelenkfunktion

POLR-Modell

Lineare Vorhersage der logarithmierten Chancenfunktionen:

$$\text{log odds}_\ell(x) \stackrel{\text{def}}{=} \log \frac{q_\ell(x)}{1 - q_\ell(x)} = \log \frac{P(\mathbb{Y} \leq \ell | \mathbb{X} = x)}{P(\mathbb{Y} > \ell | \mathbb{X} = x)} = \mathbf{a}^\top \mathbf{x} + \zeta_\ell$$

Bemerkungen

1. Normierung

$$\sum_\ell p_\ell(x) = \sum_\ell (q_\ell(x) - q_{\ell-1}(x)) = q_L(x) - q_0(x) = 1 - 0$$

2. Monotonie

$$k \leq \ell \Rightarrow \zeta_k \leq \zeta_\ell \Rightarrow \text{logz}_k(x) \leq \text{logz}_\ell(x) \Rightarrow q_k(x) \leq q_\ell(x)$$

3. Proportionale Chancen

$$\log \frac{\text{odds}_\ell(x)}{\text{odds}_\ell(x')} = \mathbf{a}^\top (\mathbf{x} - \mathbf{x}') \quad \text{ist unabhängig von } \ell$$

Beispiel — Präsidentschaftswahlen USA'96

<http://www.stat.washington.edu/quinn/classes/536/data/nes96r.dat>

Datensatz

944 Versuchspersonen

11 Attribute, u.a.:

↳ Pol/ID Clinton*

↳ Alter

Bildungsgrad

Einkommen

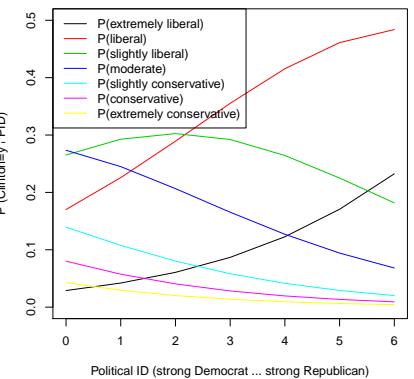
Stimme für ...

TV-News/Woche

Pol/ID selbst*

Pol/ID Dole*

*) Pol/IDs in 7 Stufen



POLR-Datenanalyse

Zielattribut $\hat{=}$ Einschätzung von Clintons politischer Haltung

Quellattribut $\hat{=}$ politische Selbsteinschätzung

Fixiert: 3 TV/Woche, 44 Jahre, 12 Schuljahre, 35–40 Kilotonnen

Lernen von Präferenzrelationen

Objektive Präferenz

Aus einer Serie gewonnener, verlorener oder unentschiedener Partien $(x_t, y_t) \in \Omega \times \Omega$ ist eine passende Qualitätsrelation (Ω, \preceq) zu lernen.



„Tourniermetapher“

Subjektive Präferenz

Aus einer Serie persönlicher Nennungen, Wertungen oder Reihungen $(s_t, x_t) \in \mathcal{S} \times \Omega$ ist eine Schar passender Qualitätsrelationen $(\Omega, \preceq_s)_{s \in \mathcal{S}}$ zu lernen.



„Jurorenmetapher“

Geschlossene Welten

Objektraum Ω und/oder Subjektraum \mathcal{S} bilden ein endliches Inventar.
(Nominalattribut)

Offene Welten

Objekte u/o Subjekte sind durch ihre Eigenschaften charakterisiert.

(Attributvektoren)

Objektive Präferenz durch logistische Regression

Bilaterales Ereignismodell

$\mathbb{X} \hat{=}$ Objekt #1 (Herausforderer)

$\mathbb{Y} \hat{=}$ Objekt #2 (Gegner)

$\mathbb{Z} \hat{=}$ Resultat \pm „Sieg“ oder \pm „Tor“ ...

Logistisch-lineares Erfolgsmodell

$$\text{log odds}(x, y) = \underbrace{\mathbf{a}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{y} + \zeta}_{g(x) - h(y)}$$

Präferenzinterpretation

x hat immer dann bessere Gewinnchancen als y wenn $g(x) > h(y)$ gilt.

Intervalordnung ?

Es gilt $p(x, x) \leq \frac{1}{2} \Leftrightarrow g_x \leq h_x$.

$$x \succ y \quad \text{gdw.} \quad [g_x, h_x] \sqsupseteq [g_y, h_y]$$

Fußballturnier

GER : BRA 3:1

USA : LBY 0:1

UK : IRAN 2:2

Punktestandbezogen

XGER XBRA +

XBRA XGER -

XUSA XLBY -

XLBY XUSA +

XUK XIRAN -

XIRAN XUK -

Torstandbezogen

XGER XBRA + 3

XGER XBRA - 87

XBRA XGER + 1

XBRA XGER - 89

... ...

XUK XIRAN + 2

XUK XIRAN - 88

XIRAN XUK + 2

XIRAN XUK - 88

Objektive Präferenz durch Proportional-Odds Regression

Trilaterales Ereignismodell

$\mathbb{X} \hat{=} \text{Objekt } \#1 \text{ (Herausforderer)}$

$\mathbb{Y} \hat{=} \text{Objekt } \#2 \text{ (Gegner)}$

$\mathbb{Z} \hat{=} \text{Resultat aus } \{\xi_1, \xi_2, \xi_3\} = \{>, \dot{=}, <\}$

POLR Erfolgsmodell

$$\log \text{odds}_\ell(\mathbf{x}, \mathbf{y}) = \underbrace{\mathbf{a}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{y}}_{\mathbf{a}^\top (\mathbf{x} - \mathbf{y})} + \zeta_\ell$$

Präferenzinterpretation

\mathbf{x} hat immer dann bessere Gewinnchancen als \mathbf{y} wenn $\log \text{odds}_1(\mathbf{x}, \mathbf{y}) > 0$ gilt, also

$$g_x := \mathbf{a}^\top \mathbf{x} > \mathbf{a}^\top \mathbf{y} + \zeta =: h_y$$

Semi-Ordnung!

$$\mathbf{x} \succ \mathbf{y} \quad \text{gdw. } [g_x, g_x + \zeta] \supseteq [g_y, g_y + \zeta]$$

Fußballturnier

GER : BRA	3:1
USA : LBY	0:1
UK : IRAN	2:2

Punktestandbezogen

XGER	XBRA	>
XBRA	XGER	<
XUSA	XLBY	<
XLBY	XUSA	>
XUK	XIRAN	=
XIRAN	XUK	=

Torstandbezogen

XGER	XBRA	>	3
XGER	XBRA	=	86
XGER	XBRA	<	1
XBRA	XGER	>	1
XBRA	XGER	=	86
XBRA	XGER	<	3
...	...		

Spezielle Form des POLR-Modells

$$\log \text{odds}_\ell(\mathbf{x}, \mathbf{y}) = \mathbf{a}^\top (\mathbf{x} - \mathbf{y}) + \begin{cases} -\infty & \ell = 0 \\ -\zeta & \ell = 1 \\ +\zeta & \ell = 2 \\ +\infty & \ell = 3 \end{cases}$$

Beweis.

Aus der strukturellen Symmetrie

$$P(> | \mathbf{x}, \mathbf{y}) = P(< | \mathbf{y}, \mathbf{x}) \text{ folgt:}$$

$$\Rightarrow p_1(\mathbf{x}, \mathbf{y}) = p_3(\mathbf{y}, \mathbf{x})$$

$$\Rightarrow q_1(\mathbf{x}, \mathbf{y}) - 0 = 1 - q_2(\mathbf{y}, \mathbf{x})$$

$$\Rightarrow \frac{q_1(\mathbf{x}, \mathbf{y})}{1 - q_1(\mathbf{x}, \mathbf{y})} = \frac{1 - q_2(\mathbf{y}, \mathbf{x})}{q_2(\mathbf{y}, \mathbf{x})}$$

$$\Rightarrow \text{odds}_1(\mathbf{x}, \mathbf{y}) = \text{odds}_2^{-1}(\mathbf{y}, \mathbf{x})$$

$$\Rightarrow + \log \text{odds}_1(\mathbf{x}, \mathbf{y}) = - \log \text{odds}_2(\mathbf{y}, \mathbf{x})$$

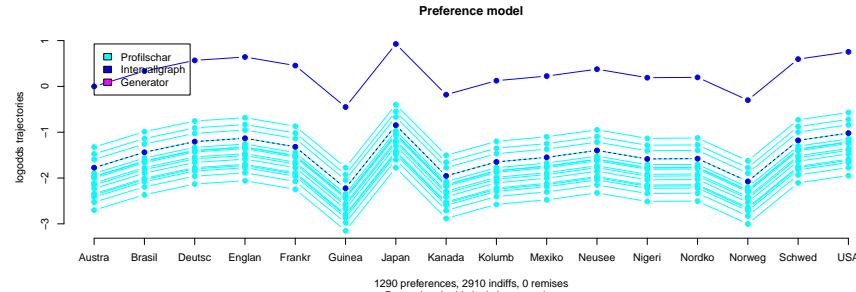
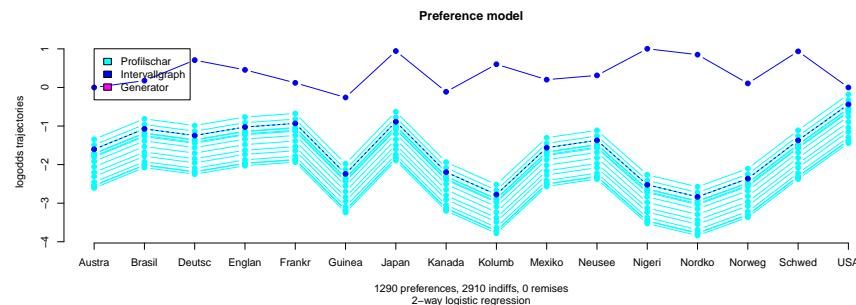
$$\Rightarrow 0 = \mathbf{a}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{y} + \zeta_1 + \mathbf{a}^\top \mathbf{y} + \mathbf{b}^\top \mathbf{x} + \zeta_2$$

$$\Rightarrow 0 = (\mathbf{a} + \mathbf{b})^\top (\mathbf{x} + \mathbf{y}) + (\zeta_1 + \zeta_2)$$

$$\Rightarrow \mathbf{b} = -\mathbf{a} \text{ und } \zeta_1 = -\zeta_2$$

□

Beispiel — Frauenfußball-WM 2011



Versionenräume

Naive Bayesregel

Multivariate lineare Regression

Logistische Regression

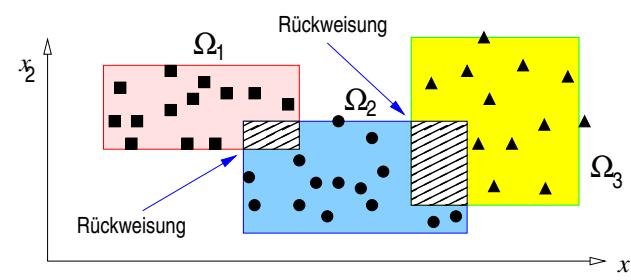
Ordinalen Regression und Präferenzmodelle

Statistische Entscheidungsbäume

Zusammenfassung

Parallelepiped-Klassifikator

Vollständige Konjunktion je zweier Literale $x_n \geq a_n, x_n \leq b_n, n = 1, \dots, N$



Vorteile

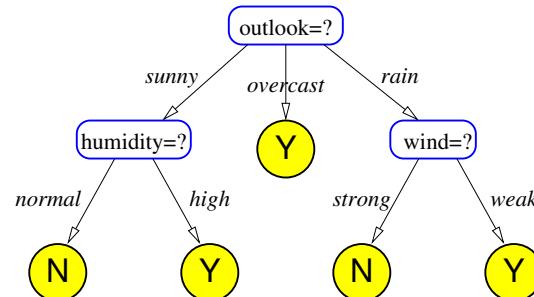
- Extrem schnelle Lernphase
- Effiziente Abrufphase
- Klassengebiete intuitiv zu deuten
- Nominalattribute handhabbar

Nachteile

- Achsenparallele Grenzen
- Unimodale Klassengebiete
- Ausgedehnte Rückweisungszonen
- Keine Rückschlüsse wahrscheinlichkeiten

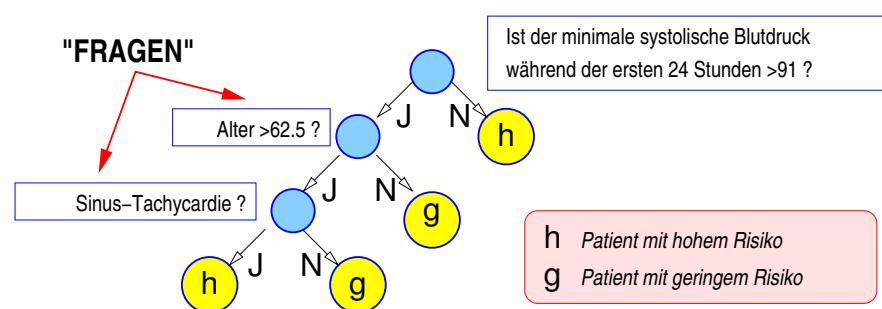
Entscheidungsbaum

Hierarchie sequentieller Auswahlfragen („multiple choice“)



Binärer Entscheidungsbaum

Hierarchie sequentieller Ja/Nein-Fragen



Diagnose für Herzinfarktpatienten

- 19 Attribute gemessen bzw. erfragt
- Patienten 30 Tage unter klinischer Beobachtung

Klassifikationsziel:

„Ist ein zweiter, diesmal tödlicher Infarkt eingetreten?“

Frage Typus

- Schwellwertdichotomie
- Zielwertdichotomie

Sportwetterempfehlungen

Vier nominale Wetterlagevariablen gegeben

Klassifikationsziel:

„Ist dieses Wetter zum Tennisspielen geeignet?“

Frage Typus

Wertverzweigung

Struktur eines Entscheidungsbaumes

Knoten $\hat{=}$ Fragen

\mathcal{B} bezeichnet die Menge aller Knoten.

Innere Knoten $\beta \in \mathcal{B}$ beherbergen eine Entscheidungsfrage:

$$Q(\beta) : \Omega \rightarrow \{1, \dots, L\}$$

Kanten $\hat{=}$ Antworten

Für $\beta \in \mathcal{B}$ ist β^\uparrow der Vorgängerknoten und $\beta^{(1)}, \dots, \beta^{(L)}$ sind die unmittelbaren Nachfolger.

Wurzelknoten $\hat{=}$ Startposition

$\beta_\Delta \in \mathcal{B}$ besitzt keinen Vorgänger. In β_Δ beginnt die Befragung des Objekts.

Blattknoten $\hat{=}$ Ergebnis

Die $\beta \in \mathcal{B}_\ell$ besitzen keine Nachfolger, aber eine Klassenmarkierung:

$$\delta_\ell : \mathcal{B}_\ell \rightarrow \{1, \dots, K\}$$

Befragung der Attributwerte

Dichotomien („Yin-Yang“-Fragen) und Wertverzweigungen

Attribut-Wert-Gleichungen

$$x_i = \text{low}$$

bei nominalen Merkmalen

(das negative Literal $x_i \neq \text{low}$ ist dazu dual)

Attribut-Wert-Ungleichungen

$$x_i \leq 3.14$$

bei ordinalen Merkmalen

(auch $x_i \geq 17$ oder Intervalle $18 \leq x_i \leq 65$ denkbar)

Wertverzweigungen

$$x_i = \text{red} | \text{blue} | \text{green}$$

bei Attributen mit kleinem $|\mathcal{X}_n|$

(eine Nachfolgerkante je Attributwert)

Teilmengenzugehörigkeit

$$x_i \in \{\text{cloudy}, \text{rainy}\}$$

bei nominalen Attributen

Reguläre Ausdrücke

$$x_i = \text{ababb} * \text{c} * \text{ba}$$

bei Wort- oder Zeichenketten

Simultanes Schleusen einer Objektmenge

Definition

Ist $(\mathcal{B}, Q, \delta_\ell)$ ein Entscheidungsbaum über Ω und $\omega \subseteq \Omega$ ein Datensatz, so definieren wir die **assoziierten Objektmengen** ω_β induktiv durch:

$$\omega(\beta) \stackrel{\text{def}}{=} \begin{cases} \omega & \beta = \beta_\Delta \\ \{x \in \omega_\beta \mid Q(\beta)(x) = j\} & \beta = \beta^{(j)} \end{cases}$$

Lemma

Ist $(\mathcal{B}, Q, \delta_\ell)$ ein Entscheidungsbaum über Ω , so gilt:

$$\Omega = \biguplus_{\beta \in \mathcal{B}_\ell} \Omega(\beta)$$

Der Entscheidungsbaum definiert ferner eine vollständige Zerlegung von Ω in Klassengebiete:

$$\Omega = \biguplus_{\kappa=1}^K \Omega_\kappa, \quad \Omega_\kappa \stackrel{\text{def}}{=} \bigcup_{\delta_\ell(\beta)=\kappa} \Omega(\beta)$$

Klassifikation eines Objekts

Hierarchisches Interview — „Durchschleusen“ bis zum Blattknoten

(Algorithmus)

1 INITIALISIERUNG

Setze $\beta = \beta_\Delta(\mathcal{B})$.

2 BEFRAGUNG

Reiche \mathbf{x} gemäß $Q(\beta)$ an einen Kindknoten weiter:

$$\beta \leftarrow \beta^{(i)}, \quad i = Q(\beta)(\mathbf{x})$$

3 TERMINIERUNG

Ist β ein Blattknoten, so lautet das Resultat:

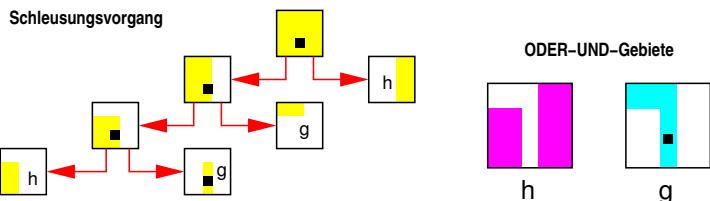
$$\delta(\mathbf{x}) = \delta_\ell(\beta)$$

Andernfalls \rightsquigarrow 2.

(Endeingabe)

Entscheidungsbäume als Hypothesen

Disjunktionen von Literalkonjunktionen



Beispiel

Intuitiv interpretierbare Klassenentscheidungen:

$$\Omega_h = (\{x_b \leq 91\} \wedge \{x_a \leq 62.5\} \wedge \{x_s \leq 0\}) \vee (\{x_b \leq 91\})$$

$$\Omega_g = (\{x_b \leq 91\} \wedge \{x_a \leq 62.5\} \wedge \{x_s \leq 0\}) \vee (\{x_b \leq 91\} \wedge \{x_a \leq 62.5\})$$

mit den Variablen (Merkmale)

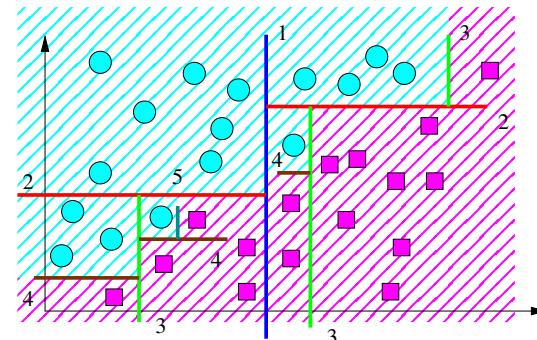
$$x_b = \text{minimaler systolischer Blutdruck}$$

$$x_a = \text{Alter des Patienten}$$

$$x_s = \text{Sinus-Tachycardie? (0 oder 1)}$$

Entscheidungsbäume für numerische Attribute?

Rekursive Halbraumbildung nach sukzessiven Schwellwertabfragen $x_n \leq \theta$



Vom Entscheidungsbaum induzierte Klassengebiete

$$\hat{\Omega}_\kappa = \bigcup_{m=1}^{M_\kappa} \hat{\Omega}_{\kappa,m}, \quad \hat{\Omega}_{\kappa,m} = \bigcap_{l=1}^{M_{\kappa,m}} H_{\kappa,m,l} = \text{Halbraum } \begin{cases} x_d \leq \theta \\ \text{oder} \\ x_d > \theta \end{cases}$$

TDI-Lernalgorithmus

Gierige Top-Down Induktion von Entscheidungsbäumen

(Algorithmus)

1 INITIALISIERUNG

Erzeuge einen Wurzelknoten $\beta = \beta_\Delta$ mit den assoziierten Stichproben $\omega_1, \dots, \omega_K$.

2 STOPPTEST

Ist β hinreichend **reinklassig**, so beende die lokale Konstruktion mit der Blattmarkierung

$$\delta_\ell(\beta) = \underset{\kappa}{\operatorname{argmax}} |\omega_\kappa(\beta)|.$$

3 FRAGEAUSWAHL

Wähle eine Frage $Q(\beta)$ mit maximaler Reduktion der **Entscheidungsunsicherheit**.

4 EXPANSION

Bilde die Nachfolgerknoten $\beta^{(1)}, \dots, \beta^{(L)}$ bezüglich $Q(\beta)$ und ihre assoziierten Stichproben

$$\omega_\kappa(\beta^{(l)}), \quad l = 1, \dots, L.$$

5 REKURSION

Fahre mit den Nachfolgern $\beta^{(l)}$ von β bei Schritt ② fort.

(summingup)

Lernen eines Entscheidungsbaumes

aus klassenetikettierten Beispielobjekten: $\omega = \omega_1 \uplus \omega_2 \uplus \dots \uplus \omega_K \subset \Omega$

Trennschärfe

Der Baum soll die Beispiele möglichst *korrekt klassifizieren*.

⇒ **Konsistenz**

Induktionskraft

Der Baum soll die Beispiele in geeigneter Weise *verallgemeinern*.

⇒ **geringe Knotenzahl**

Hypothesenraum

Welche Größe? Welche Form?

Welches Attribut? Welche Frage?

⇒ **gigantische Auswahl an E-Bäumen**

Lernen $\hat{=}$ Suche

Vollständige Suche ist NP-hart.

⇒ **lokal optimierende Suche (Bergsteiger-Algorithmus, „divide-and-conquer“)**

Stoppkriterium

Wann endet der Züchtungsvorgang?

Lokale Stoppkriterien

Wann endet die Knotenexpansion in einem Blatt?

- Wenn $\omega(\beta)$ nur noch einen Datenvektor enthält.
- Wenn $\omega(\beta)$ nur noch **Daten einer Klasse** enthält. ⇒ **Konsistenz**
- Wenn $|\omega(\beta)|$ eine gegebene Schranke unterschreitet.
- Wenn $|\omega(\beta)| - \max_\lambda |\omega_\lambda(\beta)|$ eine Schranke unterschreitet.

Überanpassung an die Lernbeispiele

Gefährlich in großen Bäumen durch *Zersplitterung* von ω auf die Blattknoten.

- Ist es wirklich weise, einen konsistenten Baum zu konstruieren ?
- ⇒ **globale a posteriori Stoppkriterien**
a.k.a. Baumschnitttechniken, „pruning“

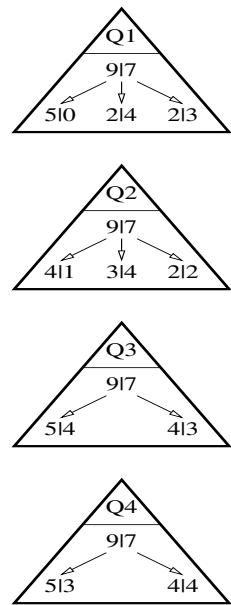
Die Frage nach der richtigen Frage

... bei Yuichiro Anzai im Autohaus ... (Beispiel)

Japanische Gebrauchtfahrzeuge

und ihre Veräußerungschancen am Markt

Objekt	cm^3	Türen	Autom.	Farbe	$x \in C$
x_1	2000	2T	ja	hell	+
x_2	2800	4T	ja	hell	+
x_3	2000	2T	nein	dunkel	-
x_4	1600	4T	ja	dunkel	-
x_5	1600	4T	ja	hell	-
x_6	2800	4T	ja	dunkel	+
x_7	2000	4T	ja	hell	+
x_8	2000	5T	nein	hell	-
x_9	1600	2T	nein	hell	+
x_{10}	2800	5T	ja	hell	+
x_{11}	2800	5T	nein	dunkel	+
x_{12}	2000	4T	ja	dunkel	-
x_{13}	1600	2T	nein	dunkel	+
x_{14}	2800	2T	nein	dunkel	+
x_{15}	1600	4T	nein	hell	-
x_{16}	2000	5T	ja	dunkel	-



Entscheidungsunsicherheit

Gütemaß für den Entmischungsgrad einer Verteilung

Definition

Es sei $K \in \mathbb{N}$ und $\{p_\kappa \mid \kappa = 1, \dots, K\}$ eine diskrete Wahrscheinlichkeitsverteilung. Eine Abbildung

$$\mathfrak{S} : \{p_1, \dots, p_K\} \mapsto u \in \mathbb{R}$$

heißt Maß für die **Entscheidungsunsicherheit** (Homogenität, „impurity“), falls gilt:

1. Die Größe $\mathfrak{S}(\cdot)$ ist nichtnegativ.
2. $\mathfrak{S}(\cdot)$ ist maximal für die Gleichverteilung $p_\kappa \equiv 1/K$
3. $\mathfrak{S}(\cdot)$ ist minimal für die definiten Verteilungen

$$\mathbf{e}_\lambda = (\underbrace{0, \dots, 0}_{\lambda-1}, 1, \underbrace{0, \dots, 0}_{K-\lambda}), \quad \lambda \in \{1, \dots, K\}$$

Auswahlregel für die „beste“ nächste Frage

Vorgehensweise

Welches ist die (lokal) zielführendste Frage ?

1. Definiere Entscheidungsunsicherheit einer Häufigkeitsverteilung
2. Definiere Entscheidungsunsicherheit eines Baumknotens
3. Definiere Entscheidungsunsicherheit einer Frage (in β)
4. Definiere Entscheidungsunsicherheit eines Teilbaums (unter β)

Relative Klassenhäufigkeit

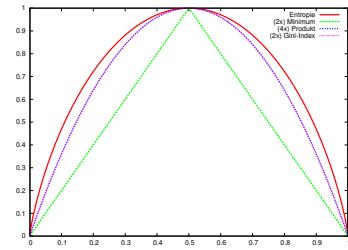
in der Teilstichprobe ω_β zum Knoten $\beta \in \mathcal{B}$:

$$\hat{p}_\kappa(\beta) = \frac{\text{Anzahl der } \Omega_\kappa\text{-Muster in } \beta}{\text{Anzahl aller Muster in } \beta} = \frac{|\omega_\kappa(\beta)|}{\sum_{\lambda=1}^K |\omega_\lambda(\beta)|}$$

⇒ ML-Schätzwert für $P(x \in \Omega_\kappa \mid x \in \Omega_\beta)$

Homogenitätsmaße

Minimum · Produkt · Gini-Index · Entropie



Extremalwerte

	\min	\max
\mathfrak{S}_m	0	$1/K$
\mathfrak{S}_p	0	$1/K^K$
\mathfrak{S}_g	0	$1 - 1/K$
\mathfrak{S}_e	0	$\log_2 K$

Lemma

Die folgenden Abbildungen sind (für festes $K \in \mathbb{N}$) Beispiele für Homogenitätsmaße:

$$\mathfrak{S}_m(\mathbf{p}) \stackrel{\text{def}}{=} \min_\kappa p_\kappa$$

$$\mathfrak{S}_g(\mathbf{p}) \stackrel{\text{def}}{=} \sum_{\lambda \neq \kappa} p_\lambda \cdot p_\kappa$$

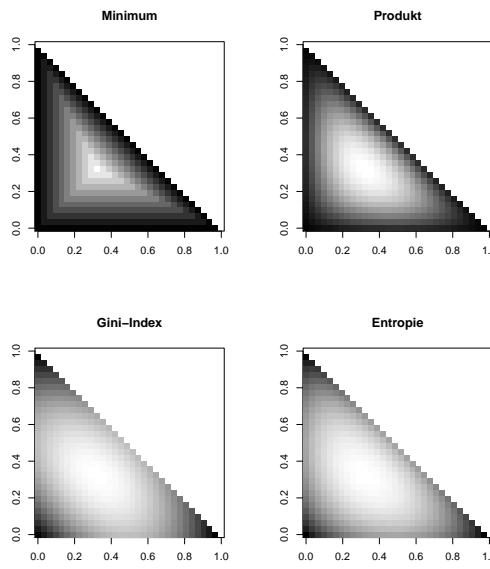
$$\mathfrak{S}_p(\mathbf{p}) \stackrel{\text{def}}{=} \prod_\kappa p_\kappa$$

$$\mathfrak{S}_e(\mathbf{p}) \stackrel{\text{def}}{=} - \sum_\kappa p_\kappa \cdot \log_2 p_\kappa$$

Für den Gini-Index gilt $\mathfrak{S}_g(\mathbf{p}) = 1 - \|\mathbf{p}\|^2$.

Homogenitätsmaße

Drei Ereignisse — Darstellung in der (p_1, p_2) -Ebene



Minimum/Produkt

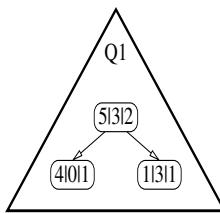
Geringe Homogenität (Unsicherheit) wird bereits dann signalisiert, wenn nur eines der drei Ereignisse unwahrscheinlich ist.

⇒ unbrauchbar

Entropie/Gini

Grundverschiedene Formeln, aber kaum unterschiedliche Funktionswerte.

⇒ praktisch äquivalent



Rechenbeispiel (Gini-Index)

Ausgangsknoten β

Der Knoten β beherbergt die Verteilung $p = (0.5, 0.3, 0.2)$, also gilt

$$\mathfrak{S}_{\text{Gini}}(p) = 1 - 0.25 - 0.09 - 0.04 = 0.62$$

Erste Frage Q_1

Die Entscheidungsunsicherheiten der Q_1 -Nachfolger lauten

$$\mathfrak{S}_{\text{Gini}}(\beta^{(1)} | Q_1) = 1 - 0.64 - 0.04 = 0.32$$

$$\mathfrak{S}_{\text{Gini}}(\beta^{(2)} | Q_1) = 1 - 0.04 - 0.36 - 0.04 = 0.56$$

Die mittlere E.U. der Nachfolger bzw. ihre Reduktion betragen also

$$\mathfrak{S}_{\text{Gini}}(\beta | Q_1) = 0.5 \cdot 0.32 + 0.5 \cdot 0.56 = 0.44$$

$$\Delta_{Q_1} \mathfrak{S}_{\text{Gini}}(\beta) = 0.62 - 0.44 = 0.18$$

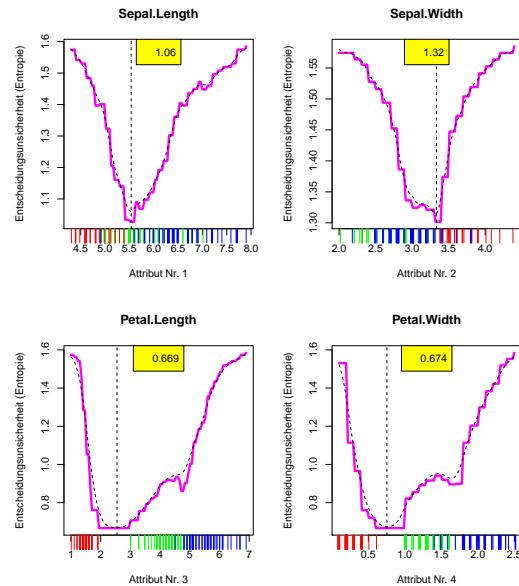
Zweite Frage Q_2

Auf dieselbe Weise errechnet sich für die konkurrierende Frage der Wert

$$\Delta_{Q_2} \mathfrak{S}_{\text{Gini}}(\beta) = 0.62 - 0.6 \cdot 0.5 - 0.4 \cdot 0.5 = 0.12$$

Folglich ist Q_1 der Frage Q_2 vorzuziehen.

Rechenbeispiel (Entropiemaß)



IRIS-Datensatz

150 Objekte
4 Attribute
3 Kategorien

$$\begin{Bmatrix} 50 \\ 50 \\ 50 \end{Bmatrix}$$

Wurzelknoten

Berechne für jedes Attribut x_n den EU-minimalen Schwellenwert θ_n

$$Q(\beta) : x_3 \stackrel{?}{\leq} 2.65$$

Reduktion der Entscheidungsunsicherheit

Entscheidungsunsicherheit im Knoten β

$$\mathfrak{S}(\beta) \stackrel{\text{def}}{=} \mathfrak{S}(\hat{p}^{(\beta)}), \quad \hat{p}_\kappa \stackrel{\text{def}}{=} \frac{|\omega_\kappa(\beta)|}{|\omega(\beta)|}$$

Verzweigungswahrscheinlichkeiten der Frage Q in β

$$\hat{P}(\beta^{(i)} | \beta) \stackrel{\text{def}}{=} \frac{|\omega(\beta^{(i)})|}{|\omega(\beta)|}, \quad i = 1, \dots, L$$

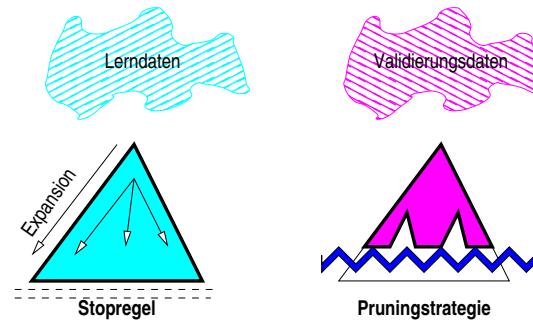
Entscheidungsunsicherheit nach der Frage Q in β

$$\mathfrak{S}(\beta | Q) \stackrel{\text{def}}{=} \sum_i \hat{P}(\beta^{(i)} | \beta) \cdot \mathfrak{S}(\beta^{(i)})$$

Reduktion der Entscheidungsunsicherheit durch Q

$$\Delta_Q(\beta) \stackrel{\text{def}}{=} \mathfrak{S}(\beta) - \mathfrak{S}(\beta | Q)$$

Überanpassung an die Lernbeispiele



Fragmentierung der Lerndaten

- β reinklassig $\sim \omega$ perfekt klassifiziert
- viele Lerndaten \sim großer Entscheidungsbaum
- Insignifikante Fragen in unteren Zweigen
- Unzuverlässige Entscheidung in den Blättern
- Stopptregeln sind „kurzsichtig“

Abhilfe

- „early stopping“
- Züchten & Zurück-schneiden

Aufspüren und Tilgen nutzloser Teilbäume

CART Pruning

Züchtung eines überangepassten Baumes
Sukzessive Vergrößerung (Entfernen schwacher Äste)
Auswahl des besten Teilbaums

Lerndaten ω
Modellstrafterm
Validierungsdaten $\tilde{\omega}$

Lokaler Resubstitutionsfehler

Relative Anzahl der Fehler bei Entscheidung in β :

$$R(\beta) \stackrel{\text{def}}{=} \frac{\# \text{ falsch klassifiziert in } \beta}{\# \text{ alle Objekte}} = \frac{|\omega(\beta)| - \max_{\kappa} |\omega_{\kappa}(\beta)|}{|\omega(\beta_{\Delta})|}$$

Kumulativer Resubstitutionsfehler

$\mathcal{B}_{\ell}^{\beta}$ = Menge aller Blattknoten in dem von β dominierten Teilbaum

$$R^*(\beta) \stackrel{\text{def}}{=} \sum_{\beta' \in \mathcal{B}_{\ell}^{\beta}} R(\beta')$$

Bemerkung Es gilt für alle $\beta \in \mathcal{B}$: $R^*(\beta) \leq R(\beta)$

Strafterme versus Kreuzvalidierung

Effizienz eines Teilbaums

Gut entmischende Teilbäume werden belohnt,
aber zersplitterungsverdächtige Teilbäume werden bestraft!

$$\Delta_{\text{eff}}(\beta) \stackrel{\text{def}}{=} \frac{\text{Fehlerzuwachs in } \beta}{\# \text{ eingesparte Knoten}} = \frac{R(\beta) - R^*(\beta)}{|\mathcal{B}_{\ell}^{\beta}| - 1}$$

Kreuzvalidierungsfehler

Jedem Objekt $x \in \tilde{\omega}$ wird durch einen Entscheidungsbaum ein Blattknoten $\beta(x)$ und damit auch eine Klassenmarkierung $\delta_{\ell}(\beta(x))$ zugeordnet.

$$\tilde{\varepsilon}(\mathcal{B}) \stackrel{\text{def}}{=} \frac{\sum_{\kappa=1}^K |\{x \in \tilde{\omega}_{\kappa} \mid \delta_{\ell}(\beta(x)) \neq \kappa\}|}{|\tilde{\omega}(\beta_{\Delta})|}$$

CART Pruning-Algorithmus

Breiman, Friedman, Olshen & Stone (1984)

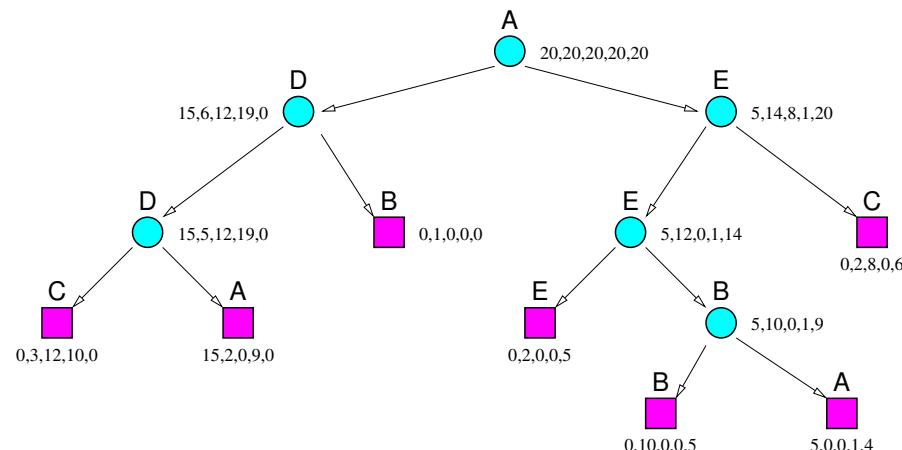
(Algorithmus)

- 1 **ZÜCHTEN**
Expandiere initialen Baum $\mathcal{B}^{(0)}$ mittels Lerndaten $\omega_1, \dots, \omega_K$ unter Einhaltung des „Reinheitsgebotes“.
- 2 **SUKZESSIVES ZURÜCKSCHNEIDEN**
Erzeuge eine Folge gestützter Teilbäume von $\mathcal{B}^{(0)}$
 - a Setze $i \rightarrow 0$.
 - b Berechne alle Effizienzwerte $\Delta_{\text{eff}}(\beta)$, $\beta \in \mathcal{B}^{(i)}$.
 - c Wähle Knoten $\beta^* \in \mathcal{B}^{(i)}$ mit minimaler Effizienz.
 - d Kappe den Teilbaum unterhalb β^* .
 - e Setze $i \leftarrow i + 1$ und bezeichne gekürzten Baum als $\mathcal{B}^{(i)}$.
 - f Ist $\mathcal{B}^{(i)} \neq \{\beta_{\Delta}\}$, dann \rightsquigarrow b.
- 3 **AUSWAHL NACH VALIDIERUNGSFEHLER**
Wähle aus $\{\mathcal{B}^{(i)} \mid i = 0, 1, 2, \dots\}$ denjenigen Baum mit geringstem Fehler auf den Validierungsdaten $\tilde{\omega}_1, \dots, \tilde{\omega}_K$.

(zum Download)

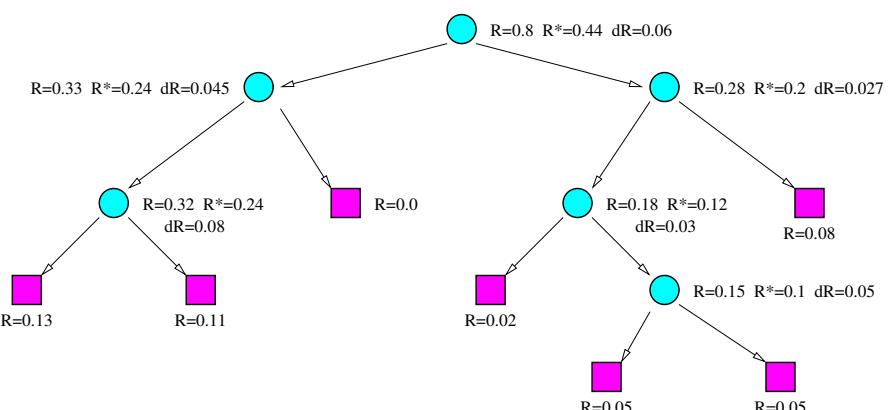
Beispiel — CART-Algorithmus

5 Klassen · 6 Fragen · 7 Blätter · 100 Objekte



Beispiel — CART-Algorithmus

5 Klassen · 6 Fragen · 7 Blätter · 100 Objekte



Lokale Resubstitutionsfehlerraten

Kumulative Resubstitutionsfehlerraten

Effizienzen — nur innere Knoten werden gezählt

Kreuzvalidierendes Stutzen der Äste

„Frühe Validierung“ — schon zur Bewertung statt erst zur Auswahl

Lokale Fehlerrate

im Knoten β nach Einschleusen der Konterdaten ω :

$$\varepsilon(\beta) \stackrel{\text{def}}{=} 1 - \frac{|\omega_\kappa(\beta)|}{|\omega(\beta)|} \quad \text{mit } \kappa := \delta_\ell(\beta) \text{ oder } \kappa := \operatorname{argmax}_\lambda |\omega_\lambda(\beta)|$$

Kumulative Fehlerrate

nach Durchschleusen von ω bis zu den Blättern:

$$\varepsilon^*(\beta) \stackrel{\text{def}}{=} \sum_{\beta' \in \mathcal{B}_\ell^\beta} \frac{|\omega(\beta')|}{|\omega(\beta)|} \cdot \varepsilon(\beta')$$

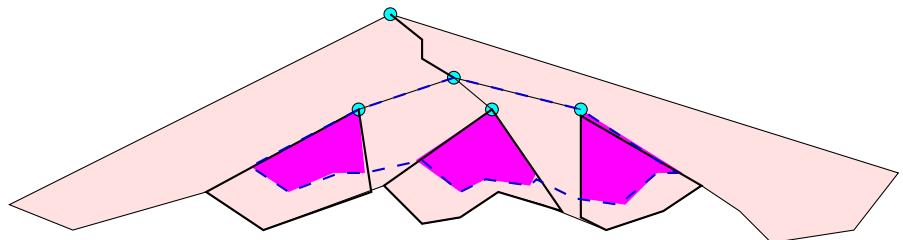
Die **Gesamtfehlerrate** ist $\varepsilon(\mathcal{B}) = \varepsilon^*(\beta_\Delta)$

Minimale Fehlerrate

aller Teilbäume \mathcal{B}^β unterm Knoten β :

$$\varepsilon^\forall(\beta) \stackrel{\text{def}}{=} \min \{ \varepsilon(\mathcal{B}') \mid \mathcal{B}' \text{ Teilbaum von } \mathcal{B}^\beta \}$$

Induktive Bottom-Up Beschneidung

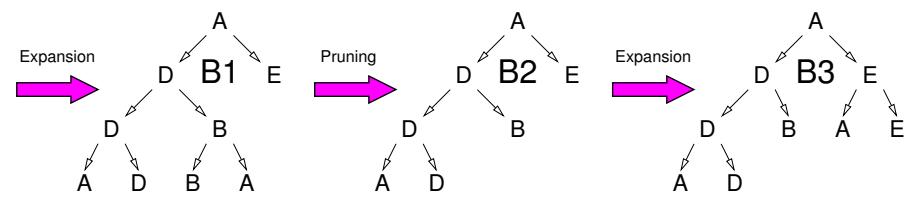


Lemma

Sei $(\mathcal{B}, Q, \delta_\ell)$ ein binärer Entscheidungsbaum über $\Omega = \mathbb{IR}^D$. Die optimale Fehlerrate des Teilbaums \mathcal{B}^β berechnet sich nach folgender Rekursion:

$$\varepsilon^\forall(\beta) = \begin{cases} \varepsilon(\beta) & \beta \in \mathcal{B}_\ell \\ \min \left(\sum_{\beta' \prec \beta} p(\beta'|\beta) \cdot \varepsilon^\forall(\beta') \right) & \beta \notin \mathcal{B}_\ell \end{cases}$$

Wiederholtes Züchten und Beschneiden



Expansionsphase · top-down

1. Schleuse die Daten ω^a bis zu den Blattknoten von $\mathcal{B}^{(i)}$.
2. Bestimme die Mengen $\omega_\kappa^a(\beta)$ für alle κ, β .
3. Züchte für alle Blattknoten $\beta \in \mathcal{B}_\ell^{(i)}$ einen Teilbaum unter β mittels $\omega^a(\beta)$.

Pruningphase · bottom-up

1. Schleuse die Daten ω^a bis zu den Blattknoten von $\mathcal{B}^{(i)}$.
2. Markiere alle $\beta \in \mathcal{B}^{(i)}$ mit neuen Klassen $\delta_\ell(\beta)$.
3. Überprüfe alle β durch Vergleich von lokaler und minimaler RFR auf Eliminierbarkeit.

Die Auswahl der besten Frage

Monothetische Knoten \rightsquigarrow keine Attributkombinationen

Problem

Die Expansion eines jeden Knotens β im TDI-Algorithmus erfordert die $\Delta_Q(\beta)$ -Bewertung jeder Frage Q zu jedem Attribut X_n !

Nominale Attribute

Wieviele Zwei- oder Mehrwege-Fragen sind zu testen?

- Wertverzweigung
- Attribut-Wert-Gleichung
- Literalkomplex

eine Frage/Attribut

$|X_n|$ Targets/Attribut

$2^{|X_n|}/2$ Mengen/Attribut

Numerische und ordinale Attribute

Wieviele Schwellenwert-Fragen sind zu testen?

- Ordinale Attribute
- Numerische Attribute

$|X_n| - 1$ Schwellen/Attribut

$|\omega(\beta)| - 1$ Schwellen/Attribut

Gelfands IEP-Algorithmus

„Iterative Expansion-Pruning“

- 1 INITIALISIERUNG
Setze $i \leftarrow 0$ und $\mathcal{B}^{(0)} \leftarrow \{\beta_\Delta\}$.

- 2 ERSTES EXPANDIEREN
Expandiere $\mathcal{B}^{(i)}$ mit den Daten ω^a .
 $\rightsquigarrow \mathcal{B}^{(i+1)}$

- 3 ERSTES STUTZEN
Beschneide $\mathcal{B}^{(i+1)}$ mit den Daten ω^b .
 $\rightsquigarrow \mathcal{B}^{(i+2)}$

- 4 ZWEITES EXPANDIEREN
Expandiere $\mathcal{B}^{(i+2)}$ mit den Daten ω^b .
 $\rightsquigarrow \mathcal{B}^{(i+3)}$

- 5 ZWEITES STUTZEN
Beschneide $\mathcal{B}^{(i+3)}$ mit den Daten ω^a .
 $\rightsquigarrow \mathcal{B}^{(i+4)}$

- 6 TERMINIERUNG
Falls $\mathcal{B}^{(i+2)} \equiv \mathcal{B}^{(i+4)}$, dann \rightsquigarrow ENDE.

- 7 WIEDERHOLUNG
Setze $i \leftarrow i + 4$ und weiter bei \rightsquigarrow 2.

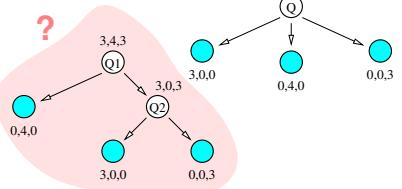
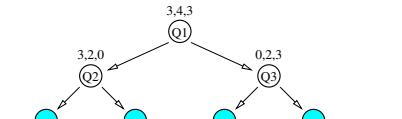
$$\text{Disjunkt: } \left\{ \begin{array}{l} \omega^a = \omega_1^a \cup \omega_2^a \cup \dots \cup \omega_k^a \\ \omega^b = \omega_1^b \cup \omega_2^b \cup \dots \cup \omega_k^b \end{array} \right\}$$

Die Befragung nominaler Attribute

Symmetrische Verzweigung $x_n = ?$ versus asymmetrische Verzweigung $x_n = \xi_\ell$

Datenfragmentierung

Die minimal zersplitternde Folge binärer Fragen wird nicht automatisch gefunden.



Unbalancierte Auswahl

Die Maximierung der Entscheidungssicherheit bevorzugt systematisch Fragen mit hohem Verzweigungsfaktor.

Gain Ratio Impurity

Abhilfe schafft Normierung auf die maximale Entropie:

$$\Delta'_Q(\beta) \stackrel{\text{def}}{=} \frac{\mathfrak{I}(\beta) - \sum_{j=1}^L p_j \cdot \mathfrak{I}(\beta_j)}{\mathcal{H}(p_1, \dots, p_L)}$$

Literalkomplexe in Zweiklassen-Szenarien

Auswahl der besten Teilmenge

Aufgabenstellung

Finde zum Attribut X_n in β diejenige Teilmengenfrage

$$Q : x_n \mapsto \begin{cases} 1 & x_n \in U \\ 0 & x_n \notin U \end{cases}, \quad U \subset X_n$$

mit der max. Reduktion $\Delta_Q(\beta)$ der Entscheidungsunsicherheit.

Premiumschlitten & Volumenmodelle

Objekte = Fahrzeuge · Klassen Ω_1 und Ω_2 · Attribut x_{19} (Hersteller)

X_{19}	VW	Benz	Alfa	Dacia	BMW	Porsche
Ω_1	112	9	3	1	28	5
Ω_2	112	1	2	4	12	0
$\hat{P}(1 \xi)$	0.5	0.9	0.6	0.2	0.7	1.0

Aufsteigende $\hat{P}(1|\xi)$ -Sortierung \Rightarrow nur 5 mögliche optimale Fragen:

$$\begin{array}{ll} x_{19} \in \{\text{Dacia, VW}\} & x_{19} \in \{\text{Dacia, VW, Alfa, BMW}\} \\ x_{19} \in \{\text{Dacia}\} & x_{19} \in \{\text{Dacia, VW, Alfa}\} \quad x_{19} \in \{\text{Dacia, VW, Alfa, BMW, Benz}\} \end{array}$$



Der Zwillingssatz („Twoing Theorem“)

Linearer Suchaufwand für entropiegesteuertes Zweiklassen-Lernen

Satz

Es sei $\mathcal{X}_n = \{\xi_1, \dots, \xi_L\}$ der (nominale) Wertebereich des n -ten Attributs, und es zerfalle die Lernstichprobe $\omega \subset \Omega$ in zwei Klassenbereiche ω_1, ω_2 . Mit den Bezeichnungen

$$\hat{P}(\kappa|\xi_\ell) \stackrel{\text{def}}{=} \frac{|\{\mathbf{x} \in \omega_\kappa \mid x_n = \xi_\ell\}|}{|\{\mathbf{x} \in \omega \mid x_n = \xi_\ell\}|}$$

für $\kappa = 1, 2$ und $\ell = 1, \dots, L$ seien infolge geeigneter Sortierung der ξ_ℓ die Häufigkeitsbeziehungen

$$\hat{P}(1|\xi_1) \leq \hat{P}(1|\xi_2) \leq \dots \leq \hat{P}(1|\xi_L)$$

gültig. Dann besitzt die Teilmengenfrage mit der maximalen Homogenitätsreduktion in Bezug auf das Entropiemaß die Gestalt

$$x_n \in \{\xi_1, \dots, \xi_\ell\}$$

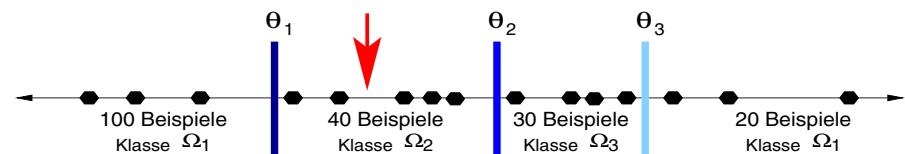
für ein geeignetes ℓ mit $1 < \ell < L$.

Schwellenwertfragen

Numerische und ordinale Attribute · zwei oder mehr Klassen

Reduzierter Suchaufwand für $\theta \in \mathcal{X}_n = \mathbb{R}$

- Nur $T_\beta = |\omega(\beta)|$ Mittelpunktschwellen zu prüfen.
- Nur klassentrennende Schwellen können $\Delta_Q(\beta)$ -maximal sein.
- Es gibt eine Rekursionsformel für $\Delta_{Q,n,\theta}(\beta)$.



Sortierung $O(T \log T)$

Aufsteigendes Sortieren der X_n -Attributwerte in $\omega(\beta)$:

$$a_1 < a_2 < a_3 \dots < a_t < \dots < a_{T_\beta}$$

Mittelpunktschwellen

Suffizienter Satz von Schwellenwerten für $Q_{n,\theta}$:

$$\theta_t = \frac{a_{t+1} - a_t}{2}, \quad t = 1, 2, \dots, T_\beta - 1$$

Separierende Schwellenwerte

Definition

Eine Mittelpunktschwelle θ_t von $\{x_n \mid x \in \omega(\beta)\}$ heißt **innere Schwelle** von \mathcal{X}_n in β , falls alle Objekte $x \in \omega(\beta)$ mit $x_n = a_t$ oder $x_n = a_{t+1}$ zu einundderselben Klasse Ω_κ gehören.

Andernfalls heißt θ_t **separierende Schwelle** oder **Klassengrenze**.

Lemma (Fayyad & Irani, 1992)

Sind $[\theta_t]$ die Mittelpunktschwellen zur assoziierten Stichprobe $[\omega_\kappa(\beta)]$ von β zum Attribut X_n , und gilt

$$\theta_{t^*} = \underset{\theta_t}{\operatorname{argmax}} \Delta_{\{x_n \leq \theta_t\}}(\beta)$$

für die entropiebezogene Entscheidungsunsicherheit, so ist θ_{t^*} notwendigerweise eine Klassengrenze.

Bemerkung

Je stärker sich die Objekte klassenweise auf der X_n -Achse häufen, desto weniger Reduktionswerte müssen berechnet werden.

Inkrementelle $\Delta_Q(\beta)$ -Berechnung

Lemma

Es seien $\theta_1 < \theta_2$ zwei benachbarte Klassengrenzen für \mathcal{X}_n in $\omega(\beta)$, zwischen denen genau m Muster der Klasse Ω_κ liegen. Dann gilt die Rekursionsformel

$$\begin{aligned}\Delta_{\{x_n \leq \theta_2\}}(\beta) &= \Delta_{\{x_n \leq \theta_1\}}(\beta) \\ &+ \frac{h(\ell, r) - h(\ell + m, r + m) + h(\ell_\kappa + m, r_\kappa + m) - h(\ell_\kappa, r_\kappa)}{T}\end{aligned}$$

mit den Abkürzungen

$$h(p, q) = p \log_2 p - q \log_2 q$$

und den Zählwerten

$$\begin{array}{lll} \ell_\kappa &= |\{x_d < \theta_1 \mid x \in \omega_\kappa\}| & \ell = \sum_\kappa \ell_\kappa \\ r_\kappa &= |\{x_d > \theta_1 \mid x \in \omega_\kappa\}| & r = \sum_\kappa r_\kappa \end{array}$$

Attribute mit Fehlanzeigen

Imputation

Wenn $x_n = ?$, so setze einen Standardwert $\hat{\xi}$ ein.

- Wähle für $\hat{\xi}$ das globale Attributmittel μ_n .
- Wähle für $\hat{\xi}$ das lokale Attributmittel $\mu_n(\beta)$.

Überlagerung

Wenn $x_n = ?$, so folge in der Abrupphase parallel allen Verzeigungen.

- Während der Lernphase werden defiziente Objekte bei der $\Delta_Q(\beta)$ -Berechnung ignoriert oder pejorisiert.

Surrogate Split

Wenn $x_i = ?$, so beantworte in der Abrupphase die/eine Ersatzfrage.

- In der Lernphase merkt man/frau sich die besten Fragen zum zweitbesten Attribut (ggf. weitere Alternativen).

Polythetische Entscheidungsfragen

Über das Züchten „schiefer“ statt achsenparalleler Entscheidungsbäume

Attributübergreifende Dichotomien (linear)

$$a_0 + \sum_{i=1}^N a_i x_i \stackrel{?}{\leq} 0$$

Trennfunktionsparameter mit guter Klassenentmischung!

CART/LC

Gradientenabstieg via $\Delta_a(\beta)$

SADT

Simulated Annealing of Decision Trees

LMDT

Linear Machine Decision Trees („ADALINE-Knoten“)

QUEST

Multivariate Variante des QUEST-Algorithmus

Leo Breimans Random Forests

Zweifache Ensembletechnik: Objekte (bagging) & Attribute

(Algorithmus) Lernprobe ω , Wälder $M \in \mathbb{N}$, Auswahl $T_b \leq |\omega|$ und $N_b \ll N$.

1 LERNPHASE

Erzeuge Bäume $\mathcal{B}^{(m)}$, $m = 1, \dots, M$:

- Lernprobe $\omega^{(m)} \subset \omega$ via T_b -Bootstrap mit Ersetzen
- Zufallsbaum $\mathcal{B}^{(m)}$ via TDI-Algorithmus
- EINGESCHRÄNKTE LOKALE FRAGEAUSWAHL: $\mathcal{A}_\beta \subset \{\mathcal{X}_1, \dots, \mathcal{X}_N\}$, $|\mathcal{A}_\beta| = N_b$ via N_b -Bootstrap ohne Ers.
- Kein Zurückstutzen!

2 ABRUPPHASE

Mehrheitsentscheidung unter allen Bäumen des Waldes:

$$\kappa^*(x) = \operatorname{argmax}_\kappa |\{\mathcal{B}^{(m)} \mid \delta_\ell(\beta^{(m)}(x)) = \kappa\}|$$

Bemerkung

Pro: Effizient, skalierbar (N, T), exzellentes Erkennungsverhalten.

Contra: Überanpassung, Reproduzierbarkeit, Präferenz stufenreicher Nominalattribute.

Versionenräume

Naive Bayesregel

Multivariate lineare Regression

Logistische Regression

Ordinale Regression und Präferenzmodelle

Statistische Entscheidungsbäume

Zusammenfassung

Zusammenfassung (4)

1. Der **Konzeptraum** enthält die **zu lernenden**, der **Hypothesenraum** die **lernbaren** Teilmengen des Objektraums.
2. Der **Versionenraum** besteht aus allen **konsistenten** Hypothesen und ist als **Halbordnungsintervall** darstellbar.
3. Die Hypothesen des **Sterns** grenzen ein Positivbeispiel gegen alle Negativbeispiele ab.
4. **Lineare Diskriminanten** approximieren die **ideale Trennfunktion** im Quadratmittelsinn.
5. **Loglineare Diskriminanten** approximieren die **a posteriori Klassenwahr'keiten**.
6. Beide Lernverfahren lassen sich **regularisieren** und **dualisieren**.
7. **Entscheidungsbäume** klassifizieren durch hierarchische Befragung **numerischer & diskreter Attribute**.
8. Sie werden durch ein **gieriges Top-Down-Verfahren** aus den Daten gelernt.
9. Für die lokale Suche nach der maximal **klassenentmischenden** Frage gibt es effiziente Verfahren.

MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2020

Prof. E.G. Schukat-Talamazzini

Stand: 10. September 2020

Teil V

Gruppierung von Objekten

Hierarchisch K-Means EM Relational Konzeptuell Spektral Clustergüte Σ

Überwachungsszenarien

Etikettierung der Lerndatenobjekte nach Klassenzugehörigkeit ?

Überwachtes Lernen

Der Lehrer stellt Zielwert **aller** Lernobjekte bereit.

{ Klassifikation
Vorhersage }

Halbüberwachtes Lernen

Der Lehrer verrät Zielwert **weniger** Lernobjekte.

{ Bootstrap
Transduktion }

Reinforcement Lernen

Der Lehrer übt **Erfolgskontrolle** („feed-back“).

{ Spielstrategie
Aktionsplanung }

Unüberwachtes Lernen

Der Lehrer stellt **keinerlei** Zielwerte bereit.

{ Gruppierung
Assoziation }

Hierarchisch K-Means EM Relational Konzeptuell Spektral Clustergüte Σ

Gruppierung a.k.a. Clusteranalyse

Partitionierung der Datenobjekte in Ballungs- oder Häufungsgebiete

Objektrepräsentation

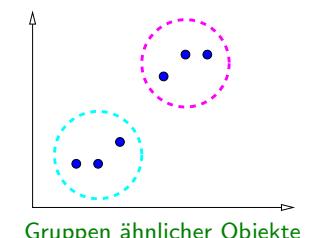
Vektorraum · Attribute · Metrik

Zielgröße

global · lokal · ad hoc

Zerlegungsstrategie

- top-down
- bottom-up
- Austausch (K-means, EM)
- split & merge



Gruppen ähnlicher Objekte

Gruppenrepräsentation

Mengen · Prototypen · Verteilungen
Formeln · Regeln



Unscharfe Gruppenzuordnung

Hierarchische Gruppierung: agglomerativ/divisiv

Austauschverfahren: (fuzzy) K-means

Mischungsidentifikation

Relationale Gruppierung

Konzeptuelle Gruppierung

Spektrale Gruppierung

Clustergütemaße

Zusammenfassung

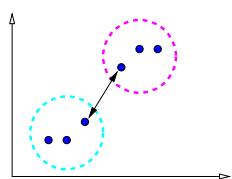
Mengendistanzfunktionen

$$d : \wp\Omega \times \wp\Omega \rightarrow \mathbb{R}_0^+$$

Single-Linkage

Kürzeste Brücke zwischen zwei Gruppen

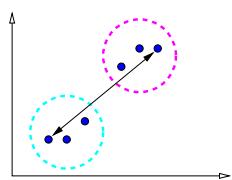
$$d_{SL}(A, B) \stackrel{\text{def}}{=} \min_{x \in A} \min_{y \in B} d(x, y)$$



Complete-Linkage

Durchmesser nach Vereinigung zweier Gruppen

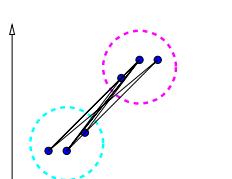
$$d_{CL}(A, B) \stackrel{\text{def}}{=} \max_{x \in A} \max_{y \in B} d(x, y)$$



Average-Linkage

Mittlerer bipartiter Punkteabstand

$$d_{AL}(A, B) \stackrel{\text{def}}{=} \frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y)$$



Agglomerative Gruppierung Generischer Bottom-up-Algorithmus

(Algorithmus)

Gegeben sind die Datenobjekte $x_1, \dots, x_T \in \Omega$

1 INITIALISIERUNG

Starte mit $K = T$ Gruppen $\omega_t = \{x_t\}$, $t = 1..T$.

2 DISTANZBERECHNUNG

Berechne für alle $1 \leq \kappa < \lambda \leq K$:

$$D_{\kappa\lambda} \stackrel{\text{def}}{=} d(\omega_\kappa, \omega_\lambda)$$

3 VEREINIGUNG

Vereinige die beiden Gruppen $\omega_{\kappa^*}, \omega_{\lambda^*}$ mit

$$(\kappa^*, \lambda^*) = \underset{\kappa, \lambda}{\operatorname{argmin}} D_{\kappa\lambda}$$

4 TERMINIERUNG

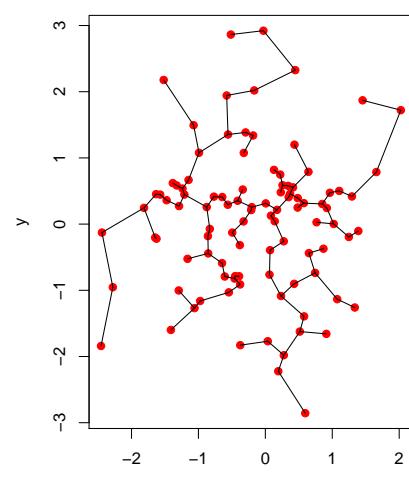
Wenn $K = 1$ dann ENDE, sonst \rightsquigarrow 2.

(Visualisierung)

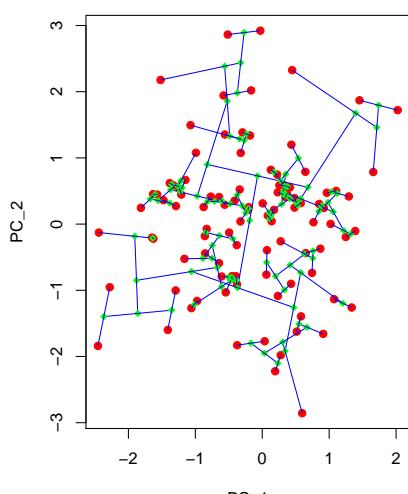
Kettenbildung und Lasso-Effekt

Beispiel mit $T = 100$ Objekten im \mathbb{R}^2

Minimum Spanning Tree



Average-linkage clusters



Mengendistanzfunktionen

Welche ist die beste?

Single-Linkage

„Ketteneffekt“

- erzeugt minimalen Spannbaum
- \oplus schwach monoton inv. monot. d -Transf.

Average-Linkage

weder Ketten- noch Lassoeffekt

- bevorzugt sphärische Ballungsgebiete
- \oplus schwach monoton Δ -invariant

Complete-Linkage

„Lassoeffekt“

- extrem anfällig gegen Ausreißer
- \oplus schwach monoton inv. monot. d -Transf.

Getrimmte Distanzen

Diese Effekte lassen sich abmildern, wenn in der Distanzformel jeweils $q > 1$ kleinste bzw. größte Distanzen eliminiert werden, wodurch die Einflußdramatik eventueller Ausreißer eingedämmt wird.

Dreiecksungleichung und Monotonie

Satz (Lance & Williams, 1967)

Es sei eine rekursive Form der Mengendistanzfunktion vorausgesetzt.

1. Gilt $\alpha_1 + \alpha_2 \geq 1$, $\beta \geq 0$ und $\gamma = 0$ und gilt die Dreiecksungleichung für alle Gruppendistanzen, so gilt sie auch noch nach der $d(\cdot, \cdot)$ -optimalen Vereinigung:

$$d(A_1 \uplus A_2, B) + d(A_1 \uplus A_2, C) \geq d(B, C)$$

2. Gilt $\alpha_1 + \alpha_2 + \beta \geq 1$ und $\gamma = 0$, so steigen die Gruppendistanzen schwach monoton an:

$$d(A_1 \uplus A_2, B) \geq d(A_1, A_2)$$

3. Erfüllt das Agglomerationsverfahren sogar die strenge Monotonie, so gelten für jede intermediäre Gruppenstruktur (mit Gruppe A) die **ultrametrischen** Ungleichungen:

$$\forall \mathbf{x}, \mathbf{y} \in A, \forall \mathbf{z} \notin A: d(\mathbf{x}, \mathbf{y}) < d(\mathbf{x}, \mathbf{z})$$

Definition

Gehorcht eine Distanzfunktion $d : \mathfrak{P}\Omega \times \mathfrak{P}\Omega \rightarrow \mathbb{IR}$ in eindeutiger Weise dem Schema

$$\begin{aligned} d(\{\mathbf{x}\}, \{\mathbf{y}\}) &= d(\mathbf{x}, \mathbf{y}) \\ d(A_1 \uplus A_2, B) &= \alpha_1 \cdot d(A_1, B) + \alpha_2 \cdot d(A_2, B) + \beta \cdot d(A_1, A_2) \\ &\quad + \gamma \cdot |d(A_1, B) - d(A_2, B)| \end{aligned}$$

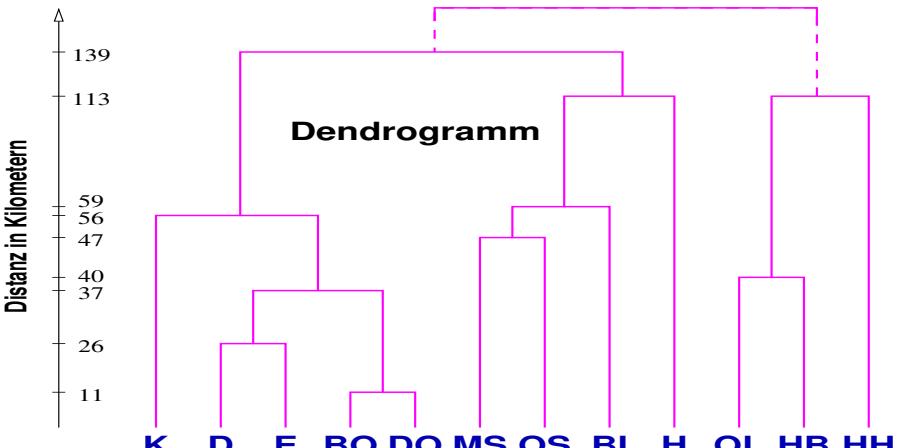
so heißt diese Vorschrift **Lance-Williams-Rekursion** mit den reellwertigen Parametern $\alpha_1 \geq 0$, $\alpha_2 \geq 0$, β und γ .

Bemerkung

Die drei X-Linkage-Funktionen besitzen alle die Lance-Williams-Gestalt:

1. Single-Linkage: $\alpha_1 = \alpha_2 = \frac{1}{2}$, $\beta = 0$, $\gamma = -\frac{1}{2}$
2. Complete-Linkage: $\alpha_1 = \alpha_2 = \frac{1}{2}$, $\beta = 0$, $\gamma = +\frac{1}{2}$
3. Average-Linkage: $\alpha_1 = \frac{|A_1|}{|A_1|+|A_2|}$, $\alpha_2 = \frac{|A_2|}{|A_1|+|A_2|}$, $\beta = 0$, $\gamma = 0$

Beispiel — Dendrogramm für Städtedistanzen



Strenge Monotonie

Je später zwei Gruppen im agglomerativen Clusteringalgorithmus vereinigt werden, desto größer ist ihre Mengendistanz.
(Nichtmonotonie \rightsquigarrow Inversionen des Dendrogramms)

Weitere Distanzfunktionen

Simple-Average

Keine globale Semantik, aber schwach monoton und Δ -invariant:

$$d_{SA}(A_1 \uplus A_2, B) \stackrel{\text{def}}{=} \frac{1}{2} \cdot d_{SA}(A_1, B) + \frac{1}{2} \cdot d_{SA}(A_2, B)$$

Lance-Williams-Parameter: $\alpha_1 = \alpha_2 = \frac{1}{2}$, $\beta = 0$, $\gamma = 0$

Zentroid-Verfahren

Für numerische Attribute; weder schwach monoton noch Δ -invariant:

$$d_{ZEN}(A, B) \stackrel{\text{def}}{=} \|\mu(A) - \mu(B)\|^2$$

Lance-Williams-Parameter: $\alpha_1 = \frac{|A_1|}{|A_1|+|A_2|}$, $\alpha_2 = \frac{|A_2|}{|A_1|+|A_2|}$, $\beta = -\alpha_1\alpha_2$, $\gamma = 0$

Median/Gower-Verfahren

Wie Zentroid; ignoriert aber die relativen Größen vereinigter Gruppen:

$$\alpha_1 = \alpha_2 = \frac{1}{2}, \quad \beta = -\frac{1}{4}, \quad \gamma = 0$$

Ward-Verfahren

Ähnelt der Zentroiddistanz · Garantiert aber Distanzmonotonie

Ward-Zielgröße

Das globale Clusterverzerrungsmaß

$$\varepsilon_{WARD}(\{\omega_1, \dots, \omega_K\}) \stackrel{\text{def}}{=} \sum_{\lambda=1}^K \sum_{x \in \omega_\lambda} \|x - \mu_\lambda\|^2, \quad \mu_\lambda = \mu(\omega_\lambda)$$

führt auf den **Heterogenitätszuwachs**

$$d_{WARD}(A, B) = \varepsilon' - \varepsilon = \frac{|A| \cdot |B|}{|A| + |B|} \cdot \|\mu(A) - \mu(B)\|^2$$

bei Vereinigung der Gruppen $A = \omega_\kappa$ und $B = \omega_\lambda$ und diese Formel wiederum auf eine Lance-Williams-Darstellung:

$$d_{WARD}(A_1 + A_2, B) = \frac{(|A_1| + |B|) \cdot d(A_1, B) + (|A_2| + |B|) \cdot d(A_2, B) - |B| \cdot d(A_1, A_2)}{|A_1| + |A_2| + |B|}$$

Divisive Gruppierung

Generischer Top-down-Algorithmus

(Algorithmus)

Gegeben sind die Datenobjekte $x_1, \dots, x_T \in \Omega$.

1 INITIALISIERUNG

Starte mit $K = 1$ Gruppe(n) $\omega_1 = \{x_1, \dots, x_T\}$.

2 HETEROGENITÄTSKRITERIUM

Berechne für alle $1 \leq \kappa \leq K$ die Gruppenheterogenität

$$H_\kappa \stackrel{\text{def}}{=} d(\omega_\kappa).$$

3 AUFPALTUNG

Zerlege diejenige Gruppe ω_{κ^*} mit

$$\kappa^* = \underset{\kappa}{\operatorname{argmax}} H_\kappa$$

in zwei disjunkte Teilgruppen (z.B. via Austauschverfahren).

4 TERMINIERUNG

Wenn $K = T$, dann Ende, sonst \rightsquigarrow 2.

(zumtinaogA)

Heterogenitätskriterien

Gruppendurchmesser

$$d_{DIAM}(\omega) \stackrel{\text{def}}{=} \max_{x, y \in \omega} d(x, y)$$

Mittlere Innergruppenspanne

$$d_{AD}(\omega) \stackrel{\text{def}}{=} \frac{1}{|\omega|^2 - |\omega|} \sum_{x, y \in \omega} d(x, y)$$

Empirische Gruppenvarianz

$$d_{VAR}(\omega) \stackrel{\text{def}}{=} \frac{1}{|\omega|} \sum_{x \in \omega} \|x - \mu(\omega)\|^2 = \operatorname{spur}(\mathbf{S}(\omega))$$

Gaußäquivalente Entropie

$$d_{CE}(\omega) \stackrel{\text{def}}{=} -\frac{2}{|\omega|} \cdot \log \mathcal{N}(\omega \mid \mu(\omega), \mathbf{S}(\omega)) = \text{const} + \log \det \mathbf{S}(\omega)$$

Hierarchische Gruppierung

Divisive Gruppierung $\hat{=}$ Top-down-Induktion

Kontrollflußregelung durch Heterogenitätsmaß; $O(T \cdot n_{\text{split}})$
Polythetische Verzweigungsfragen · extensionale Zerlegung

Blinde Gruppierung

Keine Heterogenitätsprüfung
Balancierte Aufspaltung in 2^b Gruppen

Agglomerative Gruppierung $\hat{=}$ Bottom-up-Iteration

Gierige Verschmelzung mit Aufwand $O(T \cdot T^2)$

ISODATA-Algorithmus

„Split+merge“-Strategie
Pulsierende Folge von Teilungen & Verschmelzungen

Welches ist die beste Gruppierungsstufe ?

Wähle die „richtige“ Clusteranzahl $K \in \{1, 2, \dots, T\}$

Beispiel — Agrarnationen der EU (1993)

Datensatz 'agriculture' (cluster)

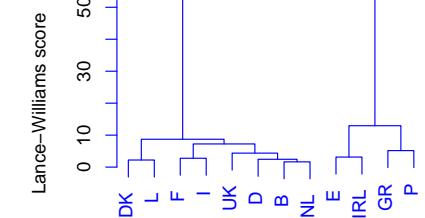
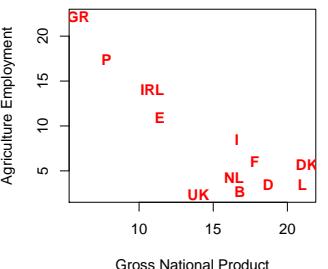
12 europäische Länder

Attribut x_1 = Bruttosozialprodukt der Hauptstadt

Attribut x_2 = Bevölkerungsanteil (%) in landwirtschaftlicher Anstellung

	B	DK	D	GR	E	F	IRL	I	L	NL	P	UK
x_1	16.8	21.3	18.7	5.9	11.4	17.8	10.9	16.6	21.0	16.4	7.8	14.0
x_2	2.7	5.7	3.5	22.2	10.9	6.0	14.0	8.5	3.5	4.3	17.4	2.3

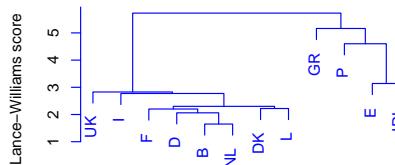
Ward-Distanz



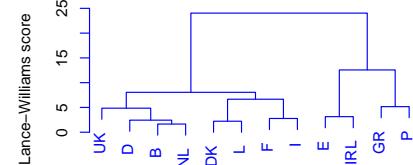
Beispiel — Agrarnationen der EU (1993)

Vergleich unterschiedlicher Lance-Williams-Distanzen

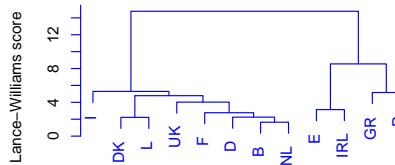
single-Distanz



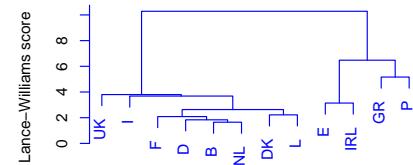
complete-Distanz



average-Distanz



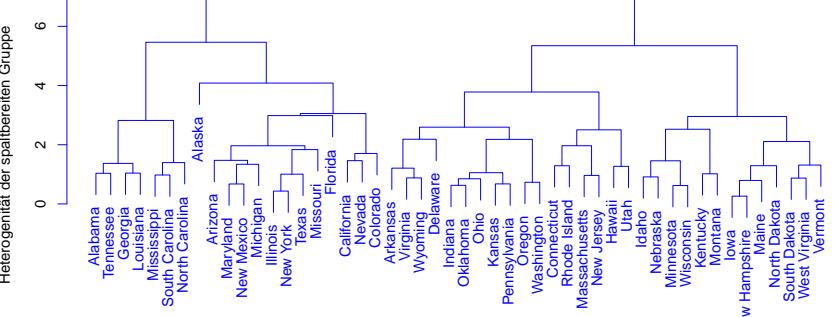
centroid-Distanz



Beispiel — Verbrechensstatistik

Divisive Gruppierung mit 'diana'/R

diana (USArrests, metric='euclidean', stand=TRUE)



Datensatz 'USArrests' (datasets)

50 Objekte: Kriminalstatistiken aller US-Bundesstaaten (1973)

3 Attribute: „Murder“, „Assault“, „Rape“ (Anzahl je 10^5 Einwohner) und

1 Attribut: „UrbanPop“ (Prozentsatz Stadtbevölkerung)

Hierarchische Gruppierung: agglomerativ/divisiv

Austauschverfahren: (fuzzy) K-means

Mischungsidentifikation

Relationale Gruppierung

Konzeptuelle Gruppierung

Spektrale Gruppierung

Clustergütemaße

Zusammenfassung

Permutationsverfahren

Gieriges Suchverfahren · extensional · alle Metriken

(Algorithmus)

1 INITIALISIERUNG

Wähle eine Startpartition $\omega_1, \dots, \omega_K$ mit vorgegebenen $|\omega_\kappa| = T_\kappa$.

2 VERZERRUNGSDIFFERENZEN

Berechne für alle $\mathbf{x} \in \omega_\kappa$ und $\mathbf{y} \in \omega_\lambda$ mit $\kappa \neq \lambda$

$$\Delta\varepsilon(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \varepsilon(\{\dots, \underbrace{\omega'_\kappa, \omega'_\lambda, \dots}\}) - \varepsilon(\{\omega_1, \dots, \omega_K\}).$$

$\mathbf{x} \leftrightarrow \mathbf{y}$

3 VERTAUSCHUNG

Vertausche innerhalb der aktuellen Partition das Datenvektorpaar

$$(\mathbf{x}^*, \mathbf{y}^*) = \operatorname{argmin} \Delta\varepsilon(\mathbf{x}, \mathbf{y}).$$

4 TERMINIERUNG

Wenn $\varepsilon \leq \theta$ dann Ende sonst \rightsquigarrow 2.

(zumdinogA)

Scharfe Gruppierung

bei vorgegebener Gruppenanzahl $K \in \mathbb{N}$

GESUCHT

ist eine K -Partition des Datensatzes $\omega \subset \Omega$.

- **extensional:** Teilmengensystem $\omega_1 \uplus \omega_2 \uplus \dots \uplus \omega_K = \omega$
- **intensional:** Gruppenprototypen $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K \in \Omega$

Verzerrung einer Gruppe

hinsichtlich einer Objektraummetrik $d : \Omega \times \Omega \rightarrow \mathbb{R}$:

$$\varepsilon(\omega_\kappa) \stackrel{\text{def}}{=} \sum_{\mathbf{x} \in \omega_\kappa} d(\mathbf{x}, \mu(\omega_\kappa)), \quad \kappa = 1, \dots, K$$

Verzerrung einer Partition

$$\varepsilon(\{\omega_1, \dots, \omega_K\}) \stackrel{\text{def}}{=} \sum_{\kappa=1}^K \varepsilon(\omega_\kappa)$$

⇒ Kombinatorische Optimierungsaufgabe

Intensionale Gruppierung

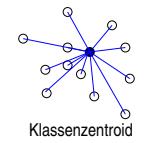
Gruppierung mit Prototypen — „Vektorquantisierung“

Lemma

Es sei $\omega_1, \dots, \omega_K$ eine Gruppierung der Elemente $\mathbf{x}_1, \dots, \mathbf{x}_T$ des metrischen Raumes (Ω, d) , welche die globale Verzerrung minimiert. Dann gibt es Gruppenprototypen $\mathbf{z}_1, \dots, \mathbf{z}_K \in \Omega$ mit

1. Jedes \mathbf{z}_κ ist Zentroid seiner Gruppe ω_κ :

$$\mathbf{z}_\kappa = \operatorname{argmin}_{\mathbf{y} \in \Omega} \sum_{\mathbf{x} \in \omega_\kappa} d(\mathbf{x}, \mathbf{y})$$



2. Jeder Datenvektor \mathbf{x}_t , $t = 1, \dots, T$ gehört zu der Gruppe des nächstliegenden Prototypen:

$$\mathbf{x}_t \in \omega_\kappa \Rightarrow d(\mathbf{x}_t, \mathbf{z}_\kappa) = \min_\lambda d(\mathbf{x}_t, \mathbf{z}_\lambda)$$



Für die euklidische Distanz gilt natürlich $\mathbf{z}_\kappa = \mu(\omega_\kappa)$ für alle $\kappa = 1, \dots, K$.

Stapelweiser K-means-Algorithmus

Lloyd 1957 · Forgy 1965

(Algorithmus)

1 INITIALISIERUNG

Wähle eine zufällige Startpartition

$$\omega_1 \uplus \omega_2 \uplus \dots \uplus \omega_K = \omega$$

2 REPRÄSENTATION

Berechne alle neuen Prototypen

$$z_\kappa = \mu_{ZEN}(\omega_\kappa) \stackrel{\text{def}}{=} \operatorname{argmin}_{y \in \Omega} \sum_{x \in \omega_\kappa} d(x, y)$$

3 REKLASSIFIKATION

Berechne alle neuen Gruppen

$$\omega_\kappa = \left\{ x_t \in \omega \mid \operatorname{argmin}_\lambda d(x_t, z_\lambda) = \kappa \right\}$$

4 TERMINIERUNG

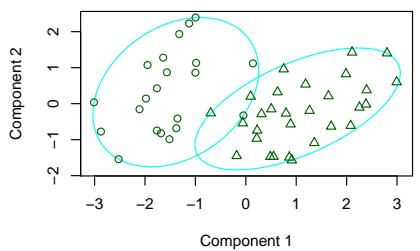
Wenn $\varepsilon(\{\omega_1, \dots, \omega_K\}) \leq \theta$ dann Ende sonst \rightsquigarrow 2.

(zumDingA)

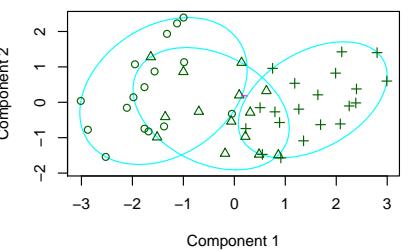
Beispiel — 'USArrests'-Datensatz

K-medoids-Algorithmus minimiert $\|\cdot\|^1$ -Summe · robuster als K-means

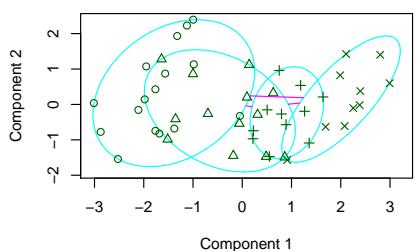
`clusplot(pam(x = USArrests, k = 2))`



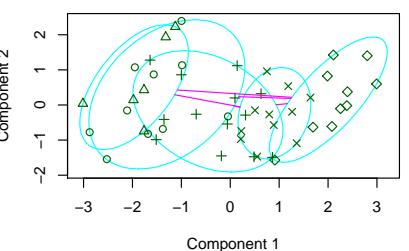
`clusplot(pam(x = USArrests, k = 3))`



`clusplot(pam(x = USArrests, k = 4))`



`clusplot(pam(x = USArrests, k = 5))`



Inkrementeller K-means-Algorithmus

MacQueen 1967

(Algorithmus)

1 INITIALISIERUNG

Wähle zufällige Startprototypen $\{z_1, \dots, z_K\}$, setze $t \leftarrow 1$.

2 REKLASSIFIKATION

Wähle $y = x_{t \bmod T}$ und setze $\kappa = \operatorname{argmin}_\lambda d(y, z_\lambda)$.

3 REPRÄSENTATION

Verschiebe $z_\kappa \leftarrow \alpha_t \cdot y + (1 - \alpha_t) \cdot z_\kappa$.

4 TERMINIERUNG

Wenn $\varepsilon(\cdot) \leq \theta$ dann Ende sonst $t \leftarrow t + 1$ und \rightsquigarrow 2.

(zumDingA)

Bemerkungen

1. Die Gewinnerprototypen z_κ werden nach jedem Einzelschritt aktualisiert.
2. Die Datenprobe wird zyklisch oder randomisiert durchlaufen.
3. Distanz $d(x, y) = \|x - y\|^2 \Rightarrow$ Zentroid $\hat{=} \text{Mittelwert}$.
4. Mittelungsgewichte exponentiell ($\alpha_t \equiv \alpha_0$) oder kumulativ ($\alpha_t = \frac{1}{|\omega_\kappa|}$).
5. Schnellerer Abstieg — aber Oszillationsgefahr!

Unscharfe Gruppierung

GESUCHT

ist eine **Zugehörigkeitsfunktion** für den Datensatz $\omega \subset \Omega$:

$$\mathbf{u} : \begin{cases} \omega & \rightarrow [0, 1]^K \\ x_t & \mapsto \{u_{\kappa, t}\}_{\kappa=1}^K \end{cases}, \quad \sum_{\kappa} u_{\kappa, t} = 1 \ (\forall t)$$

Fuzzy K-means Zielgröße

Distanzfunktion ist (hier) der quadrierte euklidische Abstand:

$$\varepsilon(\{u_\kappa\}, \{z_\kappa\}) = \sum_{\kappa=1}^K \sum_{t=1}^T (u_\kappa(x_t))^{\alpha} \cdot \|x_t - z_\kappa\|^2, \quad \alpha \geq 1$$

Opt. Prototypen/Zugehörigk.

Normierung \rightsquigarrow Lagrangemultiplikatoren

$$\sum_{x \in \omega} \beta_x \cdot \left(\sum_{\kappa=1}^K u_\kappa(x) - 1 \right)$$

Spezialfälle

$$\left\{ \begin{array}{l} \alpha = 1: \text{scharfe Datenmengen} \\ \alpha = 2: \text{unscharfe Datenmengen} \\ \alpha = \infty: \text{identische Gruppen} \end{array} \right\}$$

Fuzzy K-means-Algorithmus

1 INITIALISIERUNG

Wähle zufällige Startzugehörigkeiten $u_{\kappa,t} \in [0, 1]$.

2 PROTOTYPEN

Für alle $1 \leq \kappa \leq K$ berechne

$$\mathbf{z}_\kappa = \sum_{x \in \omega} (u_\kappa(x))^\alpha \cdot \mathbf{x} \Bigg/ \sum_{x \in \omega} (u_\kappa(x))^\alpha$$

3 ZUGEHÖRIGKEITEN

Für alle $1 \leq \kappa \leq K$ und $\mathbf{x} \in \omega$ berechne neue unscharfe Gruppen:

$$u_\kappa(\mathbf{x}) = 1 \Bigg/ \sum_{\lambda=1}^K \left(\frac{\|\mathbf{x} - \mathbf{z}_\kappa\|^2}{\|\mathbf{x} - \mathbf{z}_\lambda\|^2} \right)^{\frac{1}{\alpha-1}}$$

4 TERMINIERUNG

Wenn $\varepsilon \leq \theta$ dann Ende sonst \rightsquigarrow 2.

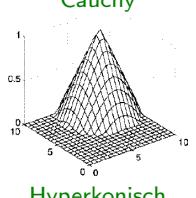
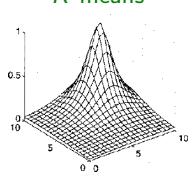
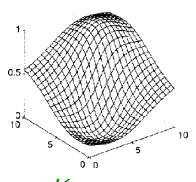
Harmonische Zugehörigkeitsfunktion

Fuzzy K-means: $u_\kappa(\mathbf{x}) \propto \|\mathbf{x} - \mathbf{z}_\kappa\|^{-2/(\alpha-1)}$

Cauchy Zugehörigkeitsfunktion

mit den Halbwertsbreiten $\eta_\kappa > 0$:

$$u_\kappa(\mathbf{x}) \stackrel{\text{def}}{=} 1 \Bigg/ \left(1 + \left(\frac{\|\mathbf{x} - \mathbf{z}_\kappa\|^2}{\eta_\kappa} \right)^{\frac{1}{\alpha-1}} \right)$$



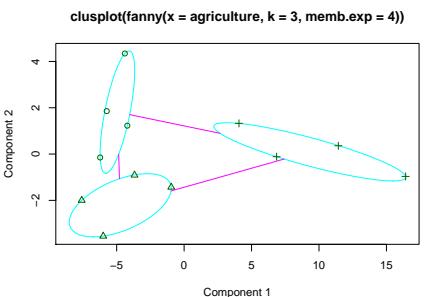
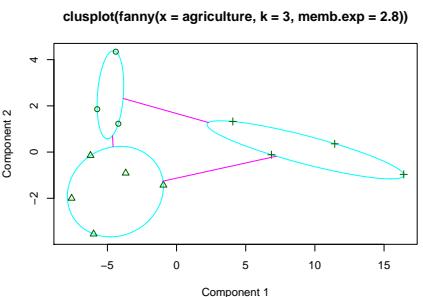
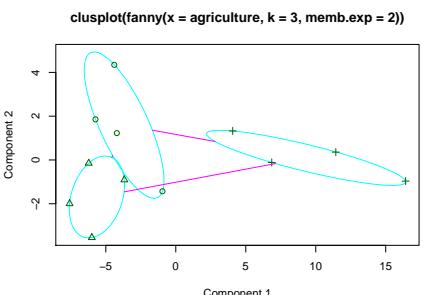
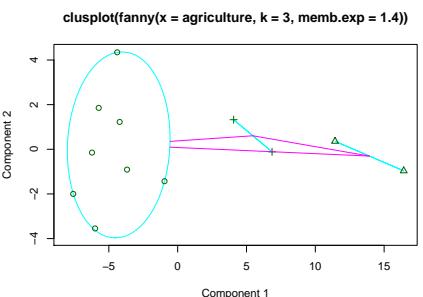
Hyperkonische Zugehörigkeitsfunktion

mit den Radien $r_\kappa > 0$:

$$u_\kappa(\mathbf{x}) \stackrel{\text{def}}{=} \begin{cases} 1 - \|\mathbf{x} - \mathbf{z}_\kappa\|/r_\kappa & \text{falls } \|\mathbf{x} - \mathbf{z}_\kappa\| \leq r_\kappa \\ 0 & \text{sonst} \end{cases}$$

Beispiel — 'agriculture'-Datensatz

Fuzzy K-means-Algorithmus ($\alpha \in \{\sqrt{2}^i \mid i = 1, 2, 3, 4\}$)



Geometrische Clusterformen

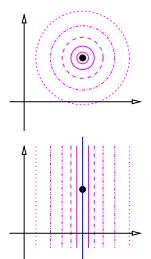
Perpendikulare Linien im \mathbb{R}^N

Punktförmiges Zentrum im \mathbb{R}^2

$$d^2(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|^2 = (x_1 - z_1)^2 + (x_2 - z_2)^2$$

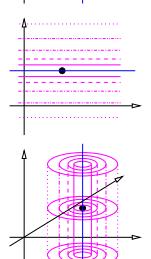
Vertikal linienförmig im \mathbb{R}^2

$$d^2(\mathbf{x}, \mathbf{z}) = (x_1 - z_1)^2$$



Horizontal linienförmig im \mathbb{R}^2

$$d^2(\mathbf{x}, \mathbf{z}) = (x_2 - z_2)^2$$



Vertikal linienförmig im \mathbb{R}^3

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{z}) &= (x_1 - z_1)^2 + (x_2 - z_2)^2 \\ &= \|\mathbf{x} - \mathbf{z}\|^2 - (x_3 - z_3)^2 \end{aligned}$$

Perpendikuläre Linienzentren

Unendlich lange, koordinatenachsenparallele Cluster

Linienförmiges Klassenzentrum

des \mathbb{R}^N in Richtung der x_n -Achse, $n \in \{1, \dots, N\}$:

$$\begin{aligned} d^2(\mathbf{x} | \mathbf{z}, n) &= \|\mathbf{x} - \mathbf{z}\|^2 - (x_n - z_n)^2 \\ &= \|\mathbf{x} - \mathbf{z}\|^2 - (\mathbf{e}_n^\top \cdot (\mathbf{x} - \mathbf{z}))^2 \\ &= \|\mathbf{x} - \mathbf{z}\|^2 - \|\mathbf{e}_n^\top \cdot (\mathbf{x} - \mathbf{z})\|^2 = d^2(\mathbf{x} | \mathbf{z}, \mathbf{e}_n) \end{aligned}$$

Vom euklidischen Abstand wird also die Norm einer Achsenprojektion subtrahiert.

Verallgemeinerung auf „schräge“ Cluster ?

Es ist naheliegend, daß dieser Zusammenhang auch für den nichtperpendikularen Fall gilt.

Achsenrotation

Datentransformation $\phi: \mathbf{x} \mapsto \mathbf{U}^\top \mathbf{x}$, $\mathbf{U}^\top \mathbf{U} = \mathbf{U} \mathbf{U}^\top = \mathbf{E}$

Rotationen sind distanzinvariant

$$\|\mathbf{x}\|^2 = \mathbf{x}^\top \mathbf{x} = \mathbf{x}^\top \mathbf{E} \mathbf{x} = \mathbf{x}^\top \mathbf{U} \mathbf{U}^\top \mathbf{x} = \|\mathbf{U}^\top \mathbf{x}\|^2 = \sum_{n=1}^N (\mathbf{u}_n^\top \mathbf{x})^2$$

Summendarstellung mit den Spaltenvektoren $\mathbf{u}_1, \dots, \mathbf{u}_N$ von \mathbf{U} .

Linienbezogener Abstand in \mathbf{u}_n -Richtung

$$\begin{aligned} d^2(\mathbf{x} | \mathbf{z}, \mathbf{u}_n) &= \|\phi \mathbf{x} - \phi \mathbf{z}\|^2 - ((\phi \mathbf{x})_n - (\phi \mathbf{z})_n)^2 \\ &= \|\phi(\mathbf{x} - \mathbf{z})\|^2 - (\mathbf{u}_n^\top \mathbf{x} - \mathbf{u}_n^\top \mathbf{z})^2 \\ &= \|\mathbf{x} - \mathbf{z}\|^2 - (\mathbf{u}_n^\top (\mathbf{x} - \mathbf{z}))^2 \end{aligned}$$

Flächenbezogener Abstand in \mathbf{u}, \mathbf{v} -Richtung

(Richtungsvektoren \mathbf{u}, \mathbf{v} normiert und senkrecht zueinander)

$$d^2(\mathbf{x} | \mathbf{z}, \mathbf{u}, \mathbf{v}) = \|\mathbf{x} - \mathbf{z}\|^2 - (\mathbf{u}^\top (\mathbf{x} - \mathbf{z}))^2 - (\mathbf{v}^\top (\mathbf{x} - \mathbf{z}))^2$$

M -dimensionale Hyperflächenzentren

Satz (Pythagoras)

Sei $0 \leq M \leq N$. Für alle $\mathbf{x} \in \mathbb{R}^N$ berechnet sich der lotrechte Abstand zwischen \mathbf{x} und der M -dimensionalen Hyperfläche mit dem

Aufpunktvektor \mathbf{z} und den orthonormalen Richtungsvektoren

$\mathbf{u}_1, \dots, \mathbf{u}_M \in \mathbb{R}^N$ gemäß

$$\min_{a_1, \dots, a_M} \left\| \mathbf{x} - \left(\mathbf{z} + \sum_{m=1}^M a_m \mathbf{u}_m \right) \right\|^2 = \|\mathbf{x} - \mathbf{z}\|^2 - \|\mathbf{U}^\top (\mathbf{x} - \mathbf{z})\|^2,$$

wenn $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_M)$ ist.

Beweis.

Der Abstand für ein M -dimensionales Zentrum, das durch die orthonormalen Vektoren $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_M) \in \mathbb{R}^{N \times M}$ aufgespannt wird:

$$d^2(\mathbf{x} | \mathbf{z}, \mathbf{U}) = \|\mathbf{x} - \mathbf{z}\|^2 - \sum_{m=1}^M (\mathbf{u}_m^\top (\mathbf{x} - \mathbf{z}))^2 = \|\mathbf{x} - \mathbf{z}\|^2 - \|\mathbf{U}^\top (\mathbf{x} - \mathbf{z})\|^2$$

□

Fuzzy K -Varieties

Definition

Das (unscharfe) Gruppierungsverfahren mit der Zielgröße

$$\varepsilon(\{\omega_\kappa\}) = \sum_{\kappa=1}^K \sum_{\mathbf{x} \in \omega} u_\kappa(\mathbf{x})^\alpha \cdot d^2(\mathbf{x} | \mathbf{z}_\kappa, \mathbf{U}_\kappa)$$

heißt **fuzzy K -varieties**-Algorithmus; im Spezialfall $M = 1$ heißt es **fuzzy K -lines**-Algorithmus.

Elliptotypzentren

(„fuzzy K -elliptotypes“)

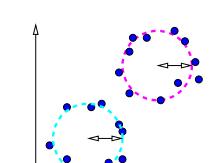
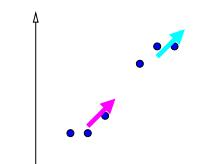
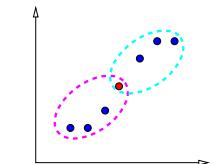
$$d^2(\mathbf{x} | \mathbf{z}, \mathbf{U}) = \|\mathbf{x} - \mathbf{z}\|^2 - \rho \cdot \|\mathbf{U}^\top (\mathbf{x} - \mathbf{z})\|^2$$

Spezialfälle: $\rho = 0$ Punktzentrum · $\rho = 1$ Hyperflächenzentrum

Hyperkugelschalen

(„fuzzy K -shells“)

$$d^2(\mathbf{x} | \mathbf{z}, r) = (\|\mathbf{x} - \mathbf{z}\| - r)^2$$



Gradientenabstieg für Fuzzy K-Varieties

Lemma

Die Minimierung der Zielgröße mit Lagrangemultiplikatoren für die Normierungsbedingungen liefert die Bestimmungsgleichungen

$$\mathbf{z}_\kappa = \cdot \sum_{\mathbf{x} \in \omega} (u_\kappa(\mathbf{x}))^\alpha \cdot \mathbf{x} \Bigg/ \sum_{\mathbf{x} \in \omega} (u_\kappa(\mathbf{x}))^\alpha$$

für die **Aufpunktvektoren**,

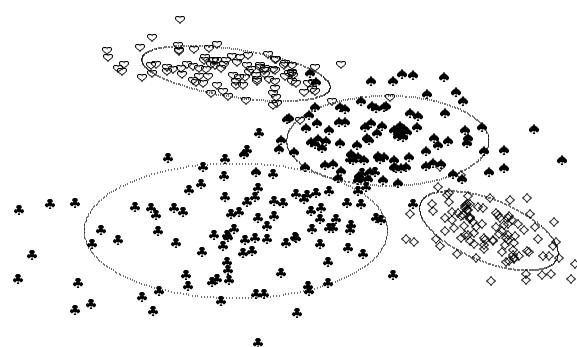
$$u_\kappa(\mathbf{x}) = 1 \Bigg/ \sum_{\lambda=1}^K \left(\frac{d^2(\mathbf{x} | \mathbf{z}_\kappa, \mathbf{U}_\kappa)}{d^2(\mathbf{x} | \mathbf{z}_\lambda, \mathbf{U}_\lambda)} \right)^{\frac{1}{\alpha-1}}$$

für die **Gruppenzugehörigkeiten** und für die **Gruppenkovarianzen**

$$\mathbf{S}_\kappa = \sum_{\mathbf{x} \in \omega} u_\kappa(\mathbf{x})^\alpha (\mathbf{x} - \mathbf{z}_\kappa)(\mathbf{x} - \mathbf{z}_\kappa)^\top.$$

Die m -te Spalte $\mathbf{u}_{\kappa,m}$ von \mathbf{U}_κ schließlich ergibt sich als Eigenvektor zum m -größten Eigenwert von \mathbf{S}_κ .

Identifikation von Mischverteilungen



Problem

Angenommen, obige Daten sind gemäß $f(\mathbf{x}) = \sum_{\kappa=1}^K \pi_\kappa \cdot g(\mathbf{x} | \boldsymbol{\theta}_\kappa)$ mischverteilt. Wie lauten die **bestpassenden** (ML) Verteilungsparameter $\hat{\pi}_\kappa, \hat{\boldsymbol{\theta}}_\kappa$, $\kappa = 1, \dots, K$ des Modells?

Lösung

Im Normalverteilungsfall $g(\mathbf{x} | \boldsymbol{\theta}_\kappa) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_\kappa, \mathbf{S}_\kappa)$ existiert eine **asymptotisch eindeutige Lösung sowie ein lokales Optimierungsverfahren** (EM).

Hierarchische Gruppierung: agglomerativ/divisiv

Austauschverfahren: (fuzzy) K-means

Mischungsidentifikation

Relationale Gruppierung

Konzeptuelle Gruppierung

Spektrale Gruppierung

Clustergütemaße

Zusammenfassung

EM-Algorithmus

zur Identifikation gaußscher Mischverteilungen

1 INITIALISIERUNG

Wähle zufällige Startparameter $(\pi_\kappa, \boldsymbol{\mu}_\kappa, \mathbf{S}_\kappa)$, $\kappa = 1, \dots, K$.

2 ERWARTUNGSWERTE

Berechne für $\kappa = 1..K$ und $t = 1..T$ die a posteriori Wahrsch'keiten

$$\gamma_{\kappa,t} \stackrel{\text{def}}{=} P(\Omega_\kappa | \mathbf{x}_t) = \frac{P(\Omega_\kappa) \cdot P(\mathbf{x}_t | \Omega_\kappa)}{P(\mathbf{x}_t)} \propto \pi_\kappa \cdot \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_\kappa, \mathbf{S}_\kappa)$$

3 MAXIMIERUNG

$$\pi_\kappa \leftarrow \frac{\sum_t \gamma_{\kappa,t}}{\sum_\lambda \sum_t \gamma_{\lambda,t}}, \quad \boldsymbol{\mu}_\kappa \leftarrow \frac{\sum_t \gamma_{\kappa,t} \mathbf{x}_t}{\sum_t \gamma_{\kappa,t}}, \quad \mathbf{S}_\kappa \leftarrow \frac{\sum_t \gamma_{\kappa,t} \mathbf{x}_t \mathbf{x}_t^\top}{\sum_t \gamma_{\kappa,t}} - \boldsymbol{\mu}_\kappa \boldsymbol{\mu}_\kappa^\top$$

4 TERMINIERUNG

Wenn die ML-Zielgröße $\ell(\dots)$ stagniert dann Ende sonst \rightsquigarrow 2.

Konvergenzeigenschaften

des EM-Algorithmus für Gaußsche Mischverteilungsmodelle (GMM)

1. Schwache Monotonie

Verfahren erreicht stationären Punkt

$$\ell(\theta_0) \leq \ell(\theta_1) \leq \ell(\theta_2) \leq \ell(\theta_3) \leq \dots \leq \ell(\theta_j) \leq \dots \leq \dots$$

2. Beschränktheit

pathologische Aufgabenstellung („ill-posed problem“)

$$\mathcal{N}(\mu_\lambda, \mathbf{S}_\lambda) = \mathcal{N}(\mathbf{x}_t, \mathbf{0})$$

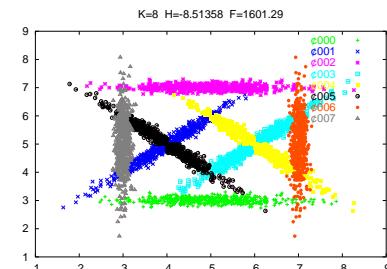
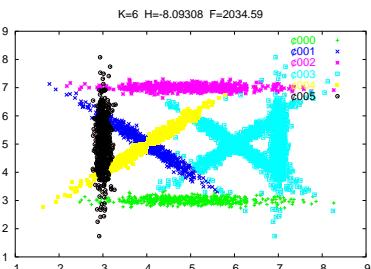
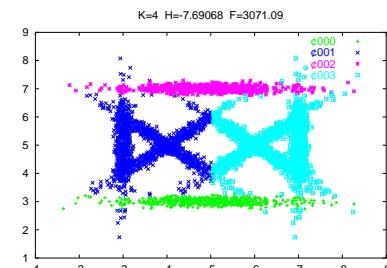
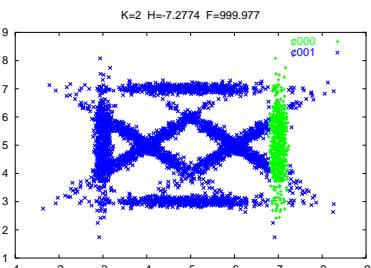
3. Lokale Maxima

viele relative Maxima mit $\ell(\theta) < \infty$ und großem Einzugsbereich

4. Zyklischer Iterationsverlauf

Kraterrandphänomen

$$\theta_1 \neq \theta_2 \neq \dots \neq \theta_m \quad \text{mit} \quad \ell(\theta_1) = \ell(\theta_2) = \dots = \ell(\theta_m)$$



Problematik des Rangdefizits

$$\text{rg}(\mathbf{S}_\kappa) < N \Rightarrow \det(\mathbf{S}_\kappa) = 0 \Rightarrow \mathbf{S}_\kappa^{-1} = ?$$

Gratregularisierung

Anisotropes Aufblasen der Konzentrationsellipse („Speckschicht“)

$$\mathbf{S}^{(\delta)} \stackrel{\text{def}}{=} \mathbf{S} + \delta \cdot \mathbf{E} = \begin{pmatrix} s_{11} + \delta & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} + \delta & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} + \delta \end{pmatrix}$$

Fixierung & Verklebung

Alle Eigenschaften des EM-Algorithmus bleiben erhalten:

- Kovarianzmatrizen fixieren ($\forall \kappa : \mathbf{S}_\kappa \stackrel{!}{=} \mathbf{S}^*$) \rightsquigarrow keine pathologische Lösung
- Kovarianzmatrizen verkleben ($\forall \kappa, \lambda : \mathbf{S}_\kappa \stackrel{!}{=} \mathbf{S}_\lambda$) \rightsquigarrow mehr Robustheit

Hintergrundkomponente

Streuungsintensives Rückweisungscluster zur Ausreißerbehandlung

$$f_0(\cdot) = \mathcal{N}(\cdot | \mu(\omega), \mathbf{S}_0) \quad \text{mit} \quad \mathbf{S}_0 = \mathbf{S}(\omega) \text{ oder} \quad \mathbf{S}_0 = \sigma_0^2 \cdot \mathbf{E}$$

Probabilistische PCA

Zerlegung des \mathbb{R}^N in systematisch und in zufällig streuende Komponenten

Normalverteilungsmodelle für rangdefizite Daten

Das homogene Faktorenanalysemodell

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{E} + \mathbf{A} \cdot \mathbf{V} \quad \text{mit}$$

$$\left\{ \begin{array}{l} \boldsymbol{\mu} \in \mathbb{R}^N \\ \mathbf{A} \in \mathbb{R}^{N \times M} \\ \mathbf{E} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{E}_N) \\ \mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{E}_M) \end{array} \right.$$

(Dimension $M \leq N$; PCA-Annahme identischer Störvarianzen)

- besitzt als Ladungsvektoren die bereits hinlänglich bekannten M führenden Hauptachsen der Verteilungsellipse,
- definiert aber gleichzeitig eine explizite Wahrscheinlichkeitsverteilung für die Daten.

PPCA-Schätzung bei bekanntem Modellrang M

Lemma (Tipping & Bishop 1999)

Der Zufallsvektor des homogenen FA-Modells ist gemäß $\mathbb{X} \sim \mathcal{N}(\mu, \mathbf{S})$ normalverteilt mit der Kovarianzmatrix

$$\mathbf{S} = \mathbf{A}\mathbf{A}^\top + \sigma^2 \cdot \mathbf{E}_N.$$

Der **ML-Schätzer** für \mathbf{S} ergibt sich durch Einsetzen der Schätzwerte

$$\begin{aligned}\hat{\mathbf{A}} &= \mathbf{U}_M \cdot (\mathbf{D}_M - \sigma^2 \cdot \mathbf{E}_M)^{1/2} \\ \hat{\sigma}^2 &= \frac{1}{N-M} \cdot \sum_{j=M+1}^N \lambda_j\end{aligned}$$

mit der $(M : N)$ -eigenzerlegten Datenkovarianzmatrix

$$\hat{\mathbf{S}}(\omega) \stackrel{!}{=} (\mathbf{U}_M, \mathbf{U}'_M) \cdot \left(\begin{array}{c c} \mathbf{D}_M & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{D}'_M \end{array} \right) \cdot (\mathbf{U}_M, \mathbf{U}'_M)^\top, \quad \left(\begin{array}{c c} \mathbf{D}_M & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{D}'_M \end{array} \right) = \text{diag}(\lambda_1, \dots, \lambda_N).$$

Invertierung der PPCA-Kovarianzmatrix

Kovarianzstruktur

Die ursprüngliche Darstellung zeigt eine rangverminderte Darstellung mit einem additiv aufgeprägtem Fehlergrat von σ^2 auf der Diagonalen.

$$\mathbf{S} = \mathbf{A}\mathbf{A}^\top + \sigma^2 \cdot \mathbf{E}_N$$

Die gleichwertige alternative Darstellung präsentiert eine vollständige Eigenzerlegung mit den kanonischen Eigenvektoren und -werten; nur die letzten $(N - M)$ Eigenwerte wurden gemittelt.

$$\mathbf{S} = \mathbf{U}_M \mathbf{D}_M \mathbf{U}_M^\top + \sigma^2 \cdot \mathbf{U}'_M \mathbf{U}'_M^\top$$

Inverse Kovarianz

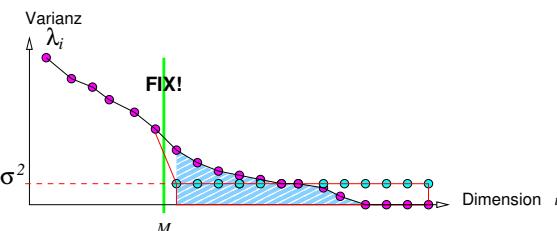
Diese Inversionsformel verwendet ausschließlich die M führenden Eigenvektoren sowie die führenden reziproken Eigenwerte.

$$\mathbf{S}^{-1} = \frac{1}{\sigma^2} \cdot \left\{ \mathbf{E}_N - \mathbf{U}_M \cdot (\mathbf{E}_M - \sigma^2 \cdot \mathbf{D}_M^{-1}) \cdot \mathbf{U}_M^\top \right\}$$

Schätzung der mittleren Reststreuung

unter ausschließlicher Verwendung der $M \ll N$ Hauptachsen

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{N-M} \cdot \text{spur}(\mathbf{D}'_M) \\ &= \frac{1}{N-M} \left\{ \text{spur} \left(\begin{array}{c c} \mathbf{D}_M & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{D}'_M \end{array} \right) - \text{spur}(\mathbf{D}_M) \right\} \\ &= \frac{1}{N-M} \left\{ \text{spur}(\hat{\mathbf{S}}(\omega)) - \text{spur}(\mathbf{D}_M) \right\} \\ &= \frac{1}{N-M} \left\{ \frac{1}{T} \sum_{t=1}^T \|\mathbf{x}_t - \mu\|^2 - \sum_{j=1}^M \lambda_j \right\}\end{aligned}$$



PPCA-Schätzung bei bekannter Störvarianz σ^2

Lemma (Meinicke & Ritter 2000)

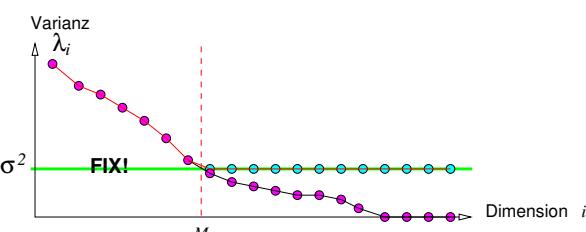
Ein Zufallsvektor sei gemäß $\mathbb{X} \sim \mathcal{N}(\mu, \mathbf{S})$ normalverteilt mit der Kovarianzmatrix

$$\mathbf{S} = \Psi + \sigma^2 \cdot \mathbf{E}_N$$

mit **bekannter** Störvarianz σ^2 und positiv-semidefinitem Ψ mit **unbekanntem** Rang $\nu = \text{ran } \Psi$, $\nu \leq N$.

Mit der empirischen Datenkovarianz $\hat{\mathbf{S}}(\omega)$ und ihrer Eigenzerlegung $\hat{\mathbf{S}}(\omega) = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{U}^\top$, $\mathbf{D} = \text{diag}(\lambda_1, \dots, \lambda_N)$ ergeben sich die ML-Schätzer

$$\hat{\nu} = |\{\lambda_i \mid \lambda_i > \sigma^2\}| \quad \text{und} \quad \hat{\Psi} = \mathbf{U}_\nu \cdot (\mathbf{D}_\nu - \sigma^2 \cdot \mathbf{E}_\nu) \cdot \mathbf{U}_\nu^\top.$$



Effiziente Berechnung der PPCA-Dichtewerte

auch in extrem hochdimensionalen ($N \gg M$) Vektorräumen

Determinante $\det(\mathbf{S})$

Determinanten sind rotationsinvariant ($\det(\mathbf{S}) = \det(\mathbf{U}^\top \mathbf{S} \mathbf{U})$); also gilt:

$$\det(\mathbf{S}) = \prod_{i=1}^N \tilde{\lambda}_i = \sigma^{2 \cdot (N-M)} \cdot \prod_{i=1}^M \lambda_i.$$

Quadratische Form $(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu})$

Wegen der Darstellung von \mathbf{S}^{-1} gilt für die quadratische Form

$$\begin{aligned} \mathbf{y}^\top \mathbf{S}^{-1} \mathbf{y} &= \frac{1}{\sigma^2} \cdot \left\{ \mathbf{y}^\top \mathbf{E}_N \mathbf{y} - \mathbf{y}^\top \mathbf{U}_M \cdot (\mathbf{E}_M - \sigma^2 \mathbf{D}_M^{-1}) \cdot \mathbf{U}_M^\top \mathbf{y} \right\} \\ &= \frac{1}{\sigma^2} \cdot \left\{ \|\mathbf{y}\|^2 - \|\tilde{\mathbf{y}}\|^2 + \sigma^2 \cdot \tilde{\mathbf{y}}^\top \mathbf{D}_M^{-1} \tilde{\mathbf{y}} \right\} \\ &= \frac{\|\mathbf{y}\|^2 - \|\tilde{\mathbf{y}}\|^2}{\sigma^2} + \sum_{i=1}^M \frac{\tilde{y}_i^2}{\lambda_i} \end{aligned}$$

unter Verwendung des Hauptachsenprojektionsvektors

$$\tilde{\mathbf{y}} = \mathbf{U}_M^\top \mathbf{y} = \mathbf{U}_M^\top \cdot (\mathbf{x} - \boldsymbol{\mu}).$$

Zweistufiges EM-Abkühlverfahren

(Algorithmus)

- 1 Vorwahl von $\sigma_{\max}^2 > 0$, $\sigma_{\min}^2 > 0$ und $\alpha \in (0, 1)$.
- 2 Setze $\theta \leftarrow (\boldsymbol{\mu}, \dots, \boldsymbol{\mu})$, $m \leftarrow 0$ und $\sigma_m^2 \leftarrow \sigma_{\max}^2$.
- 3 SPHÄRISCHE GRUPPIERUNG (EM)

$$\ell(\boldsymbol{\theta} | \omega, \sigma_m^2) = \sum_t \sum_{\kappa} \gamma_{\kappa, t} \cdot \log (\pi_{\kappa} \cdot \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_{\kappa}, \sigma_m^2 \mathbf{E})) \xrightarrow{!} \max$$

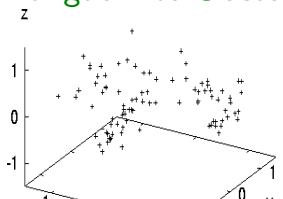
- 4 Setze $m \leftarrow m + 1$ und $\sigma_m^2 \leftarrow \alpha \cdot \sigma_{m-1}^2$.
- 5 Wenn $K_{\text{eff}} < K$ dann \rightsquigarrow 3.
- 6 LOKALADAPTIVE PPCA-GRUPPIERUNG (EM)

$$\mathcal{L}(\boldsymbol{\Theta} | \omega, \sigma_m^2) = \sum_t \sum_{\kappa} \gamma_{\kappa, t} \cdot \log (\pi_{\kappa} \cdot \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_{\kappa}, \boldsymbol{\Psi}_{\kappa} + \sigma_m^2 \mathbf{E})) \xrightarrow{!} \max$$

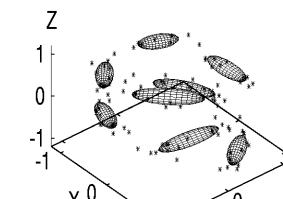
- 7 Setze $m \leftarrow m + 1$ und $\sigma_m^2 \leftarrow \alpha \cdot \sigma_{m-1}^2$.
- 8 Wenn $\sigma_m^2 > \sigma_{\min}^2$ dann \rightsquigarrow 6 sonst Ende.

PPCA-Mischverteilungsidentifikation

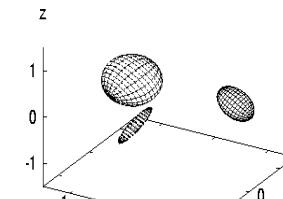
Rangdefizite Cluster



Punktmenge im \mathbb{R}^3



unterschiedliche Richtungen



unterschiedliche Ränge

Sphärische Gruppierung

Alle Gaußkomponenten sphärisch
(kugelförmige Konzentration;
konstante Streuung σ_m^2)

Klassenmittelwertvektoren $\boldsymbol{\mu}_{\kappa}$
Ende sobald Anzahl K_{eff}
Gruppenprototypen gleich
Sollgruppenzahl K ist.

Lokaladaptive Gruppierung

PPCA-Gaußkomponenten
(zeppelinförmige Konzentration;
variable Effektivdimension ν_{κ})

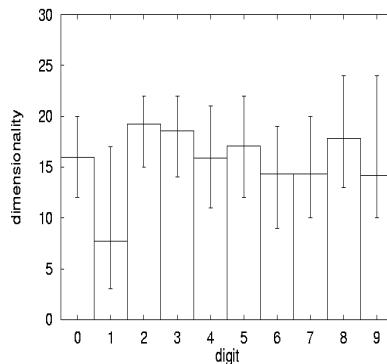
Rangdefiziente Matrizen $\boldsymbol{\Psi}_{\kappa}$
Reststreuung
Ränge (Parameterkomplexität)

Beispiel — Handgeschriebene Ziffern

MNIST Datensammlung, LeCun 1998

Datensatz

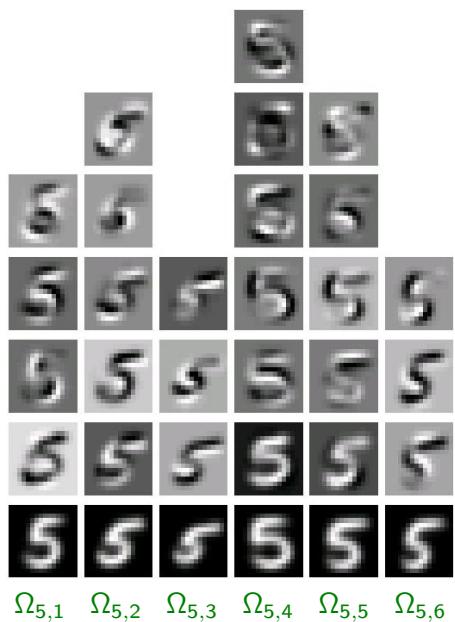
60 000 (10 000) Lern- und Testmuster
Originalziffern 28×28 Pixel zu 8 Bit \rightsquigarrow Umrasterung 8×8
Verschiedene Gruppenstärken $K \in \{1, 2, 4, 8, 16\}$ getestet



PPCA-Dimensionen

alle zehn Ziffernklassen
 ν -Durchschnitt und min/max
 $K = 16$ Mischungskomponenten
im „Gefrierpunkt“ σ_{\min}^2

Beispiel — die Ziffernklaasse „5“



PPCA-Mischung
für Ziffernklaasse Ω_5

Gruppenanzahl
 $K = 6$ gewählt

Modellrang
 $M \in \{4, 5, 6, 7\} \Rightarrow M \ll 64$

Hauptachsen
von unten nach oben
aufgetürmt
als 8×8 -Grauwertbild

Hierarchische Gruppierung: agglomerativ/divisiv

Austauschverfahren: (fuzzy) K-means

Mischungsidentifikation

Relationale Gruppierung

Konzeptuelle Gruppierung

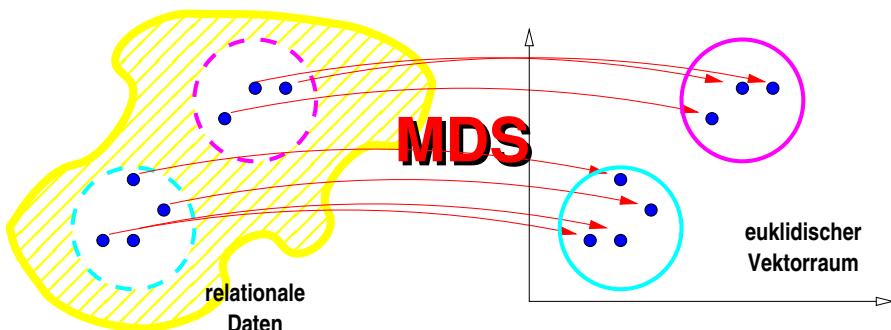
Spektrale Gruppierung

Clustergütemaße

Zusammenfassung

Relationale Gruppierung

Datenobjekte mit wechselseitiger Distanz — ohne Attribute



MDS-Gruppierung

1. Mehrdimensionale Skalierung von ω nach $\omega' = \{\mathbf{x}_1, \dots, \mathbf{x}_T\} \subseteq \mathbb{R}^N$
2. K-means Gruppierung des Datensatzes ω'
3. Aufprägung der ω' -Gruppierung auf die Urbilder ω

Fuzzy K-medoids für Metriken

Datensatz

Objektmenge $\omega = \{o_1, \dots, o_T\}$ mit der Abstandsmatrix

$$\mathbf{R} = [r_{s,t}] \in \mathbb{R}^{T \times T}, \quad r_{s,t} = d(o_s, o_t)$$

Insbesondere gelte die **Symmetrie** $\mathbf{R}^\top = \mathbf{R}$ und die **Definitheit** $\text{diag}(\mathbf{R}) = \mathbf{0}$.

Zugehörigkeitsfunktionen

- **Harmonisch:**

$$u_\kappa(o_t) = 1 / \sum_{\lambda=1}^K \left(\frac{r_{t_\kappa, t}}{r_{t_\lambda, t}} \right)^{\frac{2}{\alpha-1}}$$

- **Cauchy / possibilistisch:**

$$u_\kappa(o_t) = \left(1 + \left(\frac{r_{t_\kappa, t}}{\sqrt{\eta_\kappa}} \right)^{\frac{2}{\alpha-1}} \right)^{-1}$$

- **Hyperkonisch:**

$$u_\kappa(o_t) = \begin{cases} 1 - r_{t_\kappa, t} / \rho_\kappa & \text{falls } r_{t_\kappa, t} \leq \rho_\kappa \\ 0 & \text{sonst} \end{cases}$$

Für harmonische Zugehörigkeiten gilt die Normierungseigenschaft $\sum_\lambda u_\lambda(o_t) = 1$.

RACE — Relationaler Austauschalgorithmus

(Algorithmus)

GEGEBEN

Datenrelation $R \in \mathbb{R}^{T \times T}$, Gruppenzahl $K \in \mathbb{N}$, Iterationen $I \in \mathbb{N}$.

1 INITIALISIERUNG

Setze $i \leftarrow 1$.

Wähle zufällige Prototypenindizes $\{t_1, \dots, t_K\} \subseteq \{1, \dots, T\}$.

2 ITERATIONSSCHRITT

1. Bestimme alle Zugehörigkeiten $u_\kappa(o_t)$
2. Bestimme die „Restenergien“

$$e_{\kappa,t} = \sum_{\lambda \neq \kappa} u_\lambda(o_t)$$

3. Bestimme die neuen Prototypen

$$t_\kappa \leftarrow \operatorname{argmin}_{t=1..T} e_{\kappa,t}$$

3 TERMINIERUNG

Wenn $i = I$ dann \rightsquigarrow Ende sonst $i \leftarrow i + 1, \rightsquigarrow$ 2.

(Enddingabe)

Beispiel — Text Mining

Automatische Erstellung eines Stichwortinventars

Datensammlung

Kapitel 2 aus dem Buch „Information Mining“ (Th. Runkler)

Alle Formeln und Sonderzeichen wurden entfernt.

Großbuchstaben \mapsto Kleinbuchstaben

Objektmenge und Metrik

1605 Wortvorkommen, davon $T = 564$ verschieden

Matrix $R \in \mathbb{R}^{564 \times 564}$ der Levenshteinabstände

Verarbeitung

RACE-Algorithmus mit $K = 20$ Gruppen und $I = KT = 11280$ Schritten

ESS-Defuzzifizierung auf 28 (bzw. 29) Wörter/Gruppe

Die 20 häufigsten Wörter des Textes

die der und für in als werden ist sich mit
oder den sind ein auch daten lässt können abstand wird

Defuzzifizierung

Finales Schärfen (Aushärten) der Gruppenzugehörigkeiten

$$u_\kappa(o_t) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{falls } u_\kappa(o_t) = \max_\lambda u_\lambda(o_t) \\ 0 & \text{sonst} \end{cases}, \quad \kappa = 1, \dots, K$$

Diese Aushärtungsregel ergibt Partitionen mit variablen Gruppenstärken.

ESS-Defuzzifizierung

(equal size subset)

1 INITIALISIERUNG

Setze $\mathcal{I} \leftarrow \{1, 2, \dots, T\}$ (Indexmenge)

Setze $\mathcal{C}_\kappa \leftarrow \emptyset$ (Gruppenindexmenge) für alle $\kappa = 1..K$

2 ITERATION (für alle $t = 1, \dots, T$)

1. Setze Gruppenindex $\lambda = t \bmod K + 1$.

2. Bestimme bestpassenden Restindex $t^* = \operatorname{argmax}_{t \in \mathcal{I}} u_\lambda(o_t)$.
3. Verschiebe Index t^* von \mathcal{I} nach \mathcal{C}_λ .

3 TERMINIERUNG

Jede Gruppe enthält entweder $|T/K|$ oder $\lceil T/K \rceil$ Datenelemente.

Beispiel — Text Mining

Gruppenprototypen $o_{t_1}, o_{t_2}, \dots, o_{t_{20}}$

originalsignal	mengenschreibweise	bzw
inkompatibilität	quantisierungsschritte	wertkontinuierlich
intervallskalierten	unterschiedlichen	übereinstimmungen
matrixdarstellung	mindestabtastrate	abtastzeitpunkten
datencharakteristika	ordnungsrelation	objektdatensatz
quantisierungsfehler	polygonzug	kovarianzmatrix
speicherplatzes	kaufmännisches	

Gruppen Ω_1, Ω_2 und Ω_6

(die Wörter mit den höchsten Zugehörigkeitsbewertungen)

Gruppe 1 Gruppe 2 Gruppe 6

originalsignal	mengenschreibweise	wertkontinuierlich
zeitsignal	matrixschreibweise	zeitkontinuierlich
ordinal	beschreiben	kontinuierliche
signal	schrittweise	wertebereich
signals	schreiben	rekonstruieren
digitalen	beschreibt	nichtnumerisch
zeitsignalen	geschrieben	willkürlich
proportional	beschrieben	konstruierten

Hierarchische Gruppierung: agglomerativ/divisiv

Austauschverfahren: (fuzzy) K-means

Mischungsidentifikation

Relationale Gruppierung

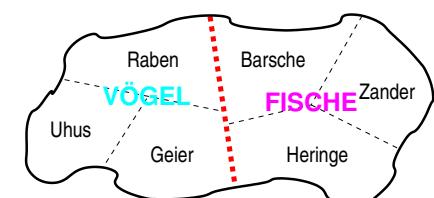
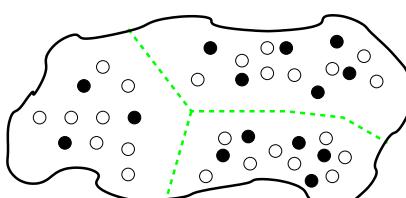
Konzeptuelle Gruppierung

Spektrale Gruppierung

Clustergütemaße

Zusammenfassung

Konzeptuelle Gruppierung



Gegeben

Objektbereich Ω · Hypothesenraum \mathcal{H} · Beispieldmenge $\omega \subseteq \Omega$

Gesucht

eine **intensionale Partition** von ω , d.h.:

Eine Folge von Hypothesen h_1, \dots, h_K , welche die beobachteten Beispiele aus ω sowie auch neue Objekte aus $\Omega \setminus \omega$ überschneidungsfrei gegeneinander abgrenzen.

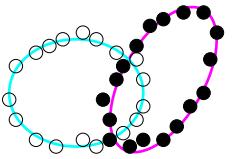
Und welches **Gütekriterium** optimiert die Partition ? ... ?

Lokale versus globale Objektähnlichkeit

Traditionelle Clusteranalyse

Lokaler Ähnlichkeitsbegriff (Hamming-Distanz)

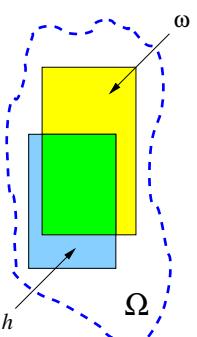
$$d(\mathbf{x}, \mathbf{y}) = \#\{\ell \mid x_\ell \neq y_\ell\}$$



Konzeptuelle Gruppierung

Globale Ähnlichkeit (engste \mathcal{H} -Umfassung)

$$d(\mathbf{x}, \mathbf{y}) = \min\{\sigma(h, \omega) \mid h \models \mathbf{x}, \mathbf{y}\}$$



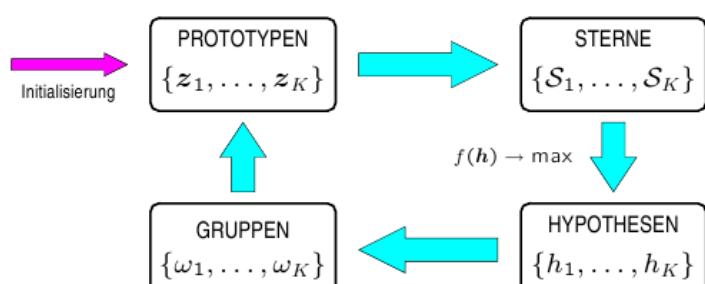
Überdeckungsgrad

der Hypothese h durch die Daten ω :

$$\sigma(h, \omega) = \frac{\#\{\mathbf{x} \in \omega \mid h \models \mathbf{x}\}}{\#\{\mathbf{x} \in \Omega \mid h \models \mathbf{x}\}} = \frac{\textcolor{green}{■}}{\textcolor{blue}{■}}$$

Wiederholtes Austauschen von Gruppenelementen

„conceptual K-means“



Gütekriterium
Kumulativer Überdeckungsgrad

$$f(\mathbf{h}) = \frac{1}{K} \cdot \sum_{k=1}^K \sigma(h_k, \omega)$$

Gruppenprototypen
müssen unbedingt ω angehören!
(Sterne, Überdeckung)

- Medoide
- Pseudomediane

Wahl der Gruppenzentren

Medoid

Dasjenige Gruppenelement mit minimaler **Exzentrizität**:

$$\mu_{\text{med}}(\omega_\kappa) = \underset{\mathbf{x} \in \omega_\kappa}{\operatorname{argmin}} \varepsilon(\mathbf{x}, \omega_\kappa) = \underset{\mathbf{x} \in \omega_\kappa}{\operatorname{argmin}} \sum_{\mathbf{y} \in \omega_\kappa} d(\mathbf{x}, \mathbf{y})$$

Pseudomedian

Kombiniere je nach Skalentyp der \mathcal{X}_n , $n = 1, \dots, N$ komponentenweise Mittelwerte, Zentroide, Mediane:

$$\mu_{\text{pseudo}}(\omega_\kappa) \stackrel{\text{def}}{=} (\mu_{\kappa,1}, \dots, \mu_{\kappa,N})^\top, \quad \mu_{\kappa,n} \stackrel{\text{def}}{=} \mu(\{x_n \mid \mathbf{x} \in \omega_\kappa\})$$

Ergibt eine hocheffizient berechenbare Näherung $\mu_{\text{pseudo}} \approx \mu_{\text{med}}$.

Reintegration

Wegen $\mu_{\text{pseudo}}(\omega_\kappa) \notin \omega_\kappa$ verwende den nächsten ω_κ -Nachbarn:

$$\tilde{\mu}_{\text{pseudo}}(\omega_\kappa) = \underset{\mathbf{x} \in \omega_\kappa}{\operatorname{argmin}} d(\mu_{\text{pseudo}}(\omega_\kappa), \mathbf{x})$$

Beispiel — Gebrauchtwagenhandel

Datensammlung

Objekt:	\mathbf{o}_1	\mathbf{o}_2	\mathbf{o}_3	\mathbf{o}_4	\mathbf{o}_5	\mathbf{o}_6	\mathbf{o}_7	\mathbf{o}_8
x_1 Geschwindigkeit	<i>h</i>	<i>m</i>	<i>h</i>	<i>l</i>	<i>m</i>	<i>l</i>	<i>h</i>	<i>m</i>
x_2 Farbe	<i>r</i>	<i>r</i>	<i>g</i>	<i>b</i>	<i>b</i>	<i>g</i>	<i>b</i>	<i>r</i>
x_3 Preis	<i>h</i>	<i>l</i>	<i>h</i>	<i>rh</i>	<i>rl</i>	<i>l</i>	<i>rh</i>	<i>rh</i>

$x_1 \in \mathcal{X}_1 = \{\text{high, medium, low}\}$

$x_2 \in \mathcal{X}_2 = \{\text{red, blue, green}\}$

$x_3 \in \mathcal{X}_3 = \{\text{high, rel_high, rel_low, low}\}$

Hypothesen als Attributkomplexe

zum Beispiel: $x_1 \in \{h\} \wedge x_2 \in \{b, g\} \wedge x_3 \in \{h, rh, rl\}$

<i>r</i>	1		
<i>b</i>	7		
<i>g</i>	3		

h rh rl l

Ebene $x_1 = h$

<i>r</i>	8		2
<i>b</i>		5	
<i>g</i>			

h rh rl l

Ebene $x_1 = m$

<i>r</i>			
<i>b</i>	4		
<i>g</i>		6	

h rh rl l

Ebene $x_1 = l$

Überdeckung

$$\sigma(h, \omega) = \frac{2}{6} = 0.\bar{3}$$

Konzeptueller Austauschalgorithmus

(Algorithmus)

1 INITIALISIERUNG

Wähle Klassenzahl K und wähle z_1, \dots, z_K zufällig aus ω .

2 PROTOTYPEN \Rightarrow STERNE

$$\mathcal{S}_\kappa = \mathcal{S}(z_\kappa \mid \{z_1, \dots, z_K\} \setminus \{z_\kappa\})$$

3 STERNE \Rightarrow HYPOTHESEN

Bestimme ω -einheitlichen Hypothesensatz $\mathbf{h} \in \mathcal{S}_1 \times \dots \times \mathcal{S}_K$ mit

$$f(\mathbf{h}) = \text{MAX}$$

4 HYPOTHESEN \Rightarrow GRUPPEN

$$\omega_\kappa = \{x \in \omega \mid h_\kappa \models x\}$$

5 GRUPPEN \Rightarrow PROTOTYPEN

$$z_\kappa = \mu_{\text{med}}(\omega_\kappa)$$

6 TERMINIERUNG Wenn $f(\mathbf{h}) \geq \theta$ dann \rightsquigarrow Ende sonst \rightsquigarrow 2.

Beispiel — Gebrauchtwagenhandel

Erster Iterationsschritt

P1 Wähle als initiale Prototypen $z_1 = o_1$ und $z_2 = o_2$ aus

S1 Stern von z_1 :

$$h_1^1 = (x_2 = r, b) \wedge (x_3 = h, rh, rl) \quad \text{und} \quad h_1^2 = (x_1 = h, l)$$

Stern von z_2 :

$$h_2^1 = (x_1 = h) \wedge (x_2 = g) \vee (x_3 = l) \quad \text{und} \quad h_2^2 = (x_1 = m)$$

G1 Dann gilt

$$\begin{aligned} h_1^1 &\models o_1, o_4, o_5, o_7, o_8 \\ h_1^2 &\models o_1, o_3, o_4, o_6, o_7 \end{aligned}$$

$$\begin{aligned} h_2^1 &\models o_2, o_3, o_6 \\ h_2^2 &\models o_2, o_5, o_8 \end{aligned}$$

H1 Nur die Kombinationen (h_1^1, h_2^1) und (h_1^2, h_2^2) bilden konzeptuelle Partitionen

$$f(h_1^1) + f(h_2^1) = 5/18 + 3/12 = 38/72$$

$$f(h_1^2) + f(h_2^2) = 5/24 + 3/12 = 33/72$$

Beispiel — Gebrauchtwagenhandel

Zweiter Iterationsschritt

P2 Objekte und ihr Median in h_1^1

Attribut	o_1	o_4	o_5	o_7	o_8	Modus
x_1	h	l	m	h	m	h, m
x_2	r	b	b	b	r	b
x_3	h	rh	rl	rh	rh	rh

→ Zentroid ist o_7

Die Hypothese h_2^1 hat Pseudomedian (m, g, l) und Median o_6

→ neue Gruppenprototypen sind $z_1 = o_7, z_2 = o_6$

S2 Stern von o_7 :

$$h_1^1 = (x_3 = h, rh)$$

$$h_1^2 = (x_1 = h, m) \wedge (x_2 = r, b) \wedge (x_3 = r, rh)$$

Stern von o_6 :

$$h_2^1 = (x_1 = m, l) \wedge (x_3 = rl, l)$$

G2 Dann gilt

$$\begin{aligned} h_1^1 &\models o_1, o_3, o_4, o_7, o_8 \\ h_1^2 &\models o_1, o_7, o_8 \end{aligned}$$

$$h_2^1 \models o_2, o_5, o_6$$

H2 Nur die Kombination (h_1^1, h_2^1) bildet eine konzeptuelle Partition

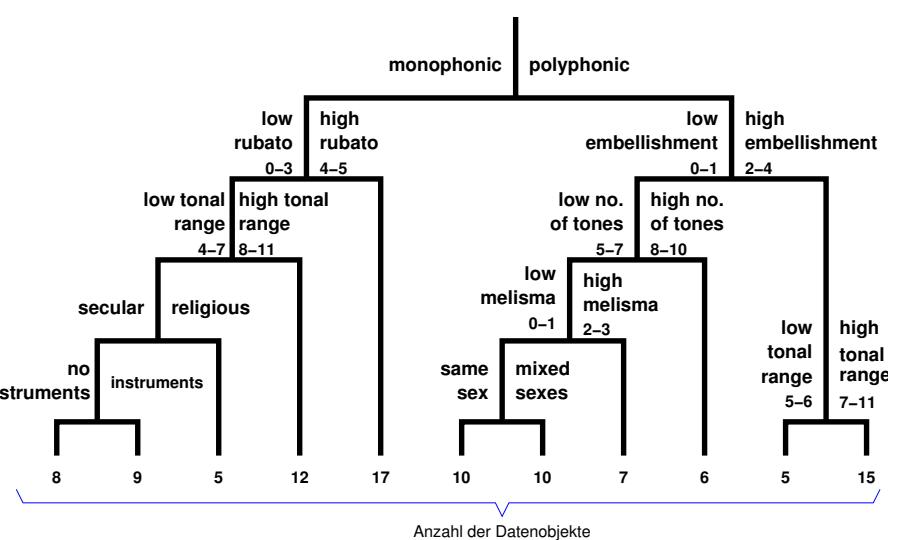
$$f(h_1^1) + f(h_2^1) = 5/18 + 3/12 = 38/72$$

Beispiel — Taxonomie spanischer Volkslieder

Gattungsdendrogramm nach konzeptueller Division

100 Lernbeispiele spanischer Volkslieder

22 Attribute mit nominalen / ordinalen Wertebereichen



Page Rank

„Gute Webseiten werden von guten Webseiten erwähnt.“

Relevanz und Qualität

Seitenbewertung = Anfragepassung + Seriositätsmaß

$$\text{score}_q^{\text{Google}}(\text{doc}) = \text{Rel}_q(\text{doc}) + \text{rank}(\text{doc})$$

Worldwide Web als gerichteter Graph

Adjazenzmatrix $A \in \{1, 0\}^{T \times T}$ mit $a_{st} = 1 \Leftrightarrow \text{doc}_i \mapsto \text{doc}_j$

Irrfahrtmodell

Der „Random Surfer“ besucht Webseiten mit W'keit p_j und der Politik

$$p_j = (1 - \beta) \cdot \frac{1}{T} + \beta \cdot \sum_i p_i \cdot a_{ij} \cdot \frac{1}{\sum_k a_{ik}}$$

Die **Gleichgewichtsverteilung** gehorcht einer Eigenwertaufgabe ($\lambda = 1$):

$$B \cdot p = \left((1 - \beta) \cdot \frac{1}{T} + \beta \cdot \tilde{A} \right) \cdot p = p = \lambda \cdot p, \quad \tilde{a}_{ij} \stackrel{\text{def}}{=} \frac{a_{ij}}{\sum_k a_{ik}}$$

Hierarchische Gruppierung: agglomerativ/divisiv

Austauschverfahren: (fuzzy) K-means

Mischungsidentifikation

Relationale Gruppierung

Konzeptuelle Gruppierung

Spektrale Gruppierung

Clustergütemaße

Zusammenfassung

Schnitte in gewichteten Graphen

Definition

Sei $(\mathcal{K}, \mathcal{E}, \mathbf{A})$ ein ungerichteter, gewichteter Graph mit nicht-negativer, symmetrischer **Affinitätsmatrix \mathbf{A}** . Für zwei Knotenmengen \mathcal{B}, \mathcal{C} sei

$$\ell_{\text{aff}}(\mathcal{B}, \mathcal{C}) = \sum_{s \in \mathcal{B}} \sum_{t \in \mathcal{C}} A_{st}, \quad \tilde{\ell}_{\text{aff}}(\mathcal{B}, \mathcal{C}) = \frac{\ell_{\text{aff}}(\mathcal{B}, \mathcal{C})}{\ell_{\text{aff}}(\mathcal{B}, \mathcal{K})}$$

definiert. Eine Menge $\mathcal{C} \subset \mathcal{K}$ mit minimalem $\ell_{\text{aff}}(\mathcal{C}, \mathcal{K} \setminus \mathcal{C})$ bzw. mit minimalem $\tilde{\ell}_{\text{aff}}(\mathcal{C}, \mathcal{K} \setminus \mathcal{C})$ heißt **Schnitt** oder **normierter Schnitt**. Eine Partition $\mathcal{C}_1, \dots, \mathcal{C}_K$ von \mathcal{K} mit minimalem

$$\tilde{\ell}_{\text{aff}}(\{\mathcal{C}_\kappa\}_{\kappa=1}^K) \stackrel{\text{def}}{=} \frac{1}{K} \cdot \sum_{\kappa=1}^K \tilde{\ell}_{\text{aff}}(\mathcal{C}_\kappa, \mathcal{K} \setminus \mathcal{C}_\kappa)$$

heißt **normierter K-Schnitt**.

Bemerkung

Für die Affinitätsmatrix \mathbf{A} gilt $A_{ss} = 0$ und $A_{st} = A_{ts}$ für alle $s, t \in \{1, \dots, T\}$.

K-NC als Spurmaximierung

„K-way normalized cut“

Matrixalgebraische Formulierung

Die **Indikatormatrix $\mathbf{C} \in \{0, 1\}^{T \times K}$** beschreibt die Zugehörigkeit der Knoten v_t zu den Gruppen \mathcal{C}_κ :

$$c_{t\kappa} \stackrel{\text{def}}{=} \begin{cases} 1 & t \in \mathcal{C}_\kappa \\ 0 & t \notin \mathcal{C}_\kappa \end{cases}$$

Die **Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{T \times T}$** enthält je Knoten die Summe seiner ausgehenden (einlaufenden) Kantengewichte:

$$\mathbf{D} = \text{diag}(\{d_s\}), \quad d_s \stackrel{\text{def}}{=} \sum_{t=1}^T A_{st}$$

Lemma

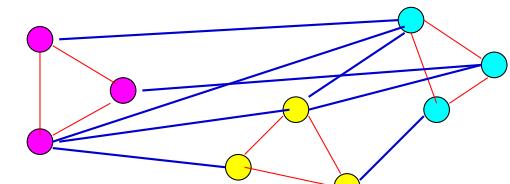
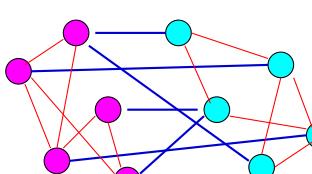
Das K-NC Kriterium ist äquivalent zur Maximierung der Größe

$$\gamma_K \cdot \text{spur}(\mathbf{Z}^\top \mathbf{A} \mathbf{Z}) \quad \text{mit} \quad \mathbf{Z} \stackrel{\text{def}}{=} \mathbf{C} \cdot (\mathbf{C}^\top \mathbf{D} \mathbf{C})^{-1/2}.$$

Gewöhnliche & normierte Schnitte

Fakt

Die Berechnung eines (normierten) 2-Schnittes ist beweisbar NP-hart.



Gewöhnlicher Schnitt

Dichotome Partition von \mathcal{K} mit **minimaler Mengenaffinität**:
Summe der Querverbindungsge wichtete zwischen \mathcal{C} und $\mathcal{K} \setminus \mathcal{C}$

Normierter Schnitt

Minimale relative Mengenaffinität:

Proportion der Querverbindungsge wichtete zu den Gewichten aller \mathcal{C} verlassenden Kanten

Normierter K-Schnitt

Minimale Summe aller relativen Affinitäten zwischen den Schnittmengen \mathcal{C}_κ und ihren **Komplementen** $\mathcal{K} \setminus \mathcal{C}_\kappa$

Relaxationslösung

F. Chung: Spectral Graph Theory, AMS 1997

Skalierung

Die umskalierte Matrix $\tilde{\mathbf{Z}} := \mathbf{D}^{1/2} \mathbf{Z} \in \mathbb{R}^{T \times K}$ besitzt offenbar orthonormale Spalten:

$$\tilde{\mathbf{Z}}^\top \tilde{\mathbf{Z}} = \mathbf{E} \in \mathbb{R}^{K \times K}$$

⇒ Spurmaximierung durch Berechnung der K ersten Eigenvektoren

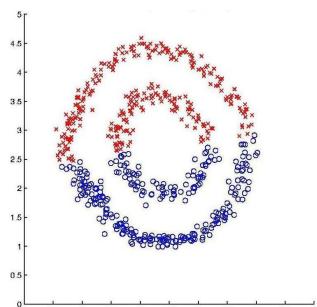
Lemma

Die Matrix $\tilde{\mathbf{Z}} \in \mathbb{R}^{T \times K}$ mit den K oberen Eigenvektoren von $\tilde{\mathbf{A}} := \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$ als Spalten maximiert die Spur

$$\text{spur} \left(\underbrace{\tilde{\mathbf{Z}}^\top \cdot \mathbf{D}^{-1/2} \cdot \mathbf{A} \cdot \mathbf{D}^{-1/2} \cdot \tilde{\mathbf{Z}}}_{\mathbf{Z}^\top} \right)$$

unter der Bedingung $\tilde{\mathbf{Z}}^\top \tilde{\mathbf{Z}} = \mathbf{E}$.

Spektrale Gruppierung statt K-means



K-means Algorithmus

modelliert ausschließlich **konvexe** Ballungsgebiete und findet nur **lokale** Verzerrungsminima.

2 Ringwolken — 2-means-Gruppierung

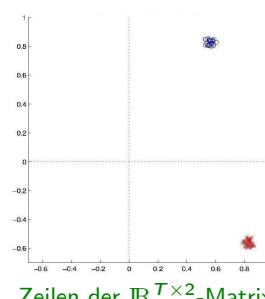
1. Bestimme $\tilde{\mathbf{A}}$
2. Berechne $\tilde{\mathbf{Z}}$ (EWP)
3. Ermittle $\mathbf{Z} = \mathbf{D}^{-\frac{1}{2}} \tilde{\mathbf{Z}}$
4. Errate (?)! \mathbf{C} aus \mathbf{Z}

Spektrale Gruppierung

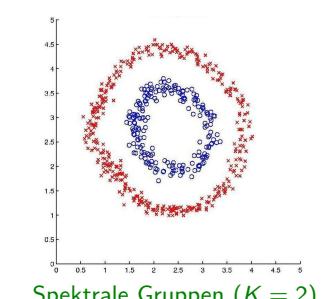
Die metrische Struktur der Datenobjekte wird in einen gewichteten Graphen transformiert; anschließend wird der Graph in Polynomialzeit durch Berechnung eines Semi-Schnittes in K Teilgraphen partitioniert.

Gruppieren im Spektralraum des Ähnlichkeitsgraphen

$$\mathbb{R}^{T \times N} \xrightarrow{\text{Affinität}} \mathbb{R}^{T \times T} \xrightarrow{\text{Eigenraum}} \mathbb{R}^{T \times K} \xrightarrow{\text{Gruppen}} \mathbb{R}^{K \times K}$$



Gruppen-ID



Eigenraummatrix $\mathbf{U} \in \mathbb{R}^{T \times K}$

Die K -dimensionalen Zeilenvektoren weisen eine hochdiskriminante Gruppenstruktur (innerer Ring — äußerer Ring) auf.

Ng-Jordan-Weiss Algorithmus

(Algorithmus)

1 AFFINITÄSMATRIX

$$\mathbf{A} \in \mathbb{R}^{T \times T} \text{ mit } A_{st} \stackrel{\text{def}}{=} \begin{cases} 0 & s = t \\ \exp\{-\|x_s - x_t\|^2 / 2\sigma^2\} & s \neq t \end{cases}$$

2 LAPLACEMATRIX

$$\mathbf{L} \in \mathbb{R}^{T \times T} \text{ via Normierung } L_{st} \stackrel{\text{def}}{=} \frac{A_{st}}{\sqrt{\rho_s \cdot \rho_t}}, \quad \rho_s \stackrel{\text{def}}{=} \sum_{r=1}^T A_{rs}$$

3 K FÜHRENDE EIGENVEKTOREN

$$\mathbf{L} = \sum_{t=1}^T d_t^2 \cdot \mathbf{u}_t \mathbf{u}_t^\top, \quad \mathbf{U} \stackrel{\text{def}}{=} (\mathbf{u}_1, \dots, \mathbf{u}_K) \in \mathbb{R}^{T \times K}$$

4 ZEILENWEISE NORMIERUNG

$$\tilde{\mathbf{U}} \stackrel{\text{def}}{=} \mathbf{C}^{-1/2} \cdot \mathbf{U}, \quad C_{st} = \begin{cases} \sum_{\kappa=1}^K U_{t\kappa}^2 & s = t \\ 0 & s \neq t \end{cases}$$

5 K-MEANS GRUPPIERUNG

der Matrixzeilen $\tilde{\pi} : \{\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_T\} \rightarrow \{1, 2, \dots, K\}$

6 PARTITIONIERUNG

der Originaldaten $\pi : \begin{cases} \{x_1, \dots, x_T\} & \rightarrow \{1, 2, \dots, k\} \\ x_t & \mapsto \tilde{\pi}(\tilde{\mathbf{v}}_t) \end{cases}$

(zurück in \mathbf{A})

Warum funktioniert der NJW-Algorithmus ?

Beobachtung (geodätische Gruppenbildung)

Die Zeilen der Matrix \mathbf{U} bilden eine K -dimensionale Repräsentation der Daten, in der Objekte mit kurzem Verbindungsweg — geodätisch, nicht Luftlinie — nahe beieinander liegen.
warum?

Idealytisches Szenario ($K = 3$)

Objekte unterschiedlicher Gruppe besitzen den euklidischen Abstand ∞ .

⇒ Affinitätsmatrix und Laplacematrix sind von **Blockdiagonalform** (geeignete Nummerierung der x_t vorausgesetzt)

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{L}_3 \end{pmatrix}$$

- Die Eigenvektoren von \mathbf{L} sind die *mit Nullen aufgefüllten* Eigenvektoren der Blockmatrizen \mathbf{L}_κ .
- Dank der Doppelnormierung von \mathbf{L} besitzt jeder Block *genau einen* maximalen **Eigenwert Eins**.

Warum funktioniert der NJW-Algorithmus ?

Idealtypisches Szenarium ($K = 3$)

Die K Haupteigenvektoren von \mathbf{L} rekrutieren sich aus den K **Blockgewinnern**.

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}'_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{u}'_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{u}'_3 \end{pmatrix} \in \mathbb{R}^{T \times 3}, \quad \tilde{\mathbf{U}} = \begin{pmatrix} \mathbf{100} \\ \mathbf{010} \\ \mathbf{001} \end{pmatrix} \in \mathbb{R}^{T \times 3}$$

- Die Matrix \mathbf{U} besitzt die K Eigenvektoren als Spalten.
- Alle Zeilen von \mathbf{U} enthalten **genau einen** Eintrag ungleich Null.
- In der *zeilensummennormierten* Matrix $\tilde{\mathbf{U}}$ wird der Eintrag zur Eins.
- ⇒ Das Gruppieren der Zeilen (\approx Einheitsvektoren) ergibt zwangsläufig genau die richtigen Cluster!

Denkfehler

Die K Haupteigenvektoren besitzen den gemeinsamen Eigenwert Eins.

- ⇒ Sie sind also keineswegs eindeutig bestimmt und voller Nullen, sondern spannen lediglich einen eindeutig bestimmten K -dimensionalen Unterraum auf.

Rettung der Argumentation

Statt \mathbf{U} erhalten wir $\mathbf{U}' = \mathbf{U}\mathbf{R}$ mit einer Rotationsmatrix $\mathbf{R} \in \mathbb{R}^{K \times K}$.

Bezeichne \mathbf{v}_t^\top die t -te Zeile von \mathbf{U} .

- ⇒ $\mathbf{v}_t^\top \mathbf{R}$ ist die t -te Zeile von \mathbf{U}' und für die Quadratnorm gilt:

$$(\mathbf{v}_t^\top \mathbf{R}) \cdot (\mathbf{R}^\top \mathbf{v}_t) = \mathbf{v}_t^\top \cdot (\mathbf{R}\mathbf{R}^\top) \cdot \mathbf{v}_t = \mathbf{v}_t^\top \cdot \mathbf{v}_t = \|\mathbf{v}_t\|^2$$

Die Zeilennormierung macht $\mathbf{U}\mathbf{R}$ zu $\tilde{\mathbf{U}}\mathbf{R}$.

Zu clustern sind nicht mehr die **Einheitsvektoren** des \mathbb{R}^K , aber immerhin noch die Vektoren einer Orthonormalbasis \mathbf{R} des Raumes.

Details zum NJW-Algorithmus

Lineare Störungstheorie

Analyse des NJW **ohne** Intercluster-Distanzen = ∞

Stewart & Sun: *Matrix Perturbation Theory*, 1990

- ⇒ **Eigengap** $\lambda_K - \lambda_{K+1}$ als untere Schranke der Gruppierungsstabilität

Abklingparameter $\sigma^2 > 0$

Minimale Endverzerrung nach dem K -means Clustering

- ⇒ Skalarer Optimierungslauf für σ^2

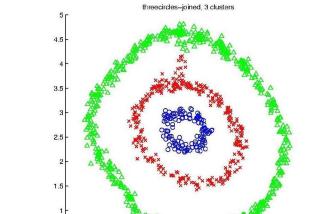
Startpartition für K -means

Die (idealen) Gruppenzentren liegen auf der Einheitssphäre.

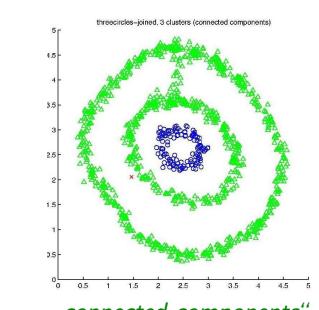
Sie stehen paarweise senkrecht aufeinander.

- ⇒ Sukzessive Auswahl derjenigen \mathbf{v}_t als Saatpunkte, die zu allen bereits selektierten Kandidaten maximal orthogonal sind.

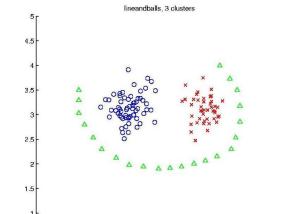
Beispiele — NJW & Wettbewerber



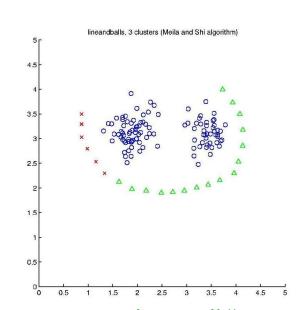
Ng-Jordan-Weiss, $K = 3$



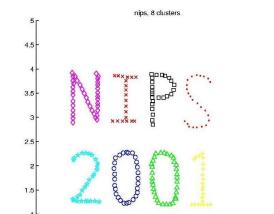
„connected components“



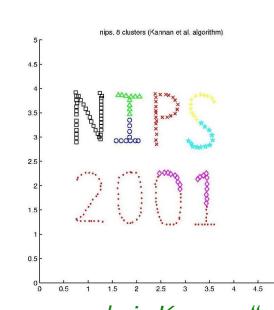
Ng-Jordan-Weiss, $K = 3$



„random walk“



Ng-Jordan-Weiss, $K = 8$



„geodesic K-means“

Sonstige spektral orientierte Verfahren

Gruppierung nach Zusammenhangskomponenten

Bilde den ungerichteten Graphen mit den Kanten

$$(s, t) \in \mathcal{E} \quad \Leftrightarrow \quad \|x_s - x_t\| \leq \theta .$$

Wähle die Schranke θ so, daß $\#(\text{ZSH-Komponenten}) = K$ ist.

Gruppierung nach dem Irrfahrtprinzip (Meila & Shi)

Unterm Strich derselbe Ablauf wie beim NJW-Algorithmus, aber:

- Nur die *Zeilen* der Affinitätsmatrix A werden normiert.
- Die Zeilen der Eigenvektormatrix U werden *nicht* normiert.

Geodätischer K-means (Kannan & Vempala & Vetta)

Wiederum derselbe Ablauf wie beim NJW-Algorithmus, aber:

- Nur die *Zeilen* der Affinitätsmatrix A werden normiert.
- Die *Urbilder* der Clusterzentroide werden als *Repräsentanten* genutzt.

K-means mit Termexpansion

Der Kernel Trick

Gruppieren in einem (impliziten) Expansionsraum $\phi(\Omega)$:

$$\phi : \Omega \rightarrow \mathbb{H}, \quad \langle \phi x, \phi y \rangle_{\mathbb{H}} = K(x, y)$$

Der **Kernoperator** $K(\cdot, \cdot)$ simuliert das „Rechnen“ im RKHS \mathbb{H} .

Optimale K-Gruppierung in $\phi(\Omega)$

Ein Kodebuch $\{\mu_1, \dots, \mu_K\}$ mit minimaler **Verzerrung**

$$\varepsilon(\{\omega_k\}_k) = \sum_{k=1}^K \sum_{x \in \omega_k} \|\phi(x) - \mu_k\|^2$$

Berechnung von
Gruppenzentren μ_k

$\mu_k \in \mathbb{H}$ mittelt (endlich
viele) expandierte Objekte.

Berechnung von
Prototypdistanzen $\|f - g\|_{\mathbb{H}}^2$

f ist ein expandiertes Objekt.
 g ist ein Gruppenzentrum.

Dualisierte Berechnungen für K-means

Lemma

Sei $\omega = \{x_1, \dots, x_T\} \subset \Omega$ und $\phi : \Omega \rightarrow \mathbb{H}$ eine Expansion mit dem zugehörigen Kernoperator $K(\cdot, \cdot)$.

Beweis.

1. Das Zentroidelement der termexpandierten Daten $\phi(\omega)$ bezüglich des quadratischen euklidischen Abstandes $\|\cdot\|_{\mathbb{H}}^2$ ist der Mittelwertvektor
$$\mu = \frac{1}{T} \cdot \sum_{t=1}^T \phi(x_t) .$$
 2. Der Abstand zwischen μ und einem expandierten Objekt $\phi(y)$, $y \in \Omega$, lässt sich mit $O(T^2)$ Kernoperatorauswertungen berechnen.
 3. Die Berechnung der Abstände von μ zu allen $\phi(x_t)$, $t = 1, \dots, T$, erfordert i.a. den Aufwand $O(T^2N)$, wenn $\Omega = \mathbb{R}^N$ ist.
- $$\begin{aligned} \|\phi y - \mu\|_{\mathbb{H}}^2 &= \left\| \phi y - \frac{1}{T} \sum_{x_t} \phi x_t \right\|_{\mathbb{H}}^2 \\ &= \langle \phi y, \phi y \rangle - 2 \cdot \frac{1}{T} \cdot \sum_{x_t} \langle \phi y, \phi x_t \rangle + \frac{1}{T^2} \cdot \sum_{x_s, x_t} \langle \phi x_s, \phi x_t \rangle \\ &= K(y, y) - 2 \cdot \frac{1}{T} \cdot \sum_{x_t} K(y, x_t) + \frac{1}{T^2} \cdot \sum_{x_s, x_t} K(x_s, x_t) \\ &= \frac{1}{T^2} \cdot \left\{ T^2 \cdot G_{rr} - 2T \cdot \sum_{x_t} G_{rt} + \sum_{x_s, x_t} G_{st} \right\} \\ &= G_{rr} - 2\bar{g}_r + \bar{G} \end{aligned}$$
3. Jede Kernoperatorauswertung kostet $O(N)$, die Berechnung der Gramschen Matrix kostet $O(T^2N)$, und die Mittelungen über G und ihre Zeilen g_r bleiben bei $O(T^2)$.

Kernel K-means Algorithmus

Kostenpunkt: $O(T^2N + T^2I)$

1 STARTWERTE

Startgruppierung $\{\omega_\kappa^{(0)}\}$ und Kernmatrix \mathbf{G} , $G_{st} = K(\mathbf{x}_s, \mathbf{x}_t)$.

2 NEUE OBJEKTZUGEHÖRIGKEIT

$$\kappa^*(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname{argmin}_{\lambda=1..K} \left\| \phi\mathbf{x} - \mu_\lambda^{(i)} \right\|$$

3 IMPLIZITE ZENTROIDBERECHNUNG

$$\omega_\lambda^{(i+1)} \leftarrow \{\mathbf{x}_t \mid \kappa^*(\mathbf{x}_t) = \lambda\}$$

(keine explizite Berechnung der Mittelwertvektoren $\mu_\lambda^{(i+1)} \in \mathbb{H}$)

4 TERMINIERUNG

Wenn $\varepsilon^{(i)}(\{\omega_\kappa\}) \leq \theta$ dann Ende sonst $i \leftarrow i + 1$ und \rightsquigarrow 2.

Gewichteter Kernel K-means

Gewichtetes Verzerrungskriterium

Objektabhängige Gewichtung der Zentrumssabstände:

$$\varepsilon(\{\omega_\kappa\}, \mathbf{w}) = \sum_{\kappa=1}^K \sum_{\mathbf{x} \in \omega_\kappa} w(\mathbf{x}) \cdot \|\phi(\mathbf{x}) - \mu_\kappa\|^2$$

\Rightarrow Gruppenzentroide $\hat{=}$ gewichtete Mittelwertvektoren.

Optimale Gruppenstruktur

Minimale Verzerrung \Leftrightarrow maximale Spur (NC):

- Kernmatrix $\mathbf{W}^{1/2} \mathbf{G} \mathbf{W}^{1/2}$ korrespondiert mit normierter Affinität $\tilde{\mathbf{A}}$.
- Mantelmatrix \mathbf{U} korrespondiert mit $\tilde{\mathbf{Z}}$.

Jede Zeile von \mathbf{U} ist ein skaliertes K -Einheitsvektor.

$\Rightarrow \mathbf{U}^\top \mathbf{U} = \mathbf{E}_{(K)}$ (\mathbf{U} besitzt orthonormale Spalten)

WKKM kann (im Relaxationssinne) auch durch NJW gelöst werden!

WKKM vs. normierter Schnitt

Minimale Verzerrung \Leftrightarrow maximale NC-Matrixspur

Lemma

Die Gruppenverzerrung des WKKM ist gleichwertig zum Ausdruck

$$\varepsilon(\{\omega_\kappa\}, \mathbf{w}) = \operatorname{spur} (\mathbf{W}^{1/2} \mathbf{G} \mathbf{W}^{1/2}) - \operatorname{spur} (\mathbf{U}^\top \cdot \mathbf{W}^{1/2} \mathbf{G} \mathbf{W}^{1/2} \cdot \mathbf{U})$$

mit

$$\Phi = (\phi\mathbf{x}_1, \dots, \phi\mathbf{x}_T)$$

$$\mathbf{G} = \Phi^\top \Phi$$

$$\mathbf{W} = \operatorname{diag}(\{w(\mathbf{x}_t)\}_t)$$

$$\mathbf{U} = \operatorname{diag}(\{s_\lambda^{-1/2} \cdot \mathbf{W}_\lambda^{1/2} \cdot \mathbf{1}_\lambda\}), \quad s_\lambda = \mathbf{1}_\lambda^\top \mathbf{W}_\lambda \mathbf{1}_\lambda.$$

Die ersten K Eigenvektoren der Matrix $\mathbf{W}^{1/2} \mathbf{G} \mathbf{W}^{1/2}$ liefern eine verzerrungsminimale Lösung unter der **Relaxationsbedingung** $\mathbf{U}^\top \mathbf{U} = \mathbf{E}$ für die blockstrukturierte $(T \times K)$ -Matrix \mathbf{U} .

(Beweis durch exzessives Nachrechnen)

Spektrale Gruppierung

Vorzügliche und nachteilige Eigenschaften

NJW-Algorithmus

- einstufiges Verfahren
- metrische Distanzen
- K -means über \mathbb{R}^K
- Eigenvektoren $O(T^2K)$
- Gruppen raten $O(TK^2I)$
- Relaxationslösung !?!

WKKM-Algorithmus

- iteratives Verfahren
- Mercer-reskalierbar
- K -means über \mathbb{H} /dual
- Gram-Matrix $O(T^2N)$
- dualer K -means $O(T^2I)$
- Startpartition ??

Hybride spektrale Gruppierung

- 1 Berechne ggf. die Affinitäten \mathbf{A} , Zeilensummen \mathbf{D} und $\tilde{\mathbf{A}} = \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$.
- 2 Startgruppierung $\{\omega_\kappa^{(0)}\}$ via NJW-Algorithmus auf $\tilde{\mathbf{A}}$.
- 3 Berechne die (virtuelle) Gram-Matrix $\mathbf{G} = \mathbf{D}^{-1} \mathbf{A} \mathbf{D}^{-1}$ und setze $\mathbf{W} = \mathbf{D}$.
- 4 Führe den i -ten Weighted-Kernel-K-Means-Doppelschritt durch.
- 5 Wenn $\varepsilon^{(i)}(\{\omega_\kappa\}) \leq \theta$ dann Ende sonst $i \leftarrow i + 1$ und \rightsquigarrow 4.

Hierarchische Gruppierung: agglomerativ/divisiv

Austauschverfahren: (fuzzy) K-means

Mischungsidentifikation

Relationale Gruppierung

Konzeptuelle Gruppierung

Spektrale Gruppierung

Clustergütemeße

Zusammenfassung

Gütemeße für Gruppen & Partitionen

Fragen über Fragen

- Mit welcher Vorgabeordnung $K \in \mathbb{N}$ starte ich K-means ?
- Welche Zerlegungsebene wähle ich als Resultat aus? (agnes, diana, pam & Co.)
- Zerfällt ω_k in noch kleinere Gruppen ?
- Trifft gelernte Partition $\{\hat{\omega}_k\}_k$ die wahre Gruppenstruktur ?

Problem

Weder Gruppenverzerrungen $\varepsilon(\omega_k)$

noch Gesamtverzerrung $\varepsilon(\{\omega_k\}_k)$

beantworten auch nur eine dieser Fragen !

Cluster Recovery Index

Externe Qualitätskriterien — Vergleich mit „Goldstandard“

Ausgangssituation

Objektmenge $\{o_1, \dots, o_T\}$ mit wahrer und hypothetischer Gruppierung:

$$\omega_1^* \uplus \omega_2^* \uplus \dots \uplus \omega_K^* \quad \text{versus} \quad \hat{\omega}_1 \uplus \hat{\omega}_2 \uplus \dots \uplus \hat{\omega}_K$$

Aufgabenstellung

Berechnen eines Übereinstimmungsmaßes zwischen $\{\omega_k^*\}$ und $\{\hat{\omega}_k\}$.

- Vergleich einer Ist-Lösung mit der Soll-Lösung
- Vergleich zweier Lösungen zweier Methoden
- Bestimmung des Medoids mehrerer Partitionen

Problem

Gruppierungen sind nur eindeutig bis auf **Indexpermutation**.

Der Rand-Index

Überschneidungsfreie, scharfe Gruppen (W. M. Rand, 1971)

Kreuzadjazenzstatistiken

Die Objekte o_1, \dots, o_T bilden $M = \binom{T}{2}$ ungeordnete Paaren $\{o_s, o_t\}$.

	gleiche $\{\hat{\omega}_k\}$ -Gruppe	verschiedene $\{\hat{\omega}_k\}$ -Gruppen	
gleiche $\{\omega_k^*\}$ -Gruppe	M_{11}	M_{10}	$M_{1\cdot}$
verschiedene $\{\omega_k^*\}$ -Gruppen	M_{01}	M_{00}	$M_{0\cdot}$
	$M_{\cdot 1}$	$M_{\cdot 0}$	M

Definition

Unter dem **Rand-Index** zweier scharfer Objektpartitionen verstehen wir den relativen Anteil

$$C_{\text{rand}} \stackrel{\text{def}}{=} \frac{M_{11} + M_{00}}{M}$$

der kohärent gruppierten Punktpaare $\{o_s, o_t\}$.

Der bereinigte Rand-Index

„adjusted Rand index“ (Hubert & Arabie, 1985)

Problem

- \oplus Maximum $C_{\text{rand}} = 1$ wird für äquivalente Partitionen angenommen.
- \ominus Hohe C_{rand} -Werte entstehen auf Grund zufälliger Korrespondenzen.

Definition

Unter dem **bereinigten Rand-Index** zweier scharfer Objektpartitionen verstehen wir den Quotienten

$$C_{\text{ari}} \stackrel{\text{def}}{=} \frac{\text{observed} - \text{expected}}{\text{maximum} - \text{expected}} = \frac{C_{\text{rand}} - \{M_1.M_1 + M_0.M_0\} / M^2}{1 - \{M_1.M_1 + M_0.M_0\} / M^2}$$

aus **beobachtetem** und **größtmöglichen** Übertreffen der allein zufallsbedingten Gruppenkohärenz.

Bemerkung

Ein *störbereinigtes* und *permutationsinvariantes* Vergleichsmaß ist auch die **Transformation** $\mathcal{H}(\mathbb{K}_1) + \mathcal{H}(\mathbb{K}_2) - \mathcal{H}(\mathbb{K}_1, \mathbb{K}_2)$ zwischen dem wahren und dem hypothetischen Gruppenindex der Objekte.

Die Heterogenität einer Punktmenge

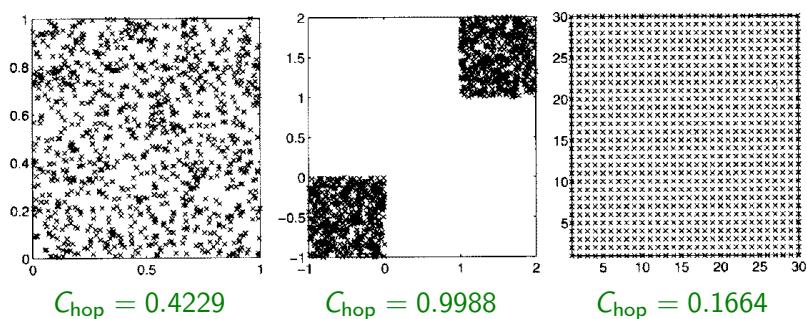
Zerfällt $\omega \subset \Omega$ in noch kleinere Gruppen?

Definition

Für die Punktmenge ω mit der konvexen Hülle $\bar{\omega} \supset \omega$ ist der **Hopkins-Index** durch die relative Punktdichte

$$C_{\text{hop}} \stackrel{\text{def}}{=} \frac{\mathcal{E}_{\bar{\omega}}[d^*(\mathbb{X}, \omega)]}{\mathcal{E}_{\bar{\omega}}[d^*(\mathbb{X}, \omega)] + \mathcal{E}_{\omega}[d^*(\mathbb{X}, \omega^{(\mathbb{X})})]}$$

im Gesamtbereich der ω -Hülle definiert.



Bestimmung des Hopkins-Index

(Algorithmus)

1 RESAMPLING ω und $\bar{\omega}$

Ziehe $S \in \mathbb{N}$ Vektoren z_1, \dots, z_S aus ω .

Ziehe $S \in \mathbb{N}$ Vektoren y_1, \dots, y_S aus der konvexen Hülle

$$\bar{\omega} \stackrel{\text{def}}{=} \left\{ \sum_{t=1}^T a_t x_t \mid \sum_{t=1}^T a_t = 1, a_t \geq 0 \right\}$$

2 PUNKTDICHTE IN $\bar{\omega}$

Berechne die kumulative Punkt-Mengen-Distanz

$$\mathcal{E}_{\bar{\omega}}[d^*(\mathbb{X}, \omega)] \approx D_{\bar{\omega}} = \sum_{s=1}^S \min_{x \in \omega} d(y_s, x)$$

3 PUNKTDICHTE IN ω

Berechne die „leave-one-out“ kumulative Punkt-Mengen-Distanz

$$\mathcal{E}_{\omega}[d^*(\mathbb{X}, \omega^{(\mathbb{X})})] \approx D_{\omega} = \sum_{s=1}^S \min_{x \in \omega \setminus \{z_s\}} d(z_s, x)$$

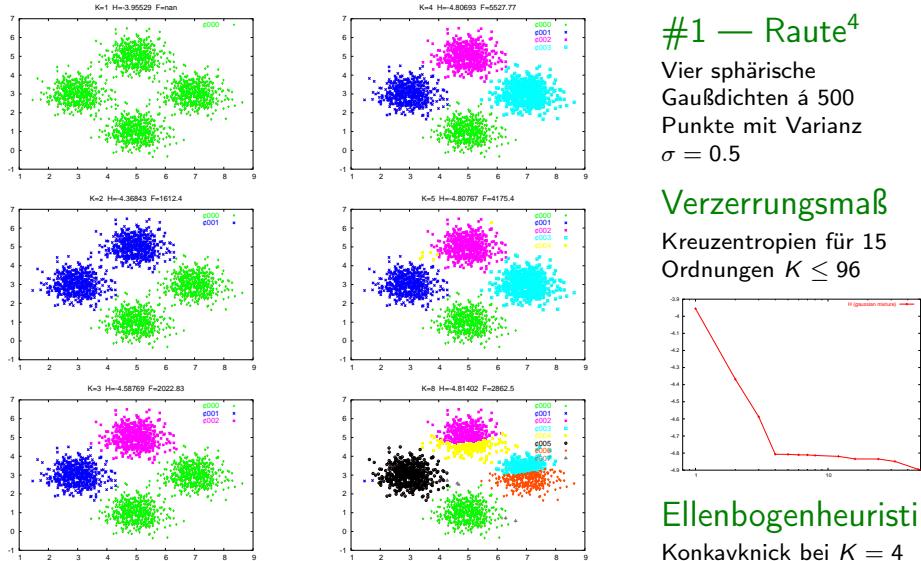
4 AUSGABE

Berechne den Quotienten $C_{\text{hop}} = D_{\bar{\omega}} / (D_{\bar{\omega}} + D_{\omega})$.

(summlingA)

Die beste Anzahl von Gruppen

Gretchenfrage $K \in \{1, 2, 3, 4, 5, 8\}$ bei der GMM-Identifikation



Ordnung und Güte einer Gruppierung

Je mehr Gruppen — desto paßgenauer das Datenmodell

Lemma

Für einen Datensatz $\omega = \{\mathbf{x}_1, \dots, \mathbf{x}_T\} \subset \mathbb{R}^N$ und die Ordnung $K \in \mathbb{N}$ bezeichne

$$\varepsilon_{VQ}^{(K)}(\omega) \stackrel{\text{def}}{=} \underset{\mathbf{z}_1, \dots, \mathbf{z}_K}{\operatorname{argmin}} \sum_{t=1}^T \min_{\kappa} \|\mathbf{x}_t - \mathbf{z}_\kappa\|^2$$

die minimale **quadratisch-euklidische Verzerrung** der K -Partition und

$$\ell_{GMM}^{(K)}(\omega) \stackrel{\text{def}}{=} \underset{\theta \in \mathcal{M}_K}{\operatorname{argmax}} \sum_{t=1}^T \log \sum_{\kappa=1}^K \left\{ \pi_\kappa^{(\theta)} \cdot \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_\kappa^{(\theta)}, \boldsymbol{\Sigma}_\kappa^{(\theta)}) \right\}$$

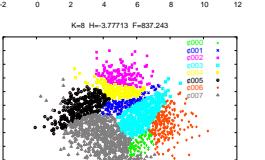
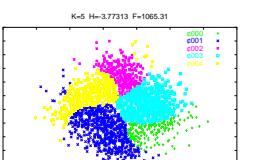
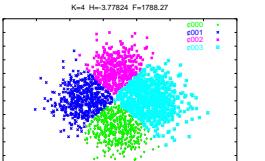
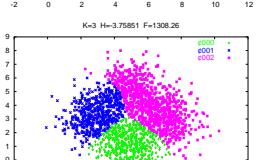
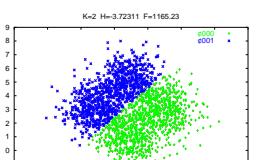
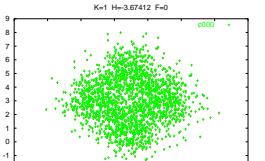
die maximale Güte eines **K -Mischverteilungsmodells** für den Datensatz.

Dann gelten die Antitonie und die Monotonie

$$K \leq K' \Rightarrow \begin{cases} \varepsilon_{VQ}^{(K)}(\omega) & \geq \varepsilon_{VQ}^{(K')}(\omega) \\ \ell_{GMM}^{(K)}(\omega) & \leq \ell_{GMM}^{(K')}(\omega) \end{cases} .$$

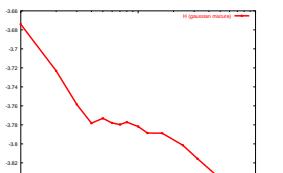
Die beste Anzahl von Gruppen

kann oft durch den Pseudo- F -Wert ermittelt werden

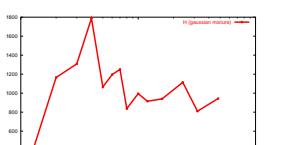


#2 — Raute⁴

Vier sphärische Gaußdichten mit Varianz $\sigma = 1.0$



Negatives $\ell_{GMM}^{(K)}$ (EM)



Der Pseudo- F -Wert

belohnt gute Gruppentrennung & bestraft große Gruppenanzahl

Problem

Die allermeisten Gruppierungskriterien **verbessern sich systematisch** mit wachsender Gruppenzahl K , sind also zur Auswahl der Gruppenzahl **völlig ungeeignet**.

Bemerkung

Die Monotonie ist in den meisten Kurven verletzt, denn K-means und GMM-Identifikation sind **lokale** Optimierungsverfahren.

Definition (Calinski-Harabasz)

Die Vergleichsgröße

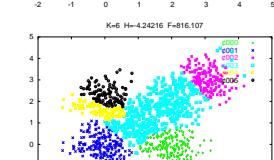
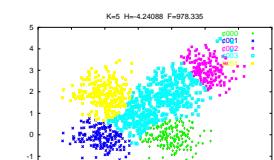
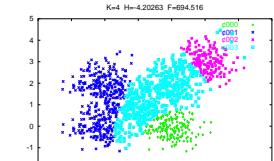
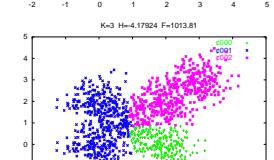
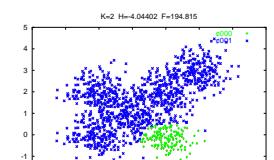
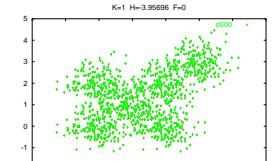
$$C_{\text{pseudo}}(\{\omega_1, \dots, \omega_K\}) \stackrel{\text{def}}{=} \frac{\text{spur}(\mathbf{S}_B) / (K-1)}{\text{spur}(\mathbf{S}_W) / (T-K)}$$

heißt **Pseudo- F -Wert** der Gruppierung $\{\omega_1, \dots, \omega_K\}$.

Es bezeichnen \mathbf{S}_B die **Zwischengruppenstreuung** und \mathbf{S}_W die **Innergruppenstreuung** der Datenpartition.

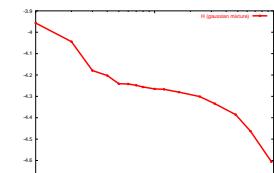
Hier versagt der Pseudo- F -Wert !

Der EM-Algorithmus zur Mischungsidentifikation findet nur suboptimale GMM

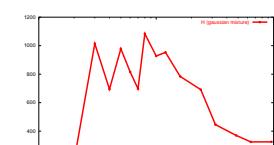


#3 - Kreuz⁶

Sechs Dichten in schrägliegender Kreuzform



Negatives $\ell_{GMM}^{(K)}$ (EM)



Dreiphasige Gruppierung

(Algorithmus)

1 DIVISIVE GRUPPIERUNG

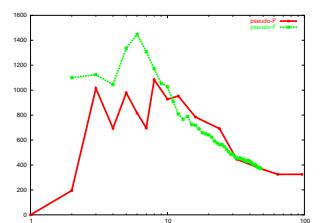
Sukzessive Zerlegung mit 2-means in $K_{\max} = 2^b$ Gruppen.

2 GMM-IDENTIFIKATION

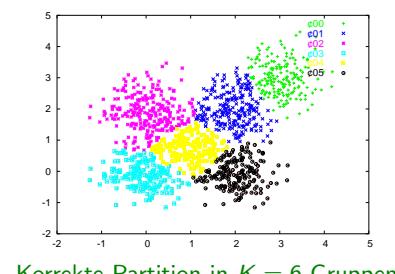
Austauschiteration mit Gaußschem Mischverteilungsmodell (EM-Schritte).

3 REAGGLOMERATION

Bottom-up Gruppierung durch sukzessive ℓ_{GMM} -Maximierung.



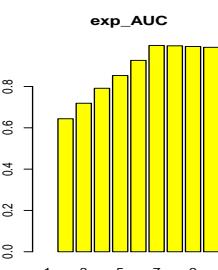
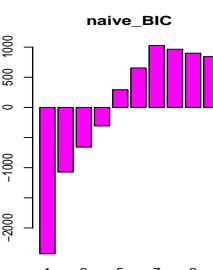
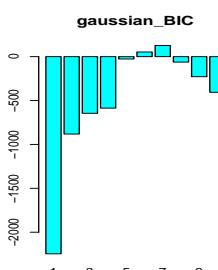
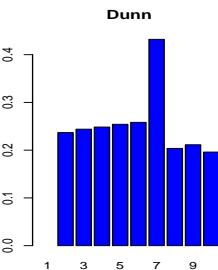
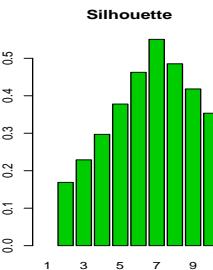
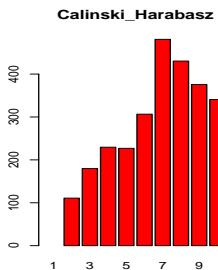
C_{pseudo} für EM & Reagglomeration



Korrekte Partition in $K = 6$ Gruppen

Beispiel: regularisierte Gütemaße im Vergleich

Sieben Cluster im \mathbb{R}^7 · Hierarchische Gruppierung für $K = 1, 2, \dots, 10$



Regularisierte Gütemaße für scharfe Gruppierungen

„cluster validity index“

Dunn's ISODATA

$$C_{\text{iso}} \stackrel{\text{def}}{=} \frac{\min_{\kappa \neq \lambda} \min_{x \in \omega_\kappa} \min_{y \in \omega_\lambda} d(x, y)}{\max_\kappa \max_{x \in \omega_\kappa} \max_{y \in \omega_\kappa} d(x, y)}$$

Rousseeuw's Silhouette

$$C_{\text{sil}} \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^T \frac{b_t - a_t}{\max(a_t, b_t)}$$

mit $D_{\kappa, t} = \overline{d(\omega_\kappa, x_t)}$ und $\begin{cases} a_t = D_{\kappa(t), t} \\ b_t = \min_{\lambda \neq \kappa(t)} D_{\lambda, t} \end{cases}$

Expected Area Under Curve

$$C_{\text{auc}} \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^T \text{AUC}([d(x_*, x_t)], [x_* \in \omega_{\kappa(t)}])$$

Uni-/multivariat gaußsches BIC

$$C_{\text{bic}} \stackrel{\text{def}}{=} -\log \ell_{\text{GMM}}^{(K)} + \log(T) \cdot \text{df}(K, N) \quad \text{mit } \text{df}(K, N) = \begin{cases} K + 2NK & (\text{naiv}) \\ K + (N+3) \frac{NK}{2} & (\text{sonst}) \end{cases}$$

Gütemaße für unscharfe Gruppierungen

Tendenz zur monotonen Verbesserung mit der Gruppenanzahl K

Partitionskoeffizient

$$C_{\text{part}} \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^T \sum_{\kappa=1}^K u_\kappa^2(x_t) \xrightarrow{!} \max$$

Proportionsexponent

$$C_{\text{prop}} \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^T \max_{\kappa} u_\kappa(x_t) \xrightarrow{!} \max$$

Klassifikationsentropie

$$C_{\text{centro}} \stackrel{\text{def}}{=} -\frac{1}{T} \sum_{t=1}^T \sum_{\kappa=1}^K u_\kappa(x_t) \cdot \log_2 u_\kappa(x_t) \xrightarrow{!} \min$$

Bemerkung

Partitionskoeffizient C_{part} , Proportionsexponent C_{prop} und Klassifikationsentropie C_{centro} nehmen ihre Optimalwerte (1/1/0) für die **scharfen** Gruppierungen $\{u_\kappa(\cdot)\}$ an.

Mischungsidentifikation in

Ein Zoo konkurrierender Modelle — Kovarianzauslegung $S_\kappa := \sigma_\kappa^2 \cdot U_\kappa D_\kappa U_\kappa^\top$

Sphärische Modelle

- EII $\mathcal{N}(\mu_\kappa, \sigma^2 \cdot E)$
- VII $\mathcal{N}(\mu_\kappa, \sigma_\kappa^2 \cdot E)$

Global oder κ -variabel?

- Volumen, Gesamtstreuung σ_κ^2
- Gestalt, Dynamik D_κ
- Orientierung, Hauptachsen U_κ

Diagonale Modelle

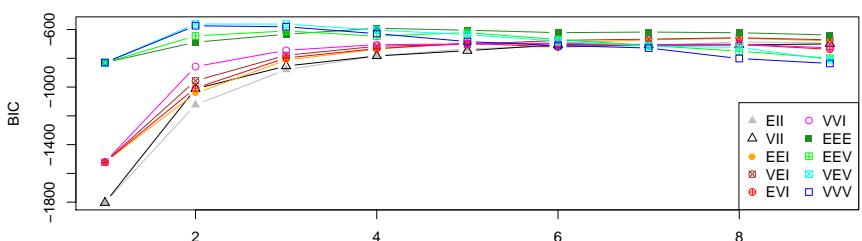
- EEI $\mathcal{N}(\mu_\kappa, \sigma^2 \cdot D)$
- VEI $\mathcal{N}(\mu_\kappa, \sigma_\kappa^2 \cdot D)$
- EVI $\mathcal{N}(\mu_\kappa, \sigma^2 \cdot D_\kappa)$
- VVI $\mathcal{N}(\mu_\kappa, \sigma_\kappa^2 \cdot D_\kappa)$

Ellipsoidale Modelle

- EEE $\mathcal{N}(\mu_\kappa, \sigma^2 \cdot U D U^\top)$
- EEV $\mathcal{N}(\mu_\kappa, \sigma^2 \cdot U_\kappa D U_\kappa^\top)$
- VEV $\mathcal{N}(\mu_\kappa, \sigma_\kappa^2 \cdot U_\kappa D U_\kappa^\top)$
- VVV $\mathcal{N}(\mu_\kappa, \sigma_\kappa^2 \cdot U_\kappa D_\kappa U_\kappa^\top)$

Mischungsidentifikation in

BIC-Entscheidung: optimales GM-Modell & optimale Gruppenzahl K



Agglomeratives Gruppieren

`hc (modelName, data, ...)`

'R'-Paket: `mclust`

EM-Iteration

`em (modelName, data, parameters, ...)`

Start mit E-Schritt

`me (modelName, data, z, ...)`

Start mit M-Schritt

Bayes-Informationsmaß

`Mclust (data, G=1:9, modelNames)` alle $K \in \{1, \dots, 9\}$, alle GMM-Typen

Hierarchische Gruppierung: agglomerativ/divisiv

Austauschverfahren: (fuzzy) K-means

Mischungsidentifikation

Relationale Gruppierung

Konzeptuelle Gruppierung

Spektrale Gruppierung

Clustergütemaße

Zusammenfassung

Zusammenfassung (5)

- Das Ziel der **Gruppierung (Clusteranalyse)** ist die **unüberwachte** Zerlegung eines Datensatzes in **explizit** oder **implizit** charakterisierte Teilmengen von Objekten.
- Die **hierarchischen Verfahren** arbeiten *bottom-up* (**agglomerativ**) oder *top-down* (**divisiv**); das Resultat ist ein **Gruppendendrogramm**.
- Die **Austauschverfahren** geben eine **Anzahl** $K \in \mathbb{N}$ vor und erzeugen **scharfe** (*K-means*) oder **unscharfe** (*fuzzy K-means*) K -Partitionen.
- Die **EM-Gruppierung** modelliert die Daten durch Identifikation einer gaußschen **Mischverteilung**.
- Neben **sphärischen** Gruppen lassen sich auch **rangdefiziente** Ballungsgebiete ermitteln (**K-Elliotypes** oder **Probabilistic PCA**).
- Relationale** Datensätze werden entweder *agglomerativ* gruppiert oder — wie auch **nominal** skalierte Objekte — mit einem **K-medoids**-Austauschverfahren (*RACE*, *K-Sterne*).
- Die **spektrale** Methode löst eine **Minimalschnittaufgabe** im Affinitätsgraphen der Datenobjekte und läuft auf eine Art gaußschen **Kernel-K-means**-Algorithmus hinaus.
- Ermittlung der **Clusteranzahl** durch **Ellenbogenheuristik** oder **Pseudo-F-Wert**.

MASCHINELLES LERNEN & DATAMINING

Vorlesung im Wintersemester 2020

Prof. E.G. Schukat-Talamazzini

Stand: 9. Februar 2021

Teil VI

Attributabhängigkeiten: graphische & kausale Modelle

Analyse von Attributabhängigkeiten

Dependenzanalyse $\stackrel{?}{=}$ Spaltengruppierung

	X_1	X_2	X_3	X_4
o_1	+	low	1.2	+4
o_2	-	hi	0.5	-3
o_3	+	hi	2.3	-3
o_4	+	med	2.1	-7

Abhängigkeit $\not\equiv$ Ähnlichkeit

- Lineare Abhängigkeiten
 $E = m \cdot c^2$
- Skalenempfindlichkeit
Temperatur in $^{\circ}\text{C}$ oder $^{\circ}\text{K}$
- Skalenübergreifend
 X_i Geschlecht, X_j Gehalt

Struktur $\not\equiv$ Partition

- keine Äquivalenzrelation
Zeitreihen, Ortsgitter
- keine binäre Relation
Alter, Geschlecht, Größe
- Kausalitätsrichtung ?
 X_n Niederschlag, X_m Ertrag

Analyse von Attributabhängigkeiten

Mit welchem Ziel — zu welchem Zweck ?

Attributwerteprädiktion

Starke Abhängigkeit \Rightarrow $\begin{cases} \text{geringe a posteriori Streuung} \\ \text{kleiner Regressionsfehler} \end{cases}$

- Voraussage · Imputation · Klassifikation

Strukturaufklärung

Lernen des am einfachsten strukturierten Datenmodells (*Occams Razor*)

- Interaktionen · Kausalitäten · Assoziationsregeln

Robuste Datenmodelle

Netzwerk ausgewählter Abhängigkeiten statt **saturierter** W-Modelle

- geringe Kapazität · hohe Effizienz (Zeit/Speicher) · gute Induktivität

Korrelation, Regression und Transformation

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Statistische Unabhängigkeit von Zufallsvariablen

Statistische Unabhängigkeit

von Zufallsvariablen $\mathbb{X}_1, \dots, \mathbb{X}_N$, falls für alle $x_1, \dots, x_N \in \mathbb{R}^N$ gilt:

$$P(\mathbb{X}_1 = x_1, \dots, \mathbb{X}_N = x_N) = \prod_{n=1}^N P(\mathbb{X}_n = x_n)$$

Statistische Unkorreliertheit

von Zufallsvariablen $\mathbb{X}_1, \dots, \mathbb{X}_N$, falls für alle $x_1, \dots, x_N \in \mathbb{R}^N$ gilt:

$$\mathcal{E}\left[\prod_{n=1}^N \mathbb{X}_n\right] = \prod_{n=1}^N \mathcal{E}[\mathbb{X}_n]$$

Bemerkungen

1. Aus der Unabhängigkeit folgt die Unkorreliertheit, aber nicht umgekehrt.
2. Für normalverteilte $\mathbb{X} \sim \mathcal{N}(\mu, \mathbf{S})$ gilt: $\mathbb{X}_i, \mathbb{X}_j$ korreliert gdw. $\sigma_{ij} \neq 0$.

Statistische Unabhängigkeit von Ereignissen

Paarweise statistische Unabhängigkeit

Faktorisierbarkeit oder (falls $P(A) \neq 0$) Neutralität:

$$A \not\sim B \quad \Leftrightarrow \quad P(A, B) = P(A) \cdot P(B) \quad \Leftrightarrow \quad P(B|A) = P(B)$$

Beispiel: der Wurf zweier fairer Würfel

$$\begin{aligned} A &= \text{"gerade Augensumme"} & P(A, B) &= \frac{1}{12} = \frac{1}{2} \cdot \frac{1}{6} = P(A) \cdot P(B) \\ B &= \text{"erster Wurf ist sechs"} & P(B, C) &= \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = P(B) \cdot P(C) \\ C &= \text{"Augensumme ist sieben"} & P(A, C) &= 0 \neq \frac{1}{2} \cdot \frac{1}{6} = P(A) \cdot P(C) \end{aligned}$$

Stat. Unabhängig. \Rightarrow paarweise s.U.
Stat. Unabhängig. $\not\Rightarrow$ paarweise s.U.

$A = \text{"erster Wurf hat gerade Augenzahl"}$
 $B = \text{"zweiter Wurf hat gerade Augenzahl"}$
 $C = \text{"Augensumme ist ungerade"}$

$$P(A, B, C) = 0 \neq \frac{1}{8} = P(A) \cdot P(B) \cdot P(C)$$

Statistische Unabhängigkeit

der Ereignisse A_1, \dots, A_l , falls für alle Indexmengen $\mathcal{I} \subseteq \{1, \dots, l\}$:

$$P\left(\bigwedge_{i \in \mathcal{I}} A_i\right) = \prod_{i \in \mathcal{I}} P(A_i)$$

Beweis.

1. Die uniforme Verteilungsdichte auf dem Träger

$$\{(x, y) \mid |x| + |y| \leq 1\} \subseteq \mathbb{R}^2$$

ist wegen

$$\mathcal{E}[\mathbb{X}\mathbb{Y}] = 0 = \mathcal{E}[\mathbb{X}] \cdot \mathcal{E}[\mathbb{Y}]$$

zwar unkorreliert, aus ihrer (hypothetischen) Unabhängigkeit folgt aber wegen

$$P(0, \cdot) \cdot P(\cdot, 0) = P(0, 0) = P(0, 1) = P(0, \cdot) \cdot P(\cdot, 1)$$

und $P(0, \cdot) \neq 0$ sofort der Widerspruch $P(\cdot, 0) = P(\cdot, 1)$.

2. Es gilt nach Kovarianzdefinition

$$\sigma_{ij} = \text{Cov}[\mathbb{X}_i, \mathbb{X}_j] = \mathcal{E}[\mathbb{X}_i \mathbb{X}_j] - \mathcal{E}[\mathbb{X}_i] \cdot \mathcal{E}[\mathbb{X}_j];$$

daraus folgt die Behauptung — auch für nicht-normal verteilte Variablen.



Korrelation und Kovarianz

Definition

Es sei $\omega \subset \mathbb{R}^N$ ein Datensatz mit der (empirischen) Kovarianzmatrix $\mathbf{S} = [\sigma_{ij}]$. Die Zahlen

$$\rho_{ij} \stackrel{\text{def}}{=} \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}} \cdot \sqrt{\sigma_{jj}}}, \quad 1 \leq i, j \leq N$$

heißen Pearson'sche **Korrelationskoeffizienten** der Attributpaare (x_i, x_j) .

Bemerkungen

1. Betragmäßig kleine (große) Werte σ_{ij} markieren einen schwachen (starken) Zusammenhang zwischen x_i und x_j .
2. Die **Kovarianzen** sind aber extrem skalierungsempfindlich (σ_{ii} , σ_{jj}).
3. Die Korrelationskoeffizienten ρ_{ij} liegen stets im Intervall $[-1, +1]$.
4. Der Wert $\rho_{ij} = 0$ markiert Unkorreliertheit, die Werte $\rho_{ij} \in \{+1, -1\}$ hingegen **deterministische Abhängigkeit** (mit positiver/negativer Steigung).

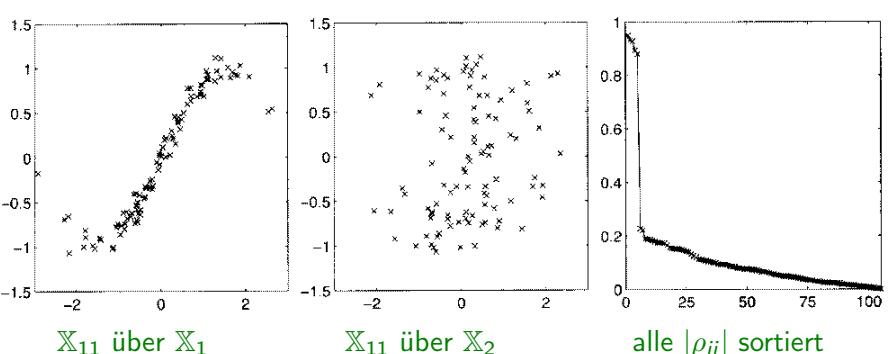
Beispiel — Korrelationsanalyse synthetischer Daten

Zufällig generierte Datenvektoren

$\omega = \{\mathbf{x}_1, \dots, \mathbf{x}_{100}\} \subset \mathbb{R}^{15}$ mit Wertetupeln der Zufallsvariablen

$$\mathbb{X}_n = \begin{cases} \mathcal{N}(0, 1) & n = 1, \dots, 10 \\ \sin(\mathbb{X}_{n-10}) + \mathcal{N}(0, 1/10) & n = 11, \dots, 15 \end{cases}$$

(10 Kanäle weißes Rauschen & 5 Kanäle verrauschte Sinuskopien)



Korrelationsgruppierung

(Algorithmus)

GEGEBEN:

Daten $\omega \subset \mathbb{R}^N$, Schwellen θ_ρ , „leere“ Gruppierung $\gamma : i \mapsto \perp$.

- 1 INITIALISIERUNG
Berechne alle Korrelationskoeffizienten ρ_{ij} .
- 2 ABSTEIGEND SORTIEREN

$$|\rho_{i_1 j_1}| \geq |\rho_{i_2 j_2}| \geq |\rho_{i_3 j_3}| \geq |\rho_{i_4 j_4}| \geq \dots \geq \dots \geq$$

- 3 FÜR ALLE $r = 1, 2, \dots, N(N-1)/2$:

1. Wenn $|\rho_{i_r j_r}| < \theta_\rho$ dann \rightsquigarrow ENDE.
2. Wenn $\gamma(n) \neq \perp$ für alle n dann \rightsquigarrow ENDE.
3. Wenn $\gamma(i_r) = \perp = \gamma(j_r)$ dann erzeuge neue Gruppe $\{i_r, j_r\}$.
4. Wenn $\gamma(i_r) = \perp$ dann setze $\gamma(i_r) \leftarrow \gamma(j_r)$.
5. Wenn $\gamma(j_r) = \perp$ dann setze $\gamma(j_r) \leftarrow \gamma(i_r)$.
6. Wenn $\gamma(i_r) \neq \gamma(j_r)$ dann vereinige die Gruppen: $\gamma(i_r) \cup \gamma(j_r)$.

Korrelationsgruppierung

Was tut dieser Algorithmus?

Single-linkage Agglomeration — aber:

Terminiert bei Unterschreiten der Korrelationsschwelle.
Terminiert sobald alle Einermengen „verbraucht“ sind.

Synthesedatenbeispiel

Für die Daten $\omega \subset \mathbb{R}^{15}$ werden in den ersten fünf Schritten die Gruppen

$$\{1, 11\}, \{2, 12\}, \{3, 13\}, \{4, 14\}, \{5, 15\}$$

gebildet; anschließend gibt es jeweils drei gleichwahrscheinliche Möglichkeiten:

1. Eine der „alten“ Gruppen wird mit einem neuen Index aufgefüllt.
2. Zwei „alte“ Gruppen werden vereinigt.
3. Aus zwei „frischen“ Indizes wird eine neue Gruppe gebildet.

Mit der Ausnahme von 1. sind all diese Optionen höchst **unerwünscht**.

Gestörte (lineare) Abhängigkeit

$\mathbb{Y} = f(\mathbb{X}) + \mathbb{E}$ mit Funktionsprototyp $f: \mathbb{R} \rightarrow \mathbb{R}$ und Residuum \mathbb{E}

Lemma

Für zwei normalverteilte Zufallsvariablen \mathbb{X}, \mathbb{Y} mit

$$\mathbb{Y} = a\mathbb{X} + b + \mathbb{E}, \quad \mathbb{X} \sim \mathcal{N}(\mu_x, \sigma_x^2), \quad \mathbb{E} \sim \mathcal{N}(0, \sigma_e^2)$$

gehört \mathbb{Y} der Verteilungsaussage

$$\mathbb{Y} \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2 + \sigma_e^2).$$

Die **Kovarianz** und die **Korrelation** zwischen \mathbb{X} und \mathbb{Y} betragen

$$\sigma_{xy} = a \cdot \sigma_x^2 \quad \text{bzw.} \quad \rho_{xy} = \text{sign}(a) \cdot \left(1 + \frac{\sigma_e^2}{a^2 \cdot \sigma_x^2}\right)^{-\frac{1}{2}}.$$

Bemerkung

Die Korrelation ρ_{xy} erbt das Vorzeichen von a .

Der Betrag wächst und fällt mit σ_e^{-2} im Einheitsintervall.

□

Beweis.

Berechnung der Kovarianz (o.B.d.A. ist $\mu_x = 0$):

$$\begin{aligned} \sigma_{xy} &= \text{Cov}[\mathbb{X}, \mathbb{Y}] = \mathbb{E}[\mathbb{X}\mathbb{Y}] - \mu_x\mu_y \\ &= \mathbb{E}[a\mathbb{X}^2 + b\mathbb{X} + \mathbb{E}\mathbb{X}] - \mu_x\mu_y \\ &= a \cdot (\sigma_x^2 + \mu_x^2) + b\mu_x + \mu_e\mu_x - \mu_x\mu_y \\ &= a\sigma_x^2 + \underbrace{a\mu_x^2 + b\mu_x + \mu_e\mu_x - \mu_x\mu_y}_{\mu_y \cdot \mu_x} \\ &= a\sigma_x^2 \end{aligned}$$

Berechnung der Korrelation:

$$\begin{aligned} \rho_{xy} &= \frac{a\sigma_x^2}{\sqrt{\sigma_x^2 \cdot (a^2\sigma_x^2 + \sigma_e^2)}} \\ &= \text{sign}(a) \cdot \frac{a\sigma_x^2}{a\sigma_x^2 \cdot \sqrt{1 + \sigma_e^2 / (a^2\sigma_x^2)}} \\ &= \text{sign}(a) \cdot \left(1 + \frac{\sigma_e^2}{a^2\sigma_x^2}\right)^{-\frac{1}{2}} \end{aligned}$$

Kausalität und Scheinzusammenhang

Verursacht Diät-Cola wirklich Übergewicht?

Ursache und Wirkung

Korrelation und Abhängigkeit haben keine Vorzugsrichtung:

$$\left\{ \begin{array}{l} \mathbb{X}_i = \text{"Körpergewicht [kg]"} \\ \mathbb{X}_j = \text{"Konsum kalorienreduzierter Getränke [\ell]"} \end{array} \right\}$$

Hohe (positive) Korrelation(en) $\rho_{ij} = \rho_{ji}$ ohne Hinweis auf Kausalrichtung.

Versteckte gemeinsame Ursache oder Lederallergie?

Das Korrelationsmaß reagiert auch auf *scheinbare* Abhängigkeiten.

$$\left\{ \begin{array}{l} \mathbb{X}_i = \text{"Kaffee geröhrt ?"} \\ \mathbb{X}_j = \text{"Kaffee schmeckt süß ?"} \\ \dots \dots \dots \\ \mathbb{X}_k = \text{"Zuckerwürfel drin ?"} \end{array} \right\}$$

$$\left\{ \begin{array}{l} \mathbb{X}_i = \text{"Lederschuhe an ?"} \\ \mathbb{X}_j = \text{"Kopfschmerzen ?"} \\ \dots \dots \dots \\ \mathbb{X}_k = \text{"Wagnergasse bis 3h ?"} \end{array} \right\}$$

Hohe (positive) Korrelation ρ_{ij} ohne jeden (direkten) kausalen Zusammenhang.

(Bi-) Partielle Korrelation

Vergleich nach Subtraktion der Ausgleichsgeraden

Definition

Es seien Zufallsvariablen $\mathbb{X}_1, \dots, \mathbb{X}_N$ gegeben; ferner bezeichne

$$\mathbb{X}_{i|k} = a_{i|k} \cdot \mathbb{X}_k + b_{i|k}, \quad (i, k \in \{1, \dots, N\}, i \neq k)$$

den linearen **Quadratmittelprädiktor** für \mathbb{X}_i aus \mathbb{X}_k („Ausgleichsgerade“).

Dann heißt

$$\rho_{ij|k} \stackrel{\text{def}}{=} \text{Corr}[\mathbb{X}_i - \mathbb{X}_{i|k}, \mathbb{X}_j - \mathbb{X}_{j|k}]$$

die **partielle Korrelation** zwischen \mathbb{X}_i und \mathbb{X}_j hinsichtlich \mathbb{X}_k und es heißt

$$\rho_{i|k,j|\ell} \stackrel{\text{def}}{=} \text{Corr}[\mathbb{X}_i - \mathbb{X}_{i|k}, \mathbb{X}_j - \mathbb{X}_{j|\ell}]$$

die **bipartiale Korrelation** zwischen \mathbb{X}_i und \mathbb{X}_j hinsichtlich \mathbb{X}_k und \mathbb{X}_ℓ .

(Bi-) Partielle Korrelation

Berechnung aus den gewöhnlichen Korrelationskoeffizienten

Lemma

Es seien die Zufallsvariablen $\mathbb{X}_1, \dots, \mathbb{X}_N$ und ihre Korrelationen ρ_{ij} , $i, j \in \{1, \dots, N\}$ gegeben.

1. Die partielle Korrelation zwischen \mathbb{X}_i und \mathbb{X}_j ohne den Einfluß von \mathbb{X}_k hat den Wert

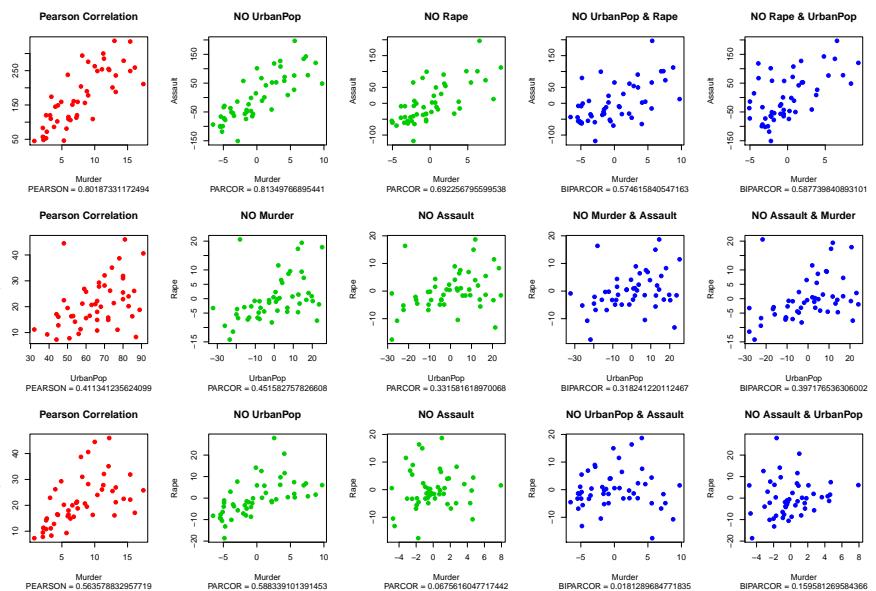
$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik} \cdot \rho_{jk}}{\sqrt{(1 - \rho_{ik}^2) \cdot (1 - \rho_{jk}^2)}}.$$

2. Die bipartiale Korrelation zwischen \mathbb{X}_i und \mathbb{X}_j ohne den Einfluß von \mathbb{X}_k bzw. \mathbb{X}_ℓ hat den Wert

$$\rho_{ijk,j|\ell} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk} - \rho_{i\ell}\rho_{j\ell} + \rho_{i\ell}\rho_{k\ell}\rho_{j\ell}}{\sqrt{(1 - \rho_{ik}^2) \cdot (1 - \rho_{j\ell}^2)}}.$$

Beispiel — U.S. Arrests

Mord/Überfall · Metropol/Vergewaltigung · Mord/Vergewaltigung



Regressionsanalyse

Definition

Eine Familie

$$[f(\cdot | \mathbf{a}) : \mathbb{R}^N \rightarrow \mathbb{R}]_{\mathbf{a} \in \mathcal{M}}$$

von Abbildungen heißt **Funktionsprototyp** der Dimension N ; ein Element $f(\cdot | \mathbf{a})$ der Familie heißt **Funktionsinstanz** zu \mathbf{a} .

Für einen Datensatz $\omega \subset \mathbb{R}^N \times \mathbb{R}$ definieren wir den **Regressionsfehler**

$$\varepsilon(f, \mathbf{a}, \omega) \stackrel{\text{def}}{=} \sum_{(\mathbf{x}, y) \in \omega} (y - f(\mathbf{x}, \mathbf{a}))^2$$

von $f(\cdot | \mathbf{a})$ über ω . Eine Funktionsinstanz $f(\cdot | \mathbf{a}^*)$ mit minimalem Regressionsfehler heißt **Regressionfunktion** von ω , ihre Parameter \mathbf{a}^* heißen **Regressionsparameter**.

Beispiel — lineare Regression

Die spezielle Familie der $f(\cdot | \mathbf{a}) : (x_1, \dots, x_N) \mapsto a_0 + \sum_{n=1}^N a_n x_n$ mit $\mathbf{a} \in \mathbb{R}^{N+1}$ heißt **affiner** oder — im Fall $a_0 \equiv 0$ — **linearer** Funktionsprototyp.

Beispiel — Ausgleichsgerade

Funktionsprototyp der Dimension $N = 1$ \Rightarrow Geradengleichungen $y = a + bx$

Regressionsparameter

für einen gegebenen Datensatz $\omega \subset \mathbb{R} \times \mathbb{R}$

$$b = \frac{\sigma_{xy}}{\sigma_{xx}} \quad \text{und} \quad a = \mu_y - b\mu_x = \mu_y - \frac{\sigma_{xy}}{\sigma_{xx}} \cdot \mu_x$$

Regressionsfehler

einer Geraden $y = a + bx$ (Verschiebung \rightsquigarrow o.B.d.A. $\mu_x = 0$)

$$\begin{aligned} \frac{1}{T} \cdot \varepsilon(a, b, \omega) &= \frac{1}{T} \sum_t (y_t - a - bx_t)^2 = \dots \\ &= \sigma_{yy} + \mu_y^2 + a^2 + b^2 \sigma_{xx} - 2a\mu_y - 2b\sigma_{xy} \\ &= \sigma_{yy} \cdot (1 - \rho_{xy}^2) \quad (\text{Einsetzen } a = \mu_y \text{ und } b = \sigma_{xy}/\sigma_{xx}) \end{aligned}$$

Aufgeklärte Varianz

Die quadrierte Korrelation $\rho_{xy}^2 \in [0, 1]$ ist der proportionale Anteil der Varianz σ_{yy} von \mathbb{Y} , der durch die ZV $\hat{\mathbb{Y}} = a + b \cdot \mathbb{X}$ aufgeklärt werden konnte.

Lineare und nichtlineare Regression

Kein Fall für Ausgleichsgeraden

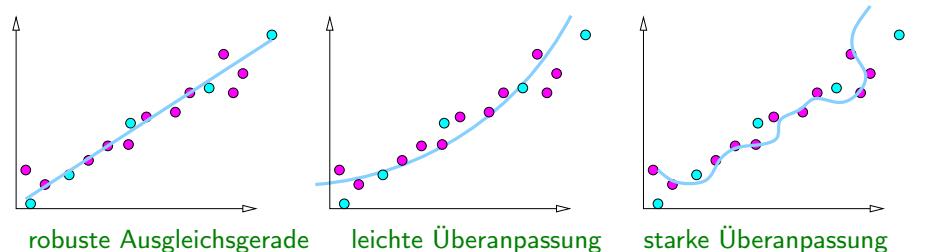
Betrachte die Taylorreihenentwicklung der **sinusoidalen** Abhängigkeit

$$y = \sin(x + \alpha) = \underbrace{\sin \alpha + x \cos \alpha}_{\text{linear}} - x^2 \frac{\sin \alpha}{2} - x^3 \frac{\cos \alpha}{6} \pm \dots$$

Ausgleichspolynome

Affiner Regressionsansatz mit Termexpansion, z.B. polynomial für $N = 3$:

$$(x_1, x_2, x_3) \mapsto (1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1 x_2, x_1 x_3, x_2 x_3, \dots)$$



Lokale Regression

Eine Frage der guten Nachbarschaft

Nächster-Nachbar-Regel

Belegmenge $\omega^{(x)} = \{x_s\}$ ist einelementig.

$$\varepsilon(f, \mathbf{a}, \omega | \mathbf{x}) \stackrel{\text{def}}{=} (y_s - f(\mathbf{x}_s | \mathbf{a}))^2, \quad s = \operatorname{argmin}_{t=1..T} d(\mathbf{x}, \mathbf{x}_t)$$

k-Nächste-Nachbarn-Regel

Scharfe Belegmenge $\omega^{(x)}$ mit genau k Elementen.

$$\varepsilon(f, \mathbf{a}, \omega | \mathbf{x}) \stackrel{\text{def}}{=} \sum_{i=1}^k (y_{s_i} - f(\mathbf{x}_{s_i} | \mathbf{a}))^2$$

Gewichtete Mittelung

Unscharfe Belegmenge $\omega^{(x)}$ mit T Elementen.

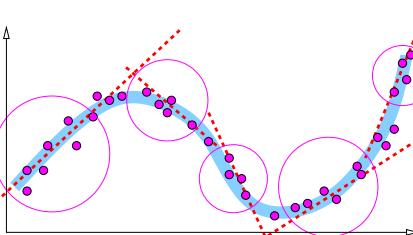
$$\varepsilon(f, \mathbf{a}, \omega | \mathbf{x}) \stackrel{\text{def}}{=} \sum_{t=1}^T w_t \cdot (y_t - f(\mathbf{x}_t | \mathbf{a}))^2, \quad w_t \propto \exp \left\{ -\frac{1}{2\sigma^2} \cdot \|\mathbf{x} - \mathbf{x}_t\|^2 \right\}$$

Lokale Regression

Verzögertes Lernen

- lokales Modell „just in time“
- kein globales Modell

$$f(\mathbf{x}_t | \mathbf{a}^*) \approx y_t (\forall t)$$



GEGEBEN:

Lerndatenprobe $\omega = [(x_t, y_t)]_1^T \subset \mathbb{R}^N \times \mathbb{R}$ und Eingabevektor $\mathbf{z} \in \mathbb{R}^N$

- 1 NACHBARSCHAFT FIXIEREN
Berechne Nachbarschaftsmenge $\omega^{(z)} \subset \omega$, eventuell mit Gewichten $\{w_t\}_1^T$.
- 2 LOKALE AUSGLEICHSSRECHNUNG
Schätze lokale Regressionsfunktion $f(\cdot | \mathbf{a}^{(z)})$ für den $\omega^{(z)}$ -Datensatz.
- 3 VORHERSAGE TREFFEN
Setze $\hat{y}(\mathbf{z}) := f(\mathbf{z} | \mathbf{a}^{(z)})$.

Lokale Regression

Konstante Funktionsprototypen · Disjunkte Nachbarschaften

Konstanter Funktionsprototyp

$$f(\cdot | a) : \begin{cases} \mathbb{R}^N & \rightarrow \mathbb{R} \\ \mathbf{x} & \mapsto a \end{cases}, \quad a \in \mathbb{R}$$

NN-Regel	k -NN-Regel	Distanzgewichte
$f_n(\mathbf{x}) = y_{t_n(\mathbf{x})}$	$f_k(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^k y_{t_i(\mathbf{x})}$	$f_g(\mathbf{x}) = \mathbf{w}^\top \mathbf{y} / \ \mathbf{w}\ _1$
„Kopie“	„Ortsmittel“	„Schwerpunkt“

Stückweise lineare Regression

- 1 GRUPPIERUNG

Lerne extensionale Partition $\omega_1, \dots, \omega_K$ von x_1, \dots, x_T (K -means).

- 2 STÜCKWEISE REGRESSION

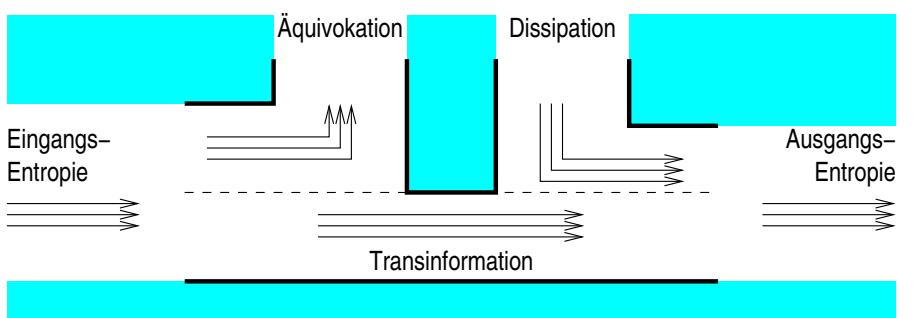
Lerne lokale Regressionsfunktionen $f(\cdot | \mathbf{a}_1), f(\cdot | \mathbf{a}_2), \dots, f(\cdot | \mathbf{a}_K)$.

- 3 VORHERSAGEPHASE

- Bestimme zu $\mathbf{x} \in \mathbb{R}^N$ den Gruppenindex λ , also mit $\mathbf{x} \in \Omega_\lambda \supset \omega_\lambda$.
- Berechne den Vorhersagewert $\hat{y}(\mathbf{x}) = f(\mathbf{x} | \mathbf{a}_\lambda)$.

Informationstheorie

Der gedächtnislose Informationskanal — Claude Shannon, 1949



Der Informationskanal

ist durch die gemeinsame Verteilung $f_{xy}(\cdot, \cdot)$ seiner **Eingangsvariablen** \mathbb{X} und seiner **Ausgangsvariablen** \mathbb{Y} charakterisiert.

Kanalentropien

Eingangsentropie

$$\mathcal{H}(\mathbb{X}) = \mathcal{E}[-\log f_x(\mathbb{X})]$$

Ausgangsentropie

$$\mathcal{H}(\mathbb{Y}) = \mathcal{E}[-\log f_y(\mathbb{Y})]$$

Gesamtentropie

$$\mathcal{H}(\mathbb{XY}) = \mathcal{E}[-\log f_{xy}(\mathbb{X}, \mathbb{Y})]$$

Bedingte Kanalentropien

Was Sie schon immer über Entropien wissen wollten, aber noch nie zu fragen wagten

Definition

Der Informationskanal sei durch f_{xy} charakterisiert.

- $\mathcal{H}(\mathbb{X}|\mathbb{Y}) = \mathcal{E}[-\log f_{x|y}(\mathbb{X}|\mathbb{Y})]$ heißt **Äquivokation** des Kanals.
- $\mathcal{H}(\mathbb{Y}|\mathbb{X}) = \mathcal{E}[-\log f_{y|x}(\mathbb{Y}|\mathbb{X})]$ heißt **Dissipation** des Kanals.
- $\mathfrak{S}(\mathbb{X}; \mathbb{Y}) = \mathcal{E}[-\log \frac{f_{xy}(\mathbb{X}, \mathbb{Y})}{f_{x|y}(\mathbb{X}, \mathbb{Y})}]$ heißt **Transinformation** des Kanals.

Lemma

In einem gedächtnislosen Informationskanal gelten die Aussagen:

1. $\mathcal{H}(\mathbb{X}|\mathbb{Y}) = \mathcal{H}(\mathbb{XY}) - \mathcal{H}(\mathbb{Y})$
2. $\mathcal{H}(\mathbb{Y}|\mathbb{X}) = \mathcal{H}(\mathbb{XY}) - \mathcal{H}(\mathbb{X})$
3. $\mathfrak{S}(\mathbb{X}; \mathbb{Y}) = \mathcal{H}(\mathbb{X}) + \mathcal{H}(\mathbb{Y}) - \mathcal{H}(\mathbb{XY})$
4. $\mathfrak{S}(\mathbb{X}; \mathbb{Y}) = \mathcal{D}(f_{xy} \| f_x \cdot f_y)$

Divergenz

(Kullback-Leibler)

$$\mathcal{D}(f \| g) = \mathcal{E}_f[\log \frac{f}{g}]$$

Beweis.

1. Univariater Fall:

$$\begin{aligned} \mathcal{H}(\mathbb{X}) &= \mathcal{E}[-\log \mathcal{N}(\mathbb{X} | \mu, \sigma^2)] = \mathcal{E}\left[\frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2} \left(\frac{\mathbb{X} - \mu}{\sigma}\right)^2\right] \\ &= \mathcal{E}\left[\frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2} \tilde{\mathbb{X}}^2\right] = \frac{1}{2} \cdot (\log \sigma^2 + 1 + \log(2\pi)) \end{aligned}$$

Beachte, daß $\tilde{\mathbb{X}} = (\mathbb{X} - \mu)/\sigma$ standardnormalverteilt ist, d.h. $\tilde{\mathbb{X}} \sim \mathcal{N}(0, 1)$.

2. Multivariater Fall:

$$\begin{aligned} \mathcal{H}(\mathbb{X}) &= \mathcal{E}[-\log \mathcal{N}(\mathbb{X} | \boldsymbol{\mu}, \mathbf{S})] = \mathcal{E}\left[\frac{1}{2} \log \det(2\pi\mathbf{S}) + \frac{1}{2} \cdot (\mathbb{X} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\mathbb{X} - \boldsymbol{\mu})\right] \\ &= \mathcal{E}\left[\frac{1}{2} \log \det(2\pi\mathbf{S}) + \frac{1}{2} \cdot \tilde{\mathbb{X}}^\top \tilde{\mathbb{X}}\right] = \frac{1}{2} \cdot (\log \det(\mathbf{S}) + N + N \log(2\pi)) \end{aligned}$$

Bivariate Fall: gilt wegen $\det \begin{pmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ji} & \sigma_{jj} \end{pmatrix} = \sigma_{ii}\sigma_{jj} - \sigma_{ij}^2$.

3. Transinformationen:

$$\begin{aligned} \mathfrak{S}(\mathbb{X}_i; \mathbb{X}_j) &= \mathcal{H}(\mathbb{X}_i) + \mathcal{H}(\mathbb{X}_j) - \mathcal{H}(\mathbb{X}_i \mathbb{X}_j) \\ &= +\frac{1}{2} \cdot \log \left(\frac{\sigma_{ii}\sigma_{jj}}{\sigma_{ii}\sigma_{jj} - \sigma_{ij}^2} \right) = -\frac{1}{2} \cdot \log (1 - \rho_{ij}^2) \end{aligned}$$

$(1 - \rho_{ij}^2)$ ist der Anteil **unaufgeklärter Varianz**.

Transformation normalverteilter Attribute

Lemma

Für die (differentiellen) Entropien und die Transinformation normalverteilter Zufallsvariablen gelten die nachfolgenden Aussagen:

1. Wenn $\mathbb{X} \sim \mathcal{N}(\mu, \sigma^2)$, so gilt:

$$\mathcal{H}(\mathbb{X}) = \frac{1}{2} \cdot (\log \sigma^2 + 1 + \log(2\pi))$$

2. Wenn $(\mathbb{X}_1, \dots, \mathbb{X}_N) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{S})$, so gilt:

$$\mathcal{H}(\mathbb{X}_1 \dots \mathbb{X}_N) = \frac{1}{2} \cdot (\log \det(\mathbf{S}) + N + N \log(2\pi))$$

$$\mathcal{H}(\mathbb{X}_i \mathbb{X}_j) = \frac{1}{2} \cdot \log(\sigma_{ii} \cdot \sigma_{jj} - \sigma_{ij}^2) + 1 + \log(2\pi)$$

3. Für jedes bivariat normale Variablenpaar $(\mathbb{X}_i, \mathbb{X}_j)$ gilt:

$$\mathfrak{S}(\mathbb{X}_i; \mathbb{X}_j) = -\frac{1}{2} \cdot \log (1 - \rho_{ij}^2)$$

Transformation diskreter Attribute

Wertebereiche und Verteilung

Es sei $\mathbb{X} \in \{\xi_1, \dots, \xi_K\}$ und $\mathbb{Y} \in \{\eta_1, \dots, \eta_L\}$ verteilt gemäß

$$p_{k\ell} = P(\mathbb{X} = \xi_k, \mathbb{Y} = \eta_\ell)$$

Marginale und gemeinsame Entropien

$$\begin{aligned} \mathcal{H}(\mathbb{X}\mathbb{Y}) &= - \sum_k \sum_\ell p_{k\ell} \cdot \log p_{k\ell} \\ \mathcal{H}(\mathbb{X}) &= - \sum_k \left(\sum_\ell p_{k\ell} \right) \cdot \log \left(\sum_\ell p_{k\ell} \right) \\ \mathcal{H}(\mathbb{Y}) &= - \sum_\ell \left(\sum_k p_{k\ell} \right) \cdot \log \left(\sum_k p_{k\ell} \right) \end{aligned} \quad \Rightarrow \quad \underbrace{- \sum_{k,\ell} p_{k\ell} \cdot \log \frac{p_{k\cdot} \cdot p_{\cdot\ell}}{p_{k\ell}}}_{\text{Transformation } \mathfrak{S}(\mathbb{X}; \mathbb{Y})}$$

Korrelation, Regression und Transformation

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Transformation gemischtskaliger Attribute

$\mathbb{X} \in \mathbb{IR}$ und $\mathbb{Y} \in \{\eta_1, \dots, \eta_L\}$

Punktweise Transformation

$$\mathfrak{S}(\mathbb{X}; \mathbb{Y}) = \int_x \sum_y f(x, y) \cdot \mathfrak{S}(x; y)$$

Die „mutual information“ zwischen korrespondierenden Werten x und y :

$$\log \frac{f(x|y)}{f(x)} = \underbrace{\log \frac{f(x, y)}{f(x) \cdot f(y)}}_{\mathfrak{S}(x; y)} = \log \frac{f(y|x)}{f(y)}$$

Faktor diskret

Gaußsche Mischverteilung

$$f(x, \eta_\ell) = \pi_\ell \cdot \mathcal{N}(x | \mu_\ell, \sigma_\ell^2)$$

Schätzformel

$$\sum_{t=1}^T \frac{1}{T} \cdot \log \frac{\mathcal{N}(x_t | \mu_{\ell(t)}, \sigma_{\ell(t)}^2)}{\sum_\ell \pi_\ell \cdot \mathcal{N}(x_t | \mu_\ell, \sigma_\ell^2)}$$

Faktor stetig

Diskriminantverteilung

$$f(x, \eta_\ell) = f_{\mathbb{X}}(x) \cdot p(\eta_\ell | x)$$

Schätzformel

$$\sum_{t=1}^T \frac{1}{T} \cdot \log \frac{p(\eta_{\ell(t)} | x_t)}{\pi_{\ell(t)}}$$

Assoziationsanalyse

Agrawal (SIGMOD Conference 1993) — mehr als 6.000× zitiert!

Warenkorbdaten

Objekte = qualitative **Stücklisten**

$$\rightsquigarrow \Omega = \mathfrak{P}(\mathfrak{A})$$

$$\omega \subset \Omega = \{0, 1\}^N$$

über einem globalen **Artikelinventar** $\mathfrak{A} = \{\mathfrak{a}_1, \dots, \mathfrak{a}_N\}$

Assoziationsregeln

„Wer alle Produkte aus A kauft, der kauft auch alle Produkte aus B.“

$$\text{IF } A \text{ THEN } B , \quad A, B \in \Omega , \quad A \cap B = \emptyset$$

Beispielregeln

IF {Windeln} THEN {Bier}

IF {Brot, Butter} THEN {Milch}

IF {Rosen, Wein, Goldbären} THEN {Kondome}

Bemerkungen

1. Warenkorbdaten haben **binäre Attribute**.

2. Assoziationsregeln formulieren **mehrere Abhängigkeiten**.

Gute und schlechte Regeln

Abdeckungs- und Geltungsgrad einer Regel · Signifikanz ihrer Prämissen

Definition

Es sei $\omega \subset \Omega$ ein Datensatz, $A, B \in \Omega$ zwei Stücklisten und
IF A THEN B (kürzer: $A \rightarrow B$) eine Assoziationsregel. Die Größe

$$\text{supp}(A \rightarrow B) \stackrel{\text{def}}{=} \text{supp}(A \cup B), \quad \text{supp}(A) \stackrel{\text{def}}{=} \frac{|\{x \in \omega \mid x \supseteq A\}|}{|\omega|}$$

heißt **Support**,

$$\rightsquigarrow \hat{P}(A \cup B)$$

$$\text{conf}(A \rightarrow B) \stackrel{\text{def}}{=} \frac{\text{supp}(A \cup B)}{\text{supp}(A)}$$

heißt **Konfidenz** und

$$\rightsquigarrow \hat{P}(B|A)$$

$$\text{lift}(A \rightarrow B) \stackrel{\text{def}}{=} \frac{\text{supp}(A \cup B)}{\text{supp}(A) \cdot \text{supp}(B)}$$

heißt **Relevanz** der Assoziation $A \rightarrow B$.

$$\rightsquigarrow \frac{\hat{P}(B|A)}{\hat{P}(B)}$$

Apriori-Basisalgorithmus

Schichtenweise Stücklisten- und Regelgenerierung

(Algorithmus)

GEGEBEN

Warenkorbdaten ω , Stückzahlgrenze N^* , Schwellen $\theta_s, \theta_c, \theta_r$.

1 INITIALISIERUNG

$$\mathcal{M}_1 \leftarrow \{\{i\} \mid \text{supp}(\{i\}) \geq \theta_s\}, \quad \mathcal{R} = \emptyset$$

2 SCHICHTEXPANSION ($n = 2, \dots, N^*$)

Erzeuge alle

$$A = B \cup \{i\} \quad \text{mit } B \in \mathcal{M}_{n-1}, \{i\} \in \mathcal{M}_1 \text{ und } i \notin B.$$

Bringe A nach \mathcal{M}_n falls $\text{supp}(A) \geq \theta_s$.

3 REGELERZEUGUNG

Für alle $C \in \mathcal{M} = \bigcup_{n=1}^{N^*} \mathcal{M}_n$:

Für alle Artikel $j \in C$ teste

$$\text{conf}(C \setminus \{j\} \rightarrow \{j\}) \geq \theta_c \wedge \text{lift}(C \setminus \{j\} \rightarrow \{j\}) \geq \theta_r$$

und verbringe die Regel im Erfolgsfall nach \mathcal{R} .

(summlösungA)

Extraktion nützlicher Assoziationsregeln

Eine Frage des Aufwandes

Aufgabenstellung

Gesucht ist — bei gegebenen Warenkorbdaten — die Teilmenge solcher Regeln $A \rightarrow B$

- mit **signifikantem Abdeckungsgrad**
- und hohem **Geltungsgrad**
- und erheblicher **Aussagekraft**.

$$\text{supp}(A \rightarrow B) \geq \theta_s$$

$$\text{conf}(A \rightarrow B) \geq \theta_c$$

$$\text{lift}(A \rightarrow B) \geq \theta_r$$

Problem

Es gibt 2^N kombinatorisch mögliche Stücklisten und es gibt 3^N mögliche Assoziationsregeln ($N = |\mathfrak{A}|$).

Lösungsansatz

Es gilt die **Antitone**

$$A \supseteq B \Rightarrow \text{supp}(A) \leq \text{supp}(B)$$

und es gibt nur N **einelementige** Stücklisten.

Regelformat und Artikelbeschreibung

Assoziationsregeln mit multipler Conclusio

Statt einfacher Regeln $C \setminus \{j\} \rightarrow \{j\}$ produziere

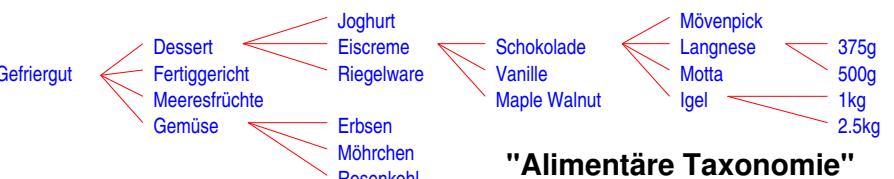
$$A \rightarrow B \quad \text{mit } A \cap B = \emptyset \text{ und } A \cup B = C.$$

Stufenweise Erzeugung („bottom-up“) unter Verwendung der Monotonie:

$$B_1 \subseteq B_2 \Rightarrow \begin{cases} \text{supp}(A_1 \rightarrow B_1) = \text{supp}(A_2 \rightarrow B_2) \\ \text{conf}(A_1 \rightarrow B_1) \geq \text{conf}(A_2 \rightarrow B_2) \\ \text{lift}(A_1 \rightarrow B_1) \geq \text{lift}(A_2 \rightarrow B_2) \end{cases}$$

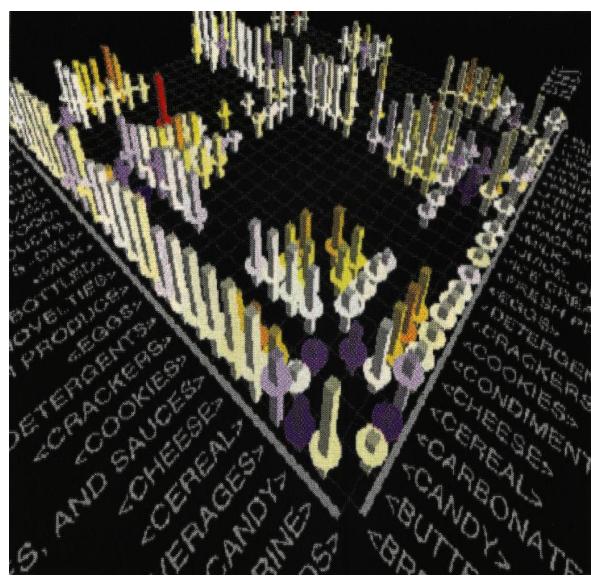
Aufwandsreduktion durch ontologische Gliederung

Artikeleinträge werden durch ihre Verallgemeinerungen aufgestockt.



Grafische Darstellung von Assoziationsregeln

Warenkorbdaten der Einzelhandelskette „WalMart“



Ikonendarstellung

SGI MineSet

L+R Regelseite

2D-Gitterposition

Konfidenz

Balkenhöhe

Support

Scheibenhöhe

Interessantheit

Farbgebung

Assoziationsanalyse für (mehrwertige) Nominalskalen

Verallgemeinerte Stücklisten

Listen von kontradiktionsfreien Attribut-Wert-Paaren:

(windy = false, play = no, outlook = sunny, humidity = high)

Es gibt $\prod_n (L_n + 1)$ Stücklisten und $\prod_n (2 \cdot L_n + 1)$ Assoziationsregeln.

Beispielregeln

Klassisch:

- IF {Spaghetti} THEN {Rotwein, Tomaten, Basilikum}
- IF {Waits, Dylan, Bush} THEN {Spektor}

Mehrwertig:

- IF {humidity = high, windy = false} THEN {outlook = sunny}

Zweiwertig:

- IF {Pommes, \neg Ketchup} THEN {Mayonnaise}
- IF {E.Jelinek, Ch.Roche} THEN { \neg U.Danella}

Beispiel — Tennisdaten mit WEKA

5 Attribute · 14 Objekte · Apriori mit $\theta_s = 15\%$, $\theta_c = 90\%$

Stücklistenaufstellung („itemsets“)

12 Einermengen · 47 Paare · 39 Tripel · 6 Quadrupel

Beste Assoziationsregeln ($\theta_c \equiv 100\%$)

- 4 IF {humidity = normal, windy = false} THEN {play = yes}
IF {temperature = cool} THEN {humidity = normal}
IF {outlook = overcast} THEN {play = yes}
- 3 IF {temperature = cool, play = yes} THEN {humidity = normal}
IF {outlook = rainy, windy = false} THEN {play = yes}
IF {outlook = rainy, play = yes} THEN {windy = false}
IF {outlook = sunny, humidity = high} THEN {play = no}
IF {outlook = sunny, play = no} THEN {humidity = high}
- 2 IF {temp = cool, windy = false} THEN {humidity = normal, play = yes}
IF {temp = cool, humidity = normal, windy = false} THEN {play = yes}

Korrelation, Regression und Transformation

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Graphische Wahrscheinlichkeitsmodelle

Regen \leftarrow **Jahreszeit**

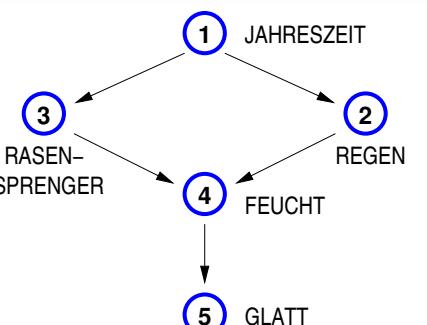
	X_1	X_1	X_1	X_1
	w	f	s	h
$X_2 = 0$	0.2	0.3	0.1	0.7
$X_2 = 1$	0.8	0.7	0.9	0.3

Feucht \leftarrow **Regen, Sprenger**

	$X_2 X_3$	$\bar{X}_2 X_3$	$X_2 \bar{X}_3$	$\bar{X}_2 \bar{X}_3$	
	\bar{X}_4	0.1	0.3	0.4	0.8
$X_4 = 0$	0.9	0.7	0.6	0.2	
$X_4 = 1$					

Wozu Graphische Modelle ?

- Visualisierung quantitativer Zusammenhänge
- Inferenz von Abhängigkeitsbeziehungen
- Berechnung kausaler Effekte
- Effiziente Auswertung multivariater Modelle



Jahreszeit \leftarrow

$X_1 = w$	0.25
$X_1 = f$	0.25
$X_1 = s$	0.25
$X_1 = h$	0.25

Glatt \leftarrow **Feucht**

	X_4	\bar{X}_4	
	\bar{X}_5	0.3	0.9
	$X_5 = 1$	0.7	0.1

Simpsons Paradoxon #1

Geschlechtsspezifische Diskriminierung an der Universität

Zweiwegetabelle: Geschlecht & Zulassungsquote

Geschlecht	#Bewerbung	#Zulassung	%
M	600	350	58.3
F	600	250	41.6

Frauen haben die geringeren Zulassungschancen!

„marginale Abhängigkeit“



„bedingte Abhängigkeit“

Zusatzvariable: Fakultätszugehörigkeit

Fakultät	Geschlecht	#Bewerbung	#Zulassung	%
TECH	M	100	25	25
TECH	F	300	75	25
PHIL	M	200	100	50
PHIL	F	200	100	50
THEO	M	300	225	75
THEO	F	100	75	75

Männer tendieren zu Fakultäten mit hoher Zulassungsquote!

Wahrscheinlichkeit und Graphstruktur

Datensatz

Wahrscheinlichkeits(dichte)werte

$$P : \Omega = \mathcal{X}_1 \times \dots \times \mathcal{X}_N \rightarrow \mathbb{IR}_0^+$$

durch **Statistiken** des Datensatzes
 $\omega \subset \Omega$ repräsentiert.

$$P(x) = P(x_1) \cdot P(x_2|x_1) \cdot P(x_3|x_1) \cdot P(x_4|x_2, x_3) \cdot P(x_5|x_4)$$

Dependenzmodell

Menge aller **bedingten**
Unabhängigkeiten zwischen
Mengen von Zufallsvariablen

$$\mathfrak{I}(\mathbb{X}_2 | \mathbb{X}_1 | \mathbb{X}_3)$$

(„Regen“ unabhängig von
„Rasensprenger“, wenn „Jahreszeit“
gegeben)

Graphisches Modell

- **Markovnetz**
ungerichteter Graph
„partielle Unabhängigkeit“
 $\mathbb{X}_i \not\leftrightarrow \mathbb{X}_j$
- **Bayesnetz**
gerichteter azyklischer Graph
„kausale Abhängigkeit“
 $\mathbb{X}_i \rightarrow \mathbb{X}_j$

Simpsons Paradoxon #2

Eine farbenfrohe Mordstatistik für den Bundesstaat Florida

Zweiwegetabelle: Hautfarbe & Strafmaß

FarbeMörder	#Todesurteil	#Haftstrafe	% T.U.
schwarz	17	149	11.4
weiß	19	141	12.5

kein Rassismus: ähnliche Todesurteilquote für Schwarz und Weiß

„marginal unabhängig“



Zusatzvariable: Hautfarbe des Opfers

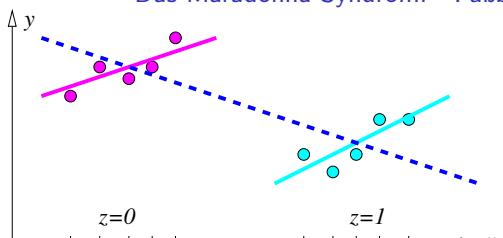
FarbeOpfer	FarbeMörder	#Tod	#Haft	% T.U.
schwarz	schwarz	6	97	5.8
schwarz	weiß	0	9	0.0
weiß	schwarz	11	52	17.5
weiß	weiß	19	132	12.6

Der Mord an einem weißen Mitbürger kommt teurer zu stehen!

„bedingt unabhängig“

Simpsons Paradoxon #3

Das Maradonna-Syndrom: Fußballspielen ist ungesund!



Drei Attribute

- \mathbb{X} = „Fußballaktivität“
- \mathbb{Y} = „Lebenserwartung“
- \mathbb{Z} = „Geschlecht“

Bedingte Abhängigkeit

Weibliche • wie männliche • Regressionsgeraden

$$\mathbb{Y} = f(\mathbb{X}|0) \quad \text{bzw.} \quad \mathbb{Y} = f(\mathbb{X}|1)$$

besitzen **positive** Steigung.

Grund:

Frauen sind tendenziell **langlebig** und stehen eher auf **Volleyball+Ayurveda**.

Marginale Abhängigkeit

Geschlechtsneutrale
Regressionsgerade

$$\mathbb{Y} = f(\mathbb{X})$$

besitzt **negative** Steigung.

Rechenregeln für bedingte Unabhängigkeiten

Lemma

Die folgenden Aussagen über die Werte a , b und z dreier Zufallsvariablen $\mathbb{X}_a, \mathbb{X}_b, \mathbb{X}_z$ sind äquivalent:

1. $P(a | b, z) = P(a | z)$ ($a \not\sim b$ wenn z bekannt)
2. $P(b | a, z) = P(b | z)$ ($b \not\sim a$ wenn z bekannt)
3. $P(a, b, z) = f(a, z) \cdot g(b, z)$ (Faktorisierbarkeit)

Diese Äquivalenz gilt entsprechend für **Mengen** von Zufallsvariablen.

Weitere äquivalente Formulierungen

für die bedingte statistische Unabhängigkeit zwischen drei Zufallsvariablen:

1. $P(a, b, z) = \frac{P(a, z) \cdot P(b, z)}{P(z)}$
2. $P(a, b | z) = P(a | z) \cdot P(b | z)$
3. $P(a, b, z) = P(a | z) \cdot P(b, z)$

Diskrete ZV

$P(\mathbb{Y} = y | \mathbb{I} = i)$ ist konstant bzgl. i .

Stetige ZV

Lineare Regression
 $\mathbb{Y} | x \sim \mathcal{N}(a + bx, \sigma^2)$
mit $b = 0$.

Bedingte statistische Unabhängigkeit

von Mengen von Zufallsvariablen

Definition

Es seien A, B, Z drei paarweise disjunkte Teilmengen der Zufallsvariablen $\{\mathbb{X}_1, \dots, \mathbb{X}_N\}$. Dann heißt A **bedingt statistisch unabhängig** von B bezogen auf Z genau dann, wenn gilt

$$P(A | B, Z) = P(A | Z)$$

und wir schreiben

$$\mathfrak{S}(A | Z | B).$$

Ferner heißen A und B **bedingt faktorisierbar** bezogen auf Z , falls es zwei geeignete Funktionen (sic!) f und g gibt mit

$$P(A, B, Z) = f(A, Z) \cdot g(B, Z).$$

Marginale statistische Unabhängigkeit

Der Spezialfall „gewöhnlicher“ statistischer Unabhängigkeit ergibt sich für UA-Postulate der Form $\mathfrak{S}(A | Z | B)$ mit $Z = \emptyset$.

Beweis.

- (1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (1)

$$\begin{aligned} P(b | a, z) &= \frac{P(a, b | z)}{P(a | z)} &=& \frac{P(a | b, z) \cdot P(b | z)}{P(a | z)} \\ & &=& \frac{P(a | z) \cdot P(b | z)}{P(a | z)} = P(b | z) \end{aligned}$$

- (1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (1)

$$P(a, b, z) = P(a, z) \cdot P(b | a, z) = P(a, z) \cdot P(b | z) =: f(a, z) \cdot g(b, z)$$

- (1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (1)

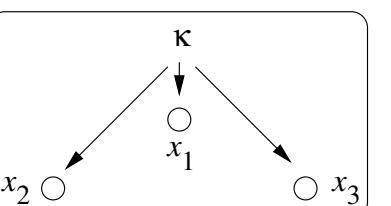
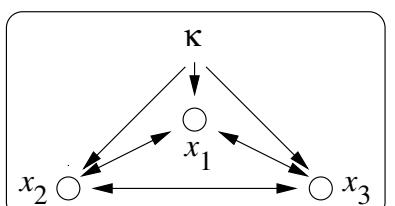
$$\begin{aligned} P(a | b, z) &= \frac{P(a, b, z)}{P(b, z)} = \frac{P(a, b, z)}{\sum_a P(a, b, z)} \\ &= \frac{f(a, z) \cdot g(b, z)}{\sum_a f(a, z) \cdot g(b, z)} = \frac{f(a, z)}{\sum_a f(a, z)} \end{aligned}$$

Der letzte Ausdruck ist offenbar unabhängig von b .

□

Beispiel — Numerische Klassifikation

Normale und naive Bayesregel



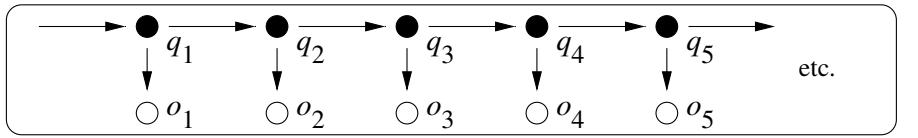
Datenerzeugungsmodell

für Merkmale $x_1, \dots, x_N \in \mathbb{R}$ und Klassenvariable $y \in \{\Omega_1, \dots, \Omega_K\}$:

$$f(\mathbf{x}, \Omega_\kappa) = P(\Omega_\kappa) \cdot f(\mathbf{x} | \Omega_\kappa)$$

- **Multivariate Normalverteilungsdichte** (saturiertes Modell):
 $x_i \leftarrow \{x_j \mid j \neq i\}$ für alle i
- **Klassenbedingte Unabhängigkeit** (ausgedünntes Modell):
 $f(\mathbf{x} | \Omega_\kappa) = \prod_i \mathcal{N}(x_i \mid \mu_i, \sigma_i^2)$ ergibt $x_i \leftarrow \emptyset$ für alle i

Beispiel — Hidden Markov Modelle



Datenerzeugungsmodell

Beobachtbare Zeichenfolge $\mathbf{o} = o_1 \dots o_T$ mit $o_t \in \mathcal{O}$

Verborgene Zustandsfolge $\mathbf{q} = q_1 \dots q_T$ mit $q_t \in \mathcal{Q}$

$$P(\mathbf{o}) = P(\mathbf{o} | \lambda) = \sum_{\mathbf{q} \in \mathcal{Q}^T} P(\mathbf{o}, \mathbf{q} | \lambda) = \sum_{\mathbf{q} \in \mathcal{Q}^T} \prod_{t=1}^T P(q_t | q_{t-1}) \cdot P(o_t | q_t)$$

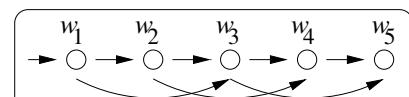
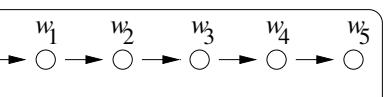
mit statistischen Abhängigkeiten $q_t \leftarrow q_{t-1}$ und $o_t \leftarrow q_t$.

Unabhängigkeitspostulate des HMM

$\mathfrak{I}(\{q_{t+1}\} \mid \{q_t\} \mid \{q_1, \dots, q_{t-1}; o_1, \dots, o_t\})$ und

$\mathfrak{I}(\{o_{t+1}\} \mid \{q_{t+1}\} \mid \{q_1, \dots, q_t; o_1, \dots, o_t\})$

Beispiel — N-Gramm-Grammatiken



Datenerzeugungsmodell

für eine Symbolfolge (Wortfolge) $\mathbf{w} = w_1 \dots w_M$ ist die **Kettenregel**

$$P(\mathbf{w}) = \prod_{m=1}^M P(w_m \mid w_1, \dots, w_{m-1}) \simeq \prod_{m=1}^M P(w_m \mid w_{m-2}, w_{m-1}) \simeq \prod_{m=1}^M P(w_m \mid w_{m-1})$$

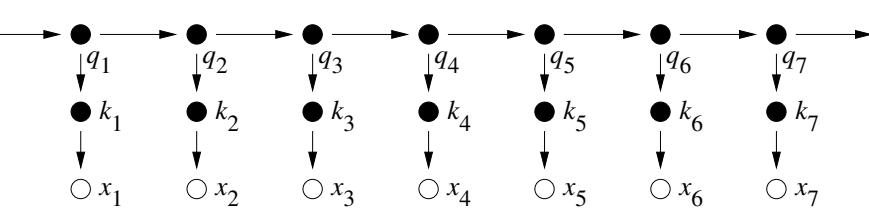
mit den statistischen Abhängigkeiten $\begin{cases} w_m \leftarrow w_{m-1} & (\text{Bigramme}) \\ w_m \leftarrow \{w_{m-2}, w_{m-1}\} & (\text{Trigramme}) \end{cases}$.

Unabhängigkeitspostulate der Bigramm-Grammatik

$\mathfrak{I}(\{w_m\} \mid \{w_{m-1}\} \mid \{w_1, \dots, w_{m-2}\})$ für alle $m = 2, \dots, M$.

Beispiel — (Semi-)kontinuierliches HMM

mit eindimensionalen Ausgabewerten



Datenerzeugungsmodell

Beobachtbare Wertefolge $\mathbf{x} = x_1, \dots, x_T$ mit $x_t \in \mathbb{R}$

Verborgene Komponentenfolge $\mathbf{k} = k_1 \dots k_T$ mit $k_t \in \mathcal{K}$

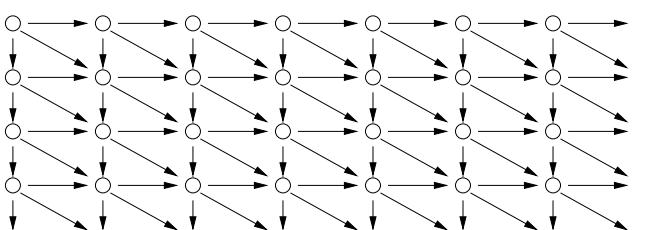
Verborgene Zustandsfolge $\mathbf{q} = q_1 \dots q_T$ mit $q_t \in \mathcal{Q}$

$$P(\mathbf{X}) = P(\mathbf{X} | \lambda) = \sum_{\mathbf{q} \in \mathcal{Q}^T} \sum_{\mathbf{k} \in \mathcal{K}^T} P(\mathbf{X}, \mathbf{k}, \mathbf{q} | \lambda)$$

mit statistischen Abhängigkeiten $q_t \leftarrow q_{t-1}$, $k_t \leftarrow q_t$ und $x_t \leftarrow k_t$.

Beispiel — 2D Markov Random Field

Texturmodelle in der Grauwertbildanalyse



Datenerzeugungsmodell

Beobachtbare Zufallsvariablen $x_{i,j}$

auf dem **Ortgitter** $i = 1, \dots, I$ und $j = 1, \dots, J$

mit statistischen Abhängigkeiten $x_{i,j} \leftarrow \{x_{i-1,j-1}, x_{i,j-1}, x_{i-1,j}\}$.

Unabhängigkeitspostulate des MRF

Für alle Gitterpunkte $(n, m) \in \mathbb{Z} \times \mathbb{Z}$ ist gefordert:

$$\mathfrak{S}(\{\mathbb{X}_{n,m}\} | \{\mathbb{X}_{n,m-1}, \mathbb{X}_{n-1,m}, \mathbb{X}_{n-1,m-1}\} | \{\mathbb{X}_{i,j} \mid i < n, j < m\})$$

Pearlsche Dependenzaxiome

Axiomatische Charakterisierung aller „erlaubten“ \mathfrak{S} -Relationen

Satz (Judea Pearl)

Es sei $P(\cdot)$ eine Wahrscheinlichkeitsverteilung über $\mathbb{X}_1, \dots, \mathbb{X}_N$ und $\mathfrak{S}(\cdot | \cdot | \cdot)$ das zugehörige Dependenzmodell.

Dann gelten für alle (paarweise disjunkten) Variablenmengen A, B, C, Z die folgenden vier Aussagen:

SYM Symmetrie

$$\mathfrak{S}(A | Z | B) \Leftrightarrow \mathfrak{S}(B | Z | A)$$

DEC Dekomposition

$$\mathfrak{S}(A | Z | B \cup C) \Rightarrow \mathfrak{S}(A | Z | B) \wedge \mathfrak{S}(A | Z | C)$$

WUN Schwache Vereinigung

$$\mathfrak{S}(A | Z | B \cup C) \Rightarrow \mathfrak{S}(A | Z \cup C | B)$$

CON Kontraktion

$$\mathfrak{S}(A | Z | B) \wedge \mathfrak{S}(A | Z \cup B | C) \Rightarrow \mathfrak{S}(A | Z | B \cup C)$$

Falls $P(\cdot)$ zudem streng positiv ($\forall x \in \Omega : P(x) > 0$) ist, gilt sogar:

INT Durchschnitt

$$\mathfrak{S}(A | Z \cup C | B) \wedge \mathfrak{S}(A | Z \cup B | C) \Rightarrow \mathfrak{S}(A | Z | B \cup C)$$

Dependenzmodelle

Algebraische Charakterisierung von Abhängigkeitsstrukturen

Definition

Sei $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$ eine Menge von Zufallsvariablen und $P(\cdot)$ eine Verteilung über V . Die Relation $\mathfrak{S} = \mathfrak{S}_P$ mit

$$\mathfrak{S} : \mathfrak{P}X \times \mathfrak{P}X \times \mathfrak{P}X \rightarrow \{0, 1\}$$

heißt **Dependenzmodell von** $P(\cdot)$, wenn für alle (disjunkten) Variablenmengen $A, B, Z \subset V$ gilt:

$$\mathfrak{S}(A | Z | B) \Leftrightarrow P(A | B, Z) = P(A | Z)$$

Bemerkungen

1. Es gibt 4^N viele Variablenkombinationen A, B, Z .
Es gibt 2^{4^N} viele dreistellige Mengenrelationen \mathfrak{S} über V .
Wieviele \mathfrak{S} davon sind ein **valides Dependenzmodell** \mathfrak{S}_P ?
2. Simpsons Paradoxa: $\mathfrak{S}(A | Z | B) \neq \mathfrak{S}(A | \emptyset | B)$ und $\mathfrak{S}(A | Z | B) \neq \mathfrak{S}(A | \emptyset | B)$

Beweis.

SYM Symmetrie

$$\mathfrak{S}(A | Z | B) \Rightarrow P(A, Z, B) = \underbrace{f(A, Z) \cdot g(B, Z)}_{P(B, Z, A)} \Rightarrow \mathfrak{S}(B | Z | A)$$

DEC Dekomposition

$$P(A, Z, B) = \sum_C P(A, Z, B, C) = \sum_C f(A, Z) \cdot g(B, C, Z) = f(A, Z) \cdot \underbrace{\sum_C g(B, C, Z)}_{\tilde{g}(B, Z)}$$

beweist $\mathfrak{S}(A | Z | B)$; analoge Herleitung von $\mathfrak{S}(A | Z | C)$.

WUN Schwache Vereinigung

$$\begin{aligned} \mathfrak{S}(A | Z | B, C) \Rightarrow P(A, Z, B, C) &= f(A, Z) \cdot g(B, C, Z) \\ &= \tilde{f}(A, Z, C) \cdot \tilde{g}(B, Z, C) \Rightarrow \mathfrak{S}(A | Z, C | B) \end{aligned}$$

CON Kontraktion

$$P(A | Z, B, C) = \underbrace{P(A | Z, B)}_{\mathfrak{S}(A | Z, B | C)} = \underbrace{P(A | Z)}_{\mathfrak{S}(A | Z | B)} \Rightarrow \mathfrak{S}(A | Z | B, C)$$

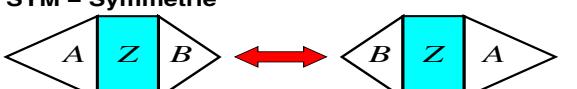
INT Durchschnitt (Beweis zu äquivalenter Formulierung INT* folgt)



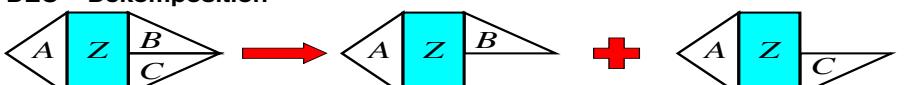
Pearlsche Dependenzaxiome

Beweis durch angestringtes Hingucken

SYM – Symmetrie



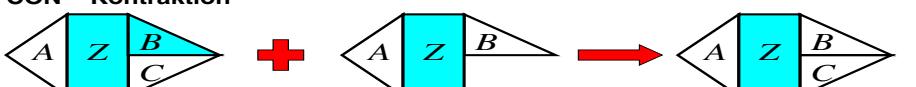
DEC – Dekomposition



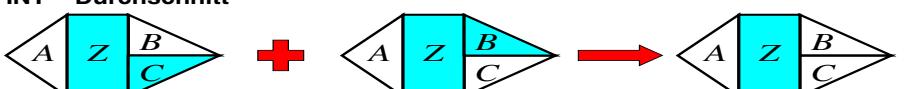
WUN – Schwache Vereinigung



CON – Kontraktion



INT – Durchschnitt



Bemerkungen zu Pearls Axiomen

1. Die logische Umkehrung der Implikation CON folgt aus den Axiomen DEC und WUN.
2. Die logische Umkehrung von INT folgt mit zweimaliger Anwendung von WUN.
3. Die Axiomatisierung kann auf nichtdisjunkte ZV-Mengen ausgedehnt werden. Aus den obengenannten Axiomen sowie der zusätzlichen Forderung $\mathfrak{S}(A | Z | Z)$ beweist man die Aussage

$$\mathfrak{S}(A | Z | B) \Leftrightarrow \mathfrak{S}(A, Z | Z | B, Z)$$
4. Die fünf Axiome sind voneinander logisch unabhängig. Beweis durch Gegenbeispiele.
5. Das Axiom INT findet sich auch in der folgenden, äquivalenten Fassung INT* (Lauritzen, $Z = \emptyset$) wieder:

$$\mathfrak{S}(A | C | B) \wedge \mathfrak{S}(A | B | C) \Rightarrow \mathfrak{S}(A | \emptyset | B, C)$$

Durchschnittsaxiom INT

Garantiert ausschließlich für streng positive Verteilungen!

Herleitung für streng positive $P(\cdot)$

Auf Grund der Prämissen von INT* gelten die Faktorisierungen

$$P(a, b, c) = k(a, c) \cdot \ell(b, c) = g(a, b) \cdot h(b, c)$$

und für beliebige Werte c — also zum Beispiel für c_0 beliebig aber fest — gilt

$$g(a, b) = k(a, c_0) \cdot \frac{\ell(b, c_0)}{h(b, c_0)} =: \pi(a) \cdot \rho(b).$$

Dann gilt die marginale Unabhängigkeit $\{a\} \not\sim \{b, c\}$ wegen der Faktorisierung

$$P(a, b, c) = \pi(a) \cdot [\rho(b) \cdot h(b, c)].$$

Gegenbeispiel

Die drei binärwertige Zufallsvariablen mit $\mathbb{X}_1 = \mathbb{X}_2 = \mathbb{X}_3$ und $P(\mathbb{X}_i = 1) = \frac{1}{2}$ für alle $i \in \{1, 2, 3\}$ sind nicht streng positiv (z.B. $P(1, 1, 0) = 0$) und widerlegen INT:

$$\mathfrak{S}(\mathbb{X}_1 | \mathbb{X}_2 | \mathbb{X}_3), \quad \mathfrak{S}(\mathbb{X}_1 | \mathbb{X}_3 | \mathbb{X}_2), \quad \neg \mathfrak{S}(\mathbb{X}_1 | \emptyset | \mathbb{X}_2, \mathbb{X}_3)$$

Vollständigkeitsvermutung (Pearl & Paz, 1985)

Trügerische Hoffnung

Wenn \mathfrak{S} die Axiome SYM, DEC, WUN & CON erfüllt, so heißt (V, \mathfrak{S}) Semigraphoid und es gibt eine Wahrscheinlichkeitsverteilung $P(\cdot)$ mit

$$P(A | B, Z) = P(A | Z) \Leftrightarrow \mathfrak{S}(A | Z | B).$$

Wenn zusätzlich das Durchschnittsaxiom (INT) erfüllt ist, so kann für das Graphoid (V, \mathfrak{S}) sogar ein streng positives $P(\cdot)$ gefunden werden.

Satz (Studeny, 1992)

Weder für die Relationenmenge

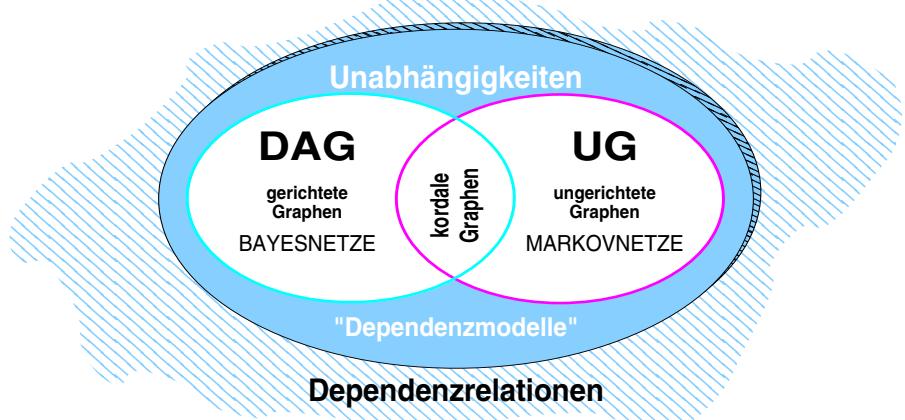
$$\{\mathfrak{S}_P \mid P \text{ Wahrscheinlichkeitsverteilung}\}$$

noch für deren Teilmenge

$$\{\mathfrak{S}_P \mid P \text{ streng positive Wahrscheinlichkeitsverteilung}\}$$

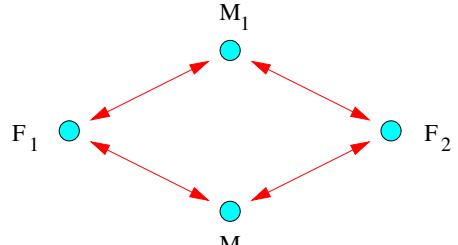
gibt es ein korrektes und vollständiges **endliches Axiomensystem**.

Dependenzmodelle und Graphen



- ? Welche Dependenzmodelle sind durch UG charakterisierbar
- ? Welche Dependenzmodelle sind durch DAG charakterisierbar
- ? Welche Dependenzmodelle liegen gleichzeitig in beiden Klassen
- ? Welche Dependenzmodelle sind komplexer als jede Graphstruktur

Trennungsrelation im ungerichteten Graphen



Attributwerte: \pm infiziert

Definition

Es sei $\mathcal{G} = (V, \mathcal{E})$ ein ungerichteter Graph und $A, B, Z \subset V$ disjunkte Knotenmengen. Die Menge Z trennt A von B genau dann, wenn alle Pfade zwischen Elementen $a \in A$ und $b \in B$ mindestens einen Knoten $z \in Z$ enthalten. Wir schreiben dafür:

$$\text{sep}(A | Z | B)$$

⇒ „Z blockiert alle Verbindungen zwischen Knoten aus A und B“

Korrelation, Regression und Transformation

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Graphische Verteilungen und Dependenzmodelle

Überrepräsentation & Unterrepräsentation von $\mathfrak{S}(\cdot | \cdot | \cdot)$ durch $\text{sep}(\cdot | \cdot | \cdot)$

Definition

Es sei $P(\cdot)$ eine Wahrscheinlichkeitsverteilung auf V und \mathfrak{S} ihr Dependenzmodell. Der ungerichtete Graph $\mathcal{G} = (V, \mathcal{E})$ heißt

- **Abhängigkeitsbild** von P gdw.

$$\mathfrak{S}(A | Z | B) \Rightarrow \text{sep}(A | Z | B)$$

- **Unabhängigkeitsbild** von P gdw.

$$\mathfrak{S}(A | Z | B) \Leftarrow \text{sep}(A | Z | B)$$

- **perfektes Bild** von P gdw.

$$\mathfrak{S}(A | Z | B) \Leftrightarrow \text{sep}(A | Z | B)$$

Die Verteilung $P(\cdot)$ (und das Modell \mathfrak{S}) heißen **graphisch**, wenn ein ungerichteter Graph existiert, der \mathfrak{S} perfekt abbildet.

Über A-Bilder, U-Bilder und P-Bilder

Bemerkungen

- Die Trennungsrelation im UG ist **monoton** in der Barriere Z :

$$\text{sep}\langle A|Z|B \rangle \text{ und } \tilde{Z} \supset Z \Rightarrow \text{sep}\langle A|\tilde{Z}|B \rangle$$

- Es gilt die „*marginale Trennung*“ $\text{sep}\langle \{a\}|\emptyset|\{b\} \rangle$ genau dann, wenn $a, b \in V$ zu verschiedenen Zusammenhangskomponenten gehören.

- A-Bild \rightsquigarrow für adjazente Knoten gilt keinerlei Unabhängigkeit
(der diskrete Graph ist A-Bild jedes P)

- U-Bild \rightsquigarrow für nichtadjazente Knoten gilt ≥ 1 Unabhängigkeit
(der vollständige Graph ist U-Bild jedes P)

- Nicht alle Dependenzmodelle \mathfrak{S} besitzen ein perfektes Bild.
Für das nichtmonotone Modell mit zwei Würfeln $\mathbb{W}_1, \mathbb{W}_2$ und die Signalglocke \mathbb{G} für Pasch gilt nämlich

$$\mathfrak{S}(\mathbb{W}_1 | \emptyset | \mathbb{W}_2) \text{ und nicht } \mathfrak{S}(\mathbb{W}_1 | \mathbb{G} | \mathbb{W}_2).$$

Die Markoveigenschaften für „Semigraphoide“

Markovnetze $\hat{=}$ minimale Unabhängigkeitsbilder

Definition

Der Graph \mathcal{G} heißt **Markovnetz** von $P(\cdot)$, wenn er minimal mit der globalen Markoveigenschaft für $P(\cdot)$ ist.

Das Markovnetz \mathcal{G} ignoriert keine Abhängigkeiten, höchstens Unabhängigkeiten, aber davon so wenige wie möglich.

Satz

Sei $\mathcal{G} = (V, \mathcal{E})$ und $P(\cdot)$ auf V gegeben. Dann gilt

$$\text{globale ME} \Rightarrow \text{lokale ME} \Rightarrow \text{paarweise ME},$$

aber es gilt im allgemeinen weder die Umkehrrichtung

$$\text{paarweise ME} \Rightarrow \text{lokale ME}$$

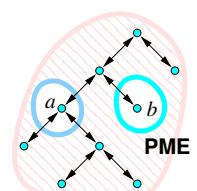
noch die Umkehrrichtung

$$\text{lokale ME} \Rightarrow \text{globale ME}$$

Die drei Markoveigenschaften

Definition

Es sei $P(\cdot)$ eine Wahrscheinlichkeitsverteilung auf V und \mathfrak{S} ihr Dependenzmodell. Der ungerichtete Graph $\mathcal{G} = (V, \mathcal{E})$ erfüllt die



- paarweise Markoveigenschaft**

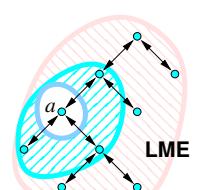
gdw. für alle nichtadjazenten $a, b \in V$ gilt:

$$\mathfrak{S}(a | V \setminus \{a, b\} | b)$$

- lokale Markoveigenschaft**

gdw. für alle jede Variable $a \in V$ gilt:

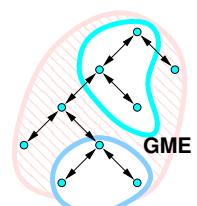
$$\mathfrak{S}(a | \text{bd}(a) | V \setminus \text{cl}(a))$$



- globale Markoveigenschaft**

gdw. für alle $A, B, Z \subset V$ mit $\text{sep}\langle A|Z|B \rangle$ gilt:

$$\mathfrak{S}(A | Z | B)$$



Beweis.

- GME \Rightarrow LME**

Es sei $a \in V$.

Offensichtlich werden die beiden Mengen $\{a\}$ und $V \setminus \text{cl}(a)$ durch den Rand $\text{bd}(a)$ von a separiert.

Damit folgt die Behauptung aus der Anwendung von GME.

- LME \Rightarrow PME**

Zunächst gilt wegen der Voraussetzung LME die Aussage

$$\mathfrak{S}(a | \text{bd}(a) | V \setminus \text{cl}(a))$$

Wegen der Teilmengenbeziehung

$$V \setminus \{a, b\} = \text{bd}(a) \cup ((V \setminus \text{cl}(a)) \setminus \{b\})$$

kann mittels Axiom WUN

$$\mathfrak{S}(a | V \setminus \{a, b\} | V \setminus \text{cl}(a))$$

gefolgert werden und mittels Axiom DEC wird verkürzt zu

$$\mathfrak{S}(a | V \setminus \{a, b\} | b).$$

- LME \Rightarrow GME (Gegenbeispiel)

$$\mathbb{U} - \mathbb{W} - \mathbb{X} - \mathbb{Y} - \mathbb{Z} \quad (\mathbb{U} = \mathbb{W}, \mathbb{Y} = \mathbb{Z}, \mathbb{X} = \mathbb{W} \cdot \mathbb{Y})$$

mit binärwertigen, gleichverteilten Variablen.

Es gilt zwar die lokale ME, aber $\mathfrak{S}(\mathbb{U}, \mathbb{W} | \mathbb{X} | \mathbb{Y}, \mathbb{Z})$ scheitert wegen

$$P(\mathbb{U} = \mathbb{W} = \mathbb{Y} = \mathbb{Z} = 1 | \mathbb{X} = 0) = 0$$

$$P(\mathbb{U} = \mathbb{W} = 1 | \mathbb{X} = 0) \cdot P(\mathbb{Y} = \mathbb{Z} = 1 | \mathbb{X} = 0) \neq 0$$

- PME \Rightarrow LME (Gegenbeispiel)

$$\mathbb{X} - \mathbb{Y} - \mathbb{Z} \quad (\mathbb{X} = \mathbb{Y} = \mathbb{Z})$$

mit binärwertigen, gleichverteilten Variablen.

Dann sind $\mathfrak{S}(X | Z | Y)$ und $\mathfrak{S}(X | Y | Z)$ trivialerweise erfüllt, überflüssigerweise sogar auch $\mathfrak{S}(Y | X | Z)$. Aber es gilt keineswegs

$$\mathfrak{S}(X | \text{bd}(X) | V \setminus \text{cl}(X))$$

denn $\text{bd}(X) = \emptyset$ und $V \setminus \text{cl}(X) = \{Y, Z\}$, und es ist \mathbb{X} natürlich nicht marginal unabhängig von $\{\mathbb{Y}, \mathbb{Z}\}$.

□

Beweis.

Es ist nur die Implikation PME \Rightarrow GME zu zeigen, die wir durch absteigende Induktion über die Größe $n = |Z|$ beweisen.

- **Induktionsanfang:**

Für $n = N - 2$ liefert PME die Behauptung (o.B.d.A. sei $|A| = |B| = 1$).

Induktionsschluß:

Wir unterscheiden die beiden Fälle $A \cup B \cup Z = V$ und $A \cup B \cup Z \neq V$.

- **Fall 1:** Sei o.B.d.A. $|A| > 1$ und $a \in A$. Dann gelten nach WUN die beiden Trennungsaussagen

$$\text{sep}(A \setminus \{a\} | Z \cup \{a\} | B), \quad \text{sep}(\{a\} | Z \cup A \setminus \{a\} | B).$$

Nach I.V. übersetzen diese in die korrespondierenden Unabhängigkeiten und mit Axiom INT folgt $\mathfrak{S}(A | Z | B)$.

- **Fall 2:** Für jedes $a \in V \setminus (A \cup B \cup Z)$ gilt $\text{sep}(A | Z \cup \{a\} | B)$ und mindestens eine der beiden Trennungsaussagen

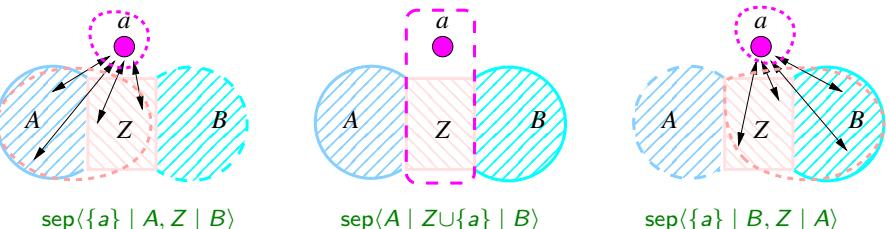
$$\text{sep}(\{a\} | A, Z | B), \quad \text{sep}(\{a\} | B, Z | A).$$

Im ersten Fall folgt das Resultat $\mathfrak{S}(A | Z | B)$ nach den Axiomen INT, DEC und im zweiten Fall nach den Axiomen SYM, INT, DEC aus den übersetzten Trennungsaussagen (I.V.).

□

Die Markoveigenschaften für „Graphoide“

Äquivalenz für strikt positive Wahrscheinlichkeitsverteilungen



Satz

Sei \mathcal{G} ein UG. Erfüllt die Dependenzrelation \mathfrak{S} von $P(\cdot)$ für alle disjunkten Mengen $A, B, C, Z \subset V$ die Eigenschaft

- **INT Durchschnitt**

$$\mathfrak{S}(A | Z \cup C | B) \wedge \mathfrak{S}(A | Z \cup B | C) \Rightarrow \mathfrak{S}(A | Z | B \cup C),$$

so gilt

$$\text{globale ME} \Leftrightarrow \text{lokale ME} \Leftrightarrow \text{paarweise ME}.$$

Markovnetzkonstruktion

(1:1)-Abbildung aller partiellen (Un-)Abhängigkeiten

Lemma

Erfüllt die Dependenzrelation \mathfrak{S} von $P(\cdot)$ die Axiome SYM, DEC und INT, so gibt es ein eindeutiges Markovnetz $\mathcal{G} = (V, \mathcal{E})$ zu \mathfrak{S} .

Für alle Variablenpaare $a, b \in V$ gilt:

$$\{a, b\} \notin \mathcal{E} \Leftrightarrow \mathfrak{S}(a | V \setminus \{a, b\} | b)$$

Satz (Pearl & Paz, 1985)

Die Dependenzrelation \mathfrak{S} ist graphisch genau dann, wenn sie die Axiome SYM, DEC, INT, SUN und TRA erfüllt.

- **SUN Starke Vereinigung**

$$\mathfrak{S}(A | Z | B) \Rightarrow \mathfrak{S}(A | Z \cup C | B)$$

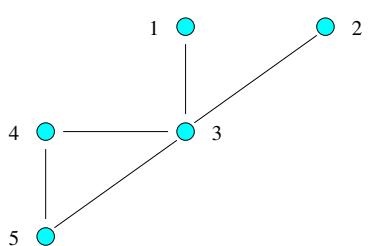
- **TRA Transitivität**

Für alle Variablen $x \in V$ gilt:

$$\mathfrak{S}(A | Z | B) \Rightarrow \mathfrak{S}(A | Z | \{x\}) \vee \mathfrak{S}(\{x\} | Z | B)$$

Beispiel — qualitative graphische Inferenz

„Vorhersage einer Reiseankunftszeit“



Uhrzeitwertige Zufallsvariable
Zwei Passanten — zwei Armbanduhren

- X_1 Zeit auf Armbanduhr I
- X_2 Zeit auf Armbanduhr II
- X_3 die wahre Uhrzeit
- X_4 die Fahrtzeit „Jena–Weimar“
- X_5 die Ankunftszeit in Weimar

Markovnetzerzeugung

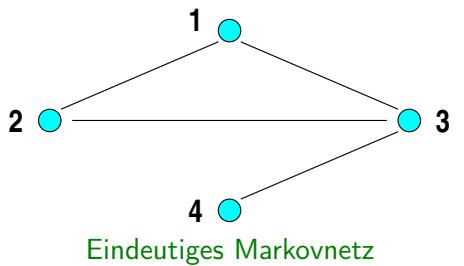
Kantenlöscherfahren mit den Vorbehalten:

$$\left\{ \begin{array}{l} \neg \Im(X_1 | X_2, X_4, X_5 | X_3) \\ \neg \Im(X_2 | X_1, X_4, X_5 | X_3) \end{array} \right. , \quad \left\{ \begin{array}{l} \neg \Im(X_3 | X_1, X_2, X_5 | X_4) \\ \neg \Im(X_3 | X_1, X_2, X_4 | X_5) \\ \neg \Im(X_4 | X_1, X_2, X_3 | X_5) \end{array} \right.$$

Inferenz durch Ablesen von Trennungseigenschaften

Bedingte, aber nicht partielle Unabhängigkeiten: $\Im(X_1, X_2 | X_3 | X_5)$

Beispiel — eine Unverteilung mit Markovnetz



Eigenschaften

- \Im erfüllt die Axiome SYM, DEC, WUN und INT.
- \Im widerspricht dem Axiom CON, denn es gelten zwar $\Im(1 | 2 | 3)$ und $\Im(1 | 2, 3 | 4)$, aber keineswegs $\Im(1 | 2 | 3, 4)$.
- \Im besitzt wegen \neg CON kein Wahrscheinlichkeitsmodell mit $\Im = \Im_P$.
- \Im besitzt aber wegen SYM, DEC, INT ein eindeutiges Markovnetz.

Dependenzstruktur

Gegeben sind die bedingten „Unabhängigkeiten“

$$\begin{aligned} \Im(1 | 2 | 3) & \quad \Im(1, 2 | 3 | 4) \\ \Im(1 | 3 | 4) & \quad \Im(1 | 2, 3 | 4) \\ \Im(2 | 3 | 4) & \quad \Im(2 | 1, 3 | 4) \end{aligned}$$

zuzüglich aller Symmetrien.

Beispiel — Würfelpaar und Glocke

Nichtgraphische Verteilungen

Viele interessante Verteilungen liegen außerhalb der Klasse ungerichteter graphischer Modelle.

- Selbst ein streng positives $P(\cdot)$ garantiert lediglich die Axiome SYM, DEC und INT, nicht aber SUN oder TRA.

Würfel-Glocken-Experiment

Es schlägt die starke Vereinigung (SUN) fehl:

$$\Im(W_1 | \emptyset | W_2) \text{ aber } \neg \Im(W_1 | G | W_2)$$

Bei **unfairen Würfeln** gilt auch keine Transitivität (TRA) mehr:

$$\Im(W_1 | \emptyset | W_2) \text{ aber weder } \Im(W_1 | \emptyset | G) \text{ noch } \Im(G | \emptyset | W_2)$$

Bemerkung

Die drei Axiome DEC, INT, SUN liefern eine beachtliche Äquivalenz:

$$\Im(A | Z | B) \iff \forall a \in A, b \in B : \Im(\{a\} | Z | \{b\})$$

Beispiel — pathologische Verteilung ohne Markovnetz

Dependenzstruktur

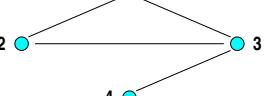
Gegeben sind bedingte Unabhängigkeiten

$$\Im(1 | 2, 3 | 4) \quad \Im(2 | 1, 3 | 4)$$

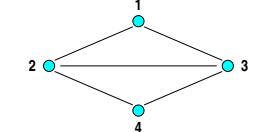
zuzüglich aller Symmetrien.

Eigenschaften

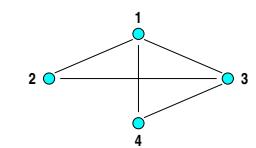
- \Im erfüllt die Axiome SYM, DEC, WUN, CON.
- \Im widerspricht dem Axiom INT, weil $\Im(1, 2 | 3 | 4)$ fehlt.
- \Im gehorcht einer Verteilung P , aber P ist wegen \neg INT nicht streng positiv!
- Das Kantenlöscherfahren ergibt kein Unabhängigkeitsbild, weil $\text{sep}(1|3|4)$ gilt, aber nicht $\Im(1|3|4)$.
- Es gibt kein eindeutiges Markovnetz!



Kantenlöscherfahren

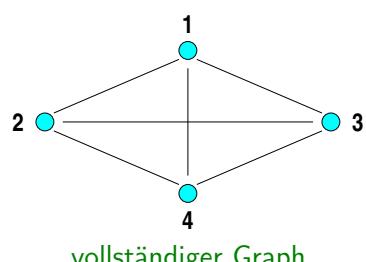


minimales U-Bild #1



minimales U-Bild #2

Beispiel — Unverteilung mit Monsternetz



Dependenzstruktur

Gegeben sind die Postulate

$$\mathfrak{S}(1 \mid 3 \mid 4) \quad \mathfrak{S}(2 \mid 3 \mid 4)$$

$$\mathfrak{S}(1, 2 \mid 3 \mid 4) \quad \mathfrak{S}(1 \mid 2 \mid 4)$$

zuzüglich aller Symmetrien.

Eigenschaften

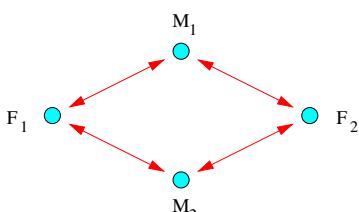
- \mathfrak{S} erfüllt die Axiome SYM, DEC, INT.
- \mathfrak{S} erfüllt nicht das Axiom WUN.
- Wegen SYM, DEC, INT gibt es ein eindeutiges Markovnetz \mathcal{G} .
- Der Graph \mathcal{G} ist offenbar (keine Löschung) **vollständig**.
- Der Graph \mathcal{G} „hilft uns nicht sparen“ ...

Faktorisierung von $P(\cdot)$

über den Cliquen eines ungerichteten Graphen

Definition

Die Menge $C \subseteq V$ heißt **Clique** von $\mathcal{G} = (V, \mathcal{E})$, wenn $(C, \mathcal{E}|_C)$ einen maximalen vollständigen Teilgraphen von \mathcal{G} bildet.



Der diamantene Graph
besitzt genau vier Cliquen:

$$C_1 = \{\mathbb{M}_1, \mathbb{F}_1\}, \quad C_2 = \{\mathbb{M}_1, \mathbb{F}_2\},$$

$$C_3 = \{\mathbb{M}_2, \mathbb{F}_1\}, \quad C_4 = \{\mathbb{M}_2, \mathbb{F}_2\}$$

Gibbs-Verteilung des PT-Modells über dem Diamanten

$$P(m_1, m_2, f_1, f_2) = \frac{1}{Z} \cdot \phi_1(m_1, f_1) \cdot \phi_2(m_1, f_2) \cdot \phi_3(m_2, f_1) \cdot \phi_4(m_2, f_2)$$

mit den **Kompatibilitäts-** oder **Kernfunktionen** (*keine Wahrscheinlichkeiten!*)

$$\phi_i(\xi_{i_1}, \xi_{i_2}) = \begin{cases} \alpha_i & \xi_{i_1} = \xi_{i_2} \text{ gleicher Gesundheitszustand} \\ \beta_i & \xi_{i_1} \neq \xi_{i_2} \text{ genau ein Partner infiziert} \end{cases}$$

FAK — die Faktorisierungseigenschaft

Definition

Die Wahrscheinlichkeitsverteilung $P(\cdot)$ **zerfällt über dem Graphen** $\mathcal{G} = (V, \mathcal{E})$, wenn es für jede vollständige Menge $A \subset V$ eine nichtnegative **Kernfunktion**

$$\phi_A : \bigotimes_{a \in A} \mathcal{X}_a \rightarrow \mathbb{R}_0^+$$

über dem kartesischen Produkt aller A -Wertebereiche gibt mit

$$P(x) = \prod_{A \text{ vollständig}} \phi_A(x_A)$$

O.B.d.A. können wir diese Faktorisierungseigenschaft (FAK) aber auch unter Beschränkung auf die Menge $\mathcal{C}(\mathcal{G})$ der **Cliquen** von \mathcal{G} definieren:

$$P(x) = \prod_{A \in \mathcal{C}(\mathcal{G})} \phi_A(x_A)$$

Faktorisierungs- und Markoveigenschaften

Lemma

Für alle ungerichteten Graphen $\mathcal{G} = (V, \mathcal{E})$ und für alle Wahrscheinlichkeitsmodelle $P : \mathcal{X}_V \rightarrow \mathbb{R}$ gilt:

$$\boxed{\text{FAK} \Rightarrow \text{GME} \Rightarrow \text{LME} \Rightarrow \text{PME}}$$

Satz (Hammersley & Clifford, 1971)

Für jede streng positive Wahrscheinlichkeitsverteilung $P(\cdot)$ und jeden ungerichteten Graphen \mathcal{G} gilt:

$$\boxed{\text{FAK} \Leftrightarrow \text{GME} \Leftrightarrow \text{LME} \Leftrightarrow \text{PME}}$$

Bemerkung

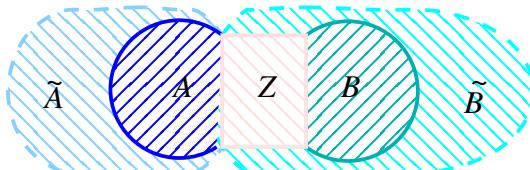
Im Falle numerischer Zufallsvariablen ist als Voraussetzung des HC-Satzes auch die **Existenz und Stetigkeit** der Dichtefunktion $f : \mathcal{X}_V \rightarrow \mathbb{R}$ zu fordern.

Beweis.

FAK \Rightarrow GME

Es seien $A, B, Z \subset V$ disjunkt mit $\text{sep}(A \mid Z \mid B)$.

Sei \tilde{A} die Zusammenhangshülle von A in $\mathcal{G}_{V \setminus Z}$ und sei $\tilde{B} = V \setminus (\tilde{A} \cup Z)$.



A, B gehören sicherlich zu verschiedenen Zusammenhangskomponenten im Restgraphen $\mathcal{G}_{V \setminus Z}$, also gilt für jede Clique $C \subset V$ genau eine der beiden Bedingungen

$$C \subseteq \tilde{A} \cup Z \quad \text{oder} \quad C \subseteq \tilde{B} \cup Z.$$

Die (garantierte: FAK) Faktorisierung gewinnt damit das folgendes Aussehen:

$$P(x) = \prod_{C \in \mathcal{C}} \phi_C(x_C) = \prod_{C \in \mathcal{C}_A} \phi_C(x_C) \cdot \prod_{C \in \mathcal{C}_B} \phi_C(x_C) = g(x_{\tilde{A} \cup Z}) \cdot h(x_{\tilde{B} \cup Z})$$

Nach Definition der bedingten Unabhängigkeit folgt daraus $\Im(\tilde{A} \mid Z \mid \tilde{B})$ und nach zweimaliger Anwendung des Axioms DEC auch die GME-Behauptung $\Im(A \mid Z \mid B)$. \square

Beweis.

GME \Rightarrow FAK

Aus völlig trivialen Gründen (auch $V \subseteq V$) gibt es eine Mammut-Faktorisierung à la

$$P(x) = \prod_{A \subset V} \phi_A(x_A).$$

Wegen der Eigenschaft $P(x) > 0$ strenger Positivität lässt sich diese Darstellung schmerzfrei logarithmieren:

$$\log P(x) = \sum_{A \subset V} \log \phi_A(x_A)$$

Nach einer sogenannten „Möbius-Inversion“ (sehr schwierig!) lassen sich in obigem Ausdruck durch Faktorisierung nach partiellen Unabhängigkeiten Zug um Zug alle Nicht-Cliquen-Summanden eliminieren. \square

Korrelation Assoziation Dependenz **Markovnetze** Bayesnetze Inferenz P-Lernen S-Lernen Gaußnetze Σ

FAK \Leftrightarrow GME für pathologische $P(\cdot)$

Moussouris (1974)

Gegenbeispiel

Betrachte $V = \{\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3, \mathbb{X}_4\}$ und die Verteilung

$$P(x) = \begin{cases} 1/8 & x \in \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \\ 0 & \text{sonst} \end{cases}$$

Für alle (x_2, x_4) ist entweder $P(\mathbb{X}_1 \mid x_2, x_4)$ oder $P(\mathbb{X}_3 \mid x_2, x_4)$ eine degenerierte Abbildung, also besteht trivialerweise keinerlei Abhängigkeit von \mathbb{X}_3 bzw. \mathbb{X}_1 . Gleiches gilt auch für alle (x_1, x_3) , also gilt insgesamt

$$\Im(\mathbb{X}_1 \mid \mathbb{X}_2, \mathbb{X}_4 \mid \mathbb{X}_3) \quad \wedge \quad \Im(\mathbb{X}_2 \mid \mathbb{X}_1, \mathbb{X}_3 \mid \mathbb{X}_4)$$

Der Diamant besitzt die GME, aber $P(\cdot)$ ist nicht \diamond -faktorisierbar:

$$\begin{aligned} 0 \neq 1/8 &= P(0, 0, 0, 0) = \phi_{1,2}(0, 0) \cdot \phi_{2,3}(0, 0) \cdot \phi_{3,4}(0, 0) \cdot \phi_{4,1}(0, 0) \\ 0 &= P(0, 0, 1, 0) = \phi_{1,2}(0, 0) \cdot \phi_{2,3}(0, 1) \cdot \phi_{3,4}(1, 0) \cdot \phi_{4,1}(0, 0) \\ 0 \neq 1/8 &= P(0, 0, 1, 1) = \phi_{1,2}(0, 0) \cdot \phi_{2,3}(0, 1) \cdot \phi_{3,4}(1, 1) \cdot \phi_{4,1}(1, 0) \\ 0 \neq 1/8 &= P(1, 1, 1, 0) = \phi_{1,2}(1, 1) \cdot \phi_{2,3}(1, 1) \cdot \phi_{3,4}(1, 0) \cdot \phi_{4,1}(0, 1) \end{aligned}$$

Korrelation Assoziation Dependenz **Markovnetze** Bayesnetze Inferenz P-Lernen S-Lernen Gaußnetze Σ

Zerlegbare graphische Modelle

Gibbs-Verteilungen

Verteilungen in Cliquenproduktform:

$$P(x) = P(x_V) = \prod_{C \in \mathcal{C}} \phi_C(x_C) / \sum_{x \in \Omega} \prod_{C \in \mathcal{C}} \phi_C(x_C)$$

Die Potentialfunktionen $\phi_C(\cdot)$ sind i.a. **keine** (\lightning) Wahrscheinlichkeiten.

Zerlegbarkeit

Wann zerfällt $P(x)$ in ein Produkt **bedingter Randverteilungen** ?

- Wenn die Cliques des Modellgraphen als Baum angeordnet sind!
- Die Baumstruktur regelt die Abhängigkeitsrichtungen.

Es besteht Freiheit in der Wahl, welche **Außencliquen** ein Blatt und welche eine Wurzel werden.

Beispiel — Markovketten I

Faktorisierung mit unterschiedlicher Variablenordnung

$$\mathbb{X}_1 \longleftrightarrow \mathbb{X}_2 \longleftrightarrow \mathbb{X}_3 \longleftrightarrow \mathbb{X}_4$$

Faktorisierung = Kettenregel + Unabhängigkeiten

$$P(x_1, x_2, x_3, x_4) = P(x_1) \cdot P(x_2|x_1) \cdot \underbrace{P(x_3|x_1, x_2)}_{P(x_3|x_2)} \cdot \underbrace{P(x_4|x_1, x_2, x_3)}_{P(x_4|x_3)}$$

Jede Variable kann als **Baumwurzel** nominiert werden — so auch \mathbb{X}_3 :

$$P(x_3, x_2, x_4, x_1) = P(x_3) \cdot P(x_2|x_3) \cdot \underbrace{P(x_4|x_3, x_2)}_{P(x_4|x_3)} \cdot \underbrace{P(x_1|x_3, x_2, x_4)}_{P(x_1|x_2)}$$

Aber **nicht jede Variablenfolge** ist mit der Baumstruktur verträglich:

$$P(x_1, x_4, x_2, x_3) = P(x_1) \cdot \underbrace{P(x_4|x_1)}_{\text{⚡}} \cdot \underbrace{P(x_2|x_1, x_4)}_{\text{⚡}} \cdot \underbrace{P(x_3|x_1, x_4, x_2)}_{\text{⚡}}$$

Beispiel — Markovketten II

Faktorisierung mit unterschiedlichen Cliquenbäumen

$$(\mathbb{X}_1, \mathbb{X}_2) \longleftrightarrow (\mathbb{X}_2, \mathbb{X}_3) \longleftrightarrow (\mathbb{X}_3, \mathbb{X}_4)$$

Faktorisierung = Cliquen + Baum + Wurzauswahl

$$P(x_1, x_2, x_3, x_4) = f(x_1, x_2) \cdot g(x_2, x_3) \cdot h(x_3, x_4)$$

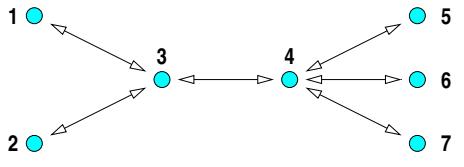
$$P(x_1, x_2) \cdot P(x_3|x_2) \cdot P(x_4|x_3)$$

$$P(x_1|x_2) \cdot P(x_2, x_3) \cdot P(x_4|x_3)$$

$$P(x_1|x_2) \cdot P(x_2|x_3) \cdot P(x_3, x_4)$$

⇒ Jede Wurzelnomierung definiert eine **valide** Modellformel.

Beispiel — Markovbäume I



Fakt

Ist \mathcal{G} ein Baum, so sind alle Cliquen zweielementig.

Die $N - 1$ Cliquen bilden selbst wieder einen Baum.

Faktorisierung im Beispiel

Kettenregel & Variablenbaumtraversierung

$$P(\mathbf{x}) = P(3) \cdot P(1|3) \cdot P(2|3) \cdot P(4|3) \cdot P(5|4) \cdot P(6|4) \cdot P(7|4)$$

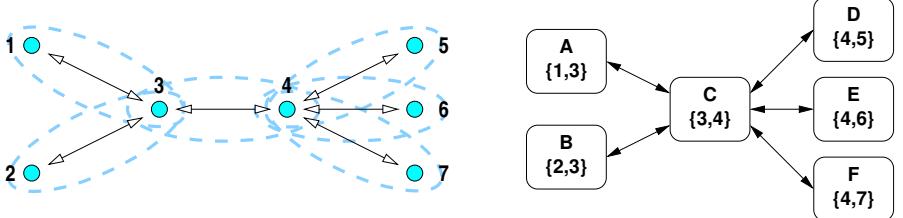
Faktorisierung allgemein

Traversieren ↪ konsistente Variablenordnung ↪ Einfachbedingungen

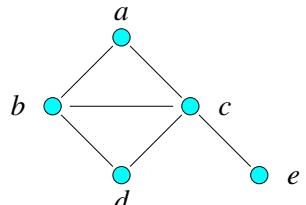
$$P(\mathbf{x}) = \prod_n P(x_n | \cdot) = \prod_n P(x_n | x_{\pi(n)})$$

Denn für alle $\mathbb{X}_n \in V$ gilt: $\text{sep}(\mathbb{X}_n | \mathbb{X}_{\pi(n)} | V \setminus \text{off}(\mathbb{X}_n))$

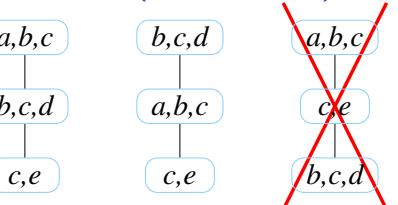
Beispiel — Markovbäume II



$$\begin{aligned} P(\mathbf{x}) &= \frac{\text{Cliquenwahrscheinlichkeit}}{\text{Cliquenschnittwahrscheinlichkeit}} \\ &= \frac{P(A) \cdot P(B) \cdot P(C) \cdot P(D) \cdot P(E) \cdot P(F)}{P(A \cap C) \cdot P(B \cap C) \cdot P(C \cap D) \cdot P(C \cap E) \cdot P(C \cap F)} \\ &= \frac{P(1,3) \cdot P(2,3) \cdot P(3,4) \cdot P(4,5) \cdot P(4,6) \cdot P(4,7)}{P(3) \cdot P(3) \cdot P(4) \cdot P(4) \cdot P(4)} \\ &= P(3) \cdot P(1|3) \cdot P(2|3) \cdot P(4|3) \cdot P(5|4) \cdot P(6|4) \cdot P(7|4) \end{aligned}$$

Beispiel — Cliquenverbundbaum

(„join tree“)

**Drei Cliquen — aber welche Baumstruktur?**

Die Cliquen $C_1 = \{a, b, c\}$, $C_2 = \{b, c, d\}$, $C_3 = \{c, e\}$ bilden paarweise einen nichtleeren Schnitt.

- ↙ $\Im(C_1|C_2|C_3) \rightsquigarrow C_1 - C_2 - C_3$ ist U-Bild von $P(\cdot)$
- ↙ $\Im(C_2|C_1|C_3) \rightsquigarrow C_2 - C_1 - C_3$ ist U-Bild von $P(\cdot)$
- ↙ $C_1 - C_3 - C_2$ ist **nicht** U-Bild von $P(\cdot)$, da $\neg \Im(C_1|C_3|C_2)$
- $\Im(C_1|C_2|C_3)$ und $\Im(C_2|C_1|C_3) \rightsquigarrow \Im(C_1, C_2|\emptyset|C_3)$ (INT) verletzt, also ≥ 2 minimale U-Bilder.

Zerlegung

Beide konsistenten Verbundbäume ergeben nach Traversierung:

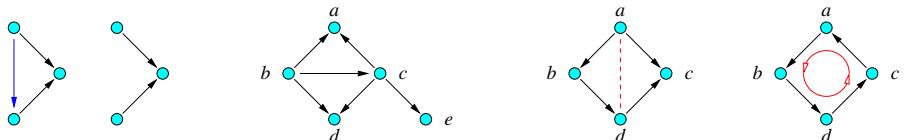
$$P(a, b, c, d, e) = P(a) \cdot P(b|a) \cdot P(c|a, b) \cdot P(d|b, c) \cdot P(e|c)$$

Moralische Graphen

„Alle Elternpaare sind (miteinander!) verheiratet.“

Definition

Ein gerichteter Graph heißt **moralisch**, wenn jedes konvergierende Kantenpaar aus zwei adjazenten Knoten entspringt.

**Satz**

Für einen ungerichteten Graphen \mathcal{G} sind die Eigenschaften äquivalent:

1. \mathcal{G} ist zerlegbar.
2. \mathcal{G} ist kordal.
3. \mathcal{G} lässt sich azyklisch und moralisch richten.
4. \mathcal{G} besitzt die Cliqueneliminationseigenschaft.
5. Es gibt einen verträglichen Verbundbaum für \mathcal{G} .

Kordalität und Zerlegbarkeit

Äquivalente Eigenschaften ungerichteter Graphen

Definition

Das Mengentripel (A, Z, B) heißt **Zerlegung des ungerichteten Graphen** $\mathcal{G} = (V, \mathcal{E})$, falls gilt:

Partition

Trennung

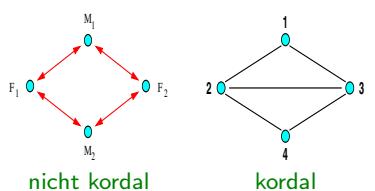
Vollständigkeit

$$A \uplus Z \uplus B = V \quad \text{sep}\langle A | Z | B \rangle \quad \mathcal{G}_Z \text{ ist vollständig}$$

Der Graph \mathcal{G} selbst heißt **zerlegbar**, wenn er vollständig ist oder aber eine Zerlegung mit zerlegbaren Teilgraphen $\mathcal{G}_{A \cup Z}$ und $\mathcal{G}_{B \cup Z}$ besitzt.

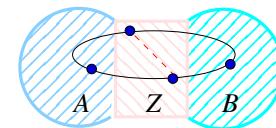
Definition

Ein ungerichteter Graph $\mathcal{G} = (V, \mathcal{E})$ heißt **kordal** oder **trianguliert** genau dann, wenn jeder Zyklus der Länge ≥ 4 mindestens eine **Sehne** besitzt.

**Beweis.**

Wir zeigen die Implikation „*zerlegbar* \Rightarrow *kordal*“

- **Induktionsanfang:** $|V| \leq 3$ impliziert trivialerweise die Kordalität.
- **Induktionsschritt:** Sei also (A, Z, B) eine Zerlegung von \mathcal{G} . Nach Induktionsvoraussetzung sind dann $\mathcal{G}_{A \cup Z}$ und $\mathcal{G}_{B \cup Z}$ kordal.



Angenommen, \mathcal{G} besitzt einen Zyklus ≥ 4 ohne Sehne. Dieser muß wegen der I.V. Knoten in A und auch in B haben, passiert also mindestens 2x die Menge Z und teilt deshalb ≥ 2 Knoten mit Z . Diese sind aber wegen der Vollständigkeit von Z verbunden — \blacksquare

 \square

Cliqueneliminationseigenschaft

Definition

Der ungerichtete Graph \mathcal{G} besitzt die **Cliqueneliminationseigenschaft**, wenn alle Knoten aller seiner Cliques durch wiederholte Anwendung folgender Operationen eliminiert werden können:

- **Unikatknoten**

Lösche einen Knoten, der nur in einer einzigen Clique auftaucht.

- **Dominierte Mengen**

Lösche eine Clique, die Teilmenge einer anderen Clique ist.

Der Schlüsselgraph besitzt die CEP

Schachmatt in sieben Zügen:

$\{a, b, c\}$	$\{b, c, d\}$	$\{c, e\}$	die 3 Cliques des Verbundbaumbeispiels
$\{b, c\}$	$\{b, c\}$	$\{c\}$	Unikate a, d und e gelöscht
$\{b, c\}$	•	•	zwei dominierte Cliques gelöscht
•	•	•	Unikate b und c gelöscht

Verträgliche Verbundbäume

Definition

Sei $\mathcal{G} = (V, \mathcal{E})$ ein ungerichteter Graph. Der Graph \mathcal{G}^* ist ein **mit \mathcal{G} verträglicher Verbundbaum**, falls gilt:

1. Die Knoten \mathcal{G}^* sind genau die Cliques $\mathcal{C}(\mathcal{G})$.
2. \mathcal{G}^* ist zusammenhängend und zyklischfrei.
3. Für jeden Knoten $a \in V$ gilt:

Je zwei a enthaltende Cliques besitzen einen Verbindungsgraphen, der ausschließlich Cliques C mit $a \in C$ enthält.

Für den Schlüsselgraphen ist VB #3 nicht verträglich

Im dritten Verbundbaum

$$\underbrace{\{a, b, c\}}_{C_1} \longrightarrow \underbrace{\{c, e\}}_{C_3} \longrightarrow \underbrace{\{b, c, d\}}_{C_2}$$

gilt $b \in C_1$ und $b \in C_2$, aber es ist $b \notin C_3$, obwohl C_3 auf dem einzigen verfügbaren Pfad von C_1 nach C_2 liegt.

Zerlegbarkeit und Faktorisierung

Lemma (Cliquenschnittformel)

Sei \mathcal{G} ein zerlegbarer ungerichteter Graph. Dann gilt für alle $P(\cdot)$

$$\text{FAK} \quad \Leftrightarrow \quad \text{GME}$$

und diese Faktorisierung besteht aus cliquenbezogenen Randverteilungen:

$$P(x) = \prod_{C \in \mathcal{C}} \frac{P(x_C)}{P(x_{C \cap \pi(C)})}$$

Dabei bezeichnet $\pi(C)$ die eindeutig bestimmte Vorgängerclique von C im (festen, aber beliebigen) verträglichen Verbundbaum.

Die beiden CSF für den Schlüsselgraphen

Die Verbundbäume $\{a, b, c\} \rightarrow \{b, c, d\} \rightarrow \{c, e\}$ und $\{b, c, d\} \rightarrow \{a, b, c\} \rightarrow \{c, e\}$ liefern die äquivalenten Faktorisierungen

$$P(\cdot) = \frac{P(a, b, c) \cdot P(b, c, d) \cdot P(c, e)}{P(b, c) \cdot P(c)} \quad \text{und} \quad P(\cdot) = \frac{P(b, c, d) \cdot P(a, b, c) \cdot P(c, e)}{P(b, c) \cdot P(c)}.$$

Beweis.

GME \Rightarrow FAK

(die Umkehrung gilt ja sowieso)

Induktion über die Zerlegungshierarchie von \mathcal{G} :

Sei $\text{sep}(A|Z|B)$ eine Zerlegung von \mathcal{G} . Dann gilt

$$\begin{aligned} P(x_V) &= P(x_{A \cup Z}) \cdot P(x_B | x_{A \cup Z}) \\ &= P(x_{A \cup Z}) \cdot P(x_B | x_Z) \\ &= \frac{P(x_{A \cup Z}) \cdot P(x_{B \cup Z})}{P(x_Z)} \end{aligned}$$

wegen $\text{G}(A|Z|B)$ nach GME.

Der Nenner $P(x_Z)$ ist bereits ein Cliquenfaktor, weil Z vollständig ist.

Die beiden Zählerterme sind nach Induktionsvoraussetzung über $\mathcal{G}_{A \cup Z}$ bzw. $\mathcal{G}_{B \cup Z}$ faktorisierbar, bestehen also ausschließlich aus Cliquentermen.

Die behauptete Faktorisierung ergibt sich durch Zusammenfassen und Umgruppieren nach \mathcal{G} -Cliques. □

Beweis.

Cliquenschnittformel

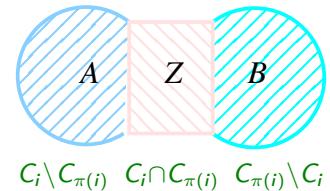
Es sei C_1, \dots, C_M eine mit der Nachfolgerrelation eines Cliquenverbundbaums von \mathcal{G} verträgliche Cliquenordnung.

Für jede Clique C_i bezeichne $C_{\pi(i)}$ die eindeutig bestimmte Elterclique ($\pi(i) < i$). Dann gilt (für alle i) die Trennungsrelation

$$\text{sep}\langle C_i \mid C_{\pi(i)} \mid C_1, \dots, C_{i-1} \rangle$$

und wegen GME auch die entsprechende bedingte Unabhängigkeit.

$$\begin{aligned} P(x) = P(x_1, \dots, x_N) &= \prod_{i=1}^M P(x_{C_i} \mid x_{C_1}, \dots, x_{C_{i-1}}) \\ &= \prod_{i=1}^M P(x_{C_i} \mid x_{C_{\pi(i)}}) \\ &= \prod_{i=1}^M P(x_{C_i} \mid x_{C_i \cap C_{\pi(i)}}) \\ &= \prod_{i=1}^M \frac{P(x_{C_i})}{P(x_{C_i \cap C_{\pi(i)}})} \end{aligned}$$



□

Graphtriangulierung & Verbundbaumkonstruktion

1 KNOTENORDNUNG

Ordne Knoten nach maximalem Rang; setze sukzessiv:

$$v_{i+1} \stackrel{\text{def}}{=} \underset{v \notin V(i)}{\operatorname{argmax}} |\{v' \in V \mid (v, v') \in \mathcal{E}, v' \in V(i)\}|$$

2 KANTENERZEUGUNG

Für $i = N, \dots, 1$

$$\mathcal{E} \leftarrow \mathcal{E} \cup \{v', v''\}$$

falls $v', v'' \in V(i-1)$ und falls $\{v_i, v'\}, \{v_i, v''\} \in \mathcal{E}$.

3 CLIQUENORDNUNG

Fixiere Reihenfolge C_1, \dots, C_M nach dem maximalen Knotenrang.

4 KANTENERZEUGUNG

Für $i = 2, \dots, M$ erzeuge neue Kante $C_{\pi(i)} \rightarrow C_i$ mit

$$\pi(i) < i \quad \text{und } |C_{\pi(i)} \cap C_i| \text{ ist maximal.}$$

Beispiel

Knotenfolge:
 $a^0 b^1 c^2 d^2 e^1$

Neue
Kanten:
(keine)

Cliquenfolge:
 $C_1 : abc^{012}$
 $C_2 : bcd^{122}$
 $C_3 : ce^{21}$
(befiebig)

VB-Kanten:
 $1 \rightarrow 2,$
 $1 \rightarrow 3$

Zwischenbilanz

für ungerichtete graphische Modelle

1. Nicht jede Verteilung ist graphisch.
2. Streng positive Verteilungen erlauben aber, mit dem Kantenlöschverfahren ein Markovnetz (minimales U-Bild) zu erzeugen.
3. Markovnetze faktorisieren gemäß ihrer Cliquenstruktur, aber nicht zwingend in Wahrscheinlichkeiten.
4. Durch Triangulieren des Markovnetzes werden einige Unabhängigkeiten außer Gefecht gesetzt, aber dafür gewinnen wir eine Kettenregel (CSF).

Korrelation, Regression und Transformation

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Ursache und Wirkung

Gerichtete azyklische Graphen

Kausalrichtung

Drei Attribute · zwei Interaktionen · drei Wirkkonfigurationen:

kaskadierend „Wetter“ → „Ernte“ → „Preis“

$$\Im(\mathbb{X}_1 | \mathbb{X}_2 | \mathbb{X}_3)$$

divergierend „Größe“ ← „Alter“ → „Lesefähigkeit“

$$\Im(\mathbb{X}_1 | \mathbb{X}_2 | \mathbb{X}_3)$$

konvergierend „Würfel₁“ → „Glocke“ ← „Würfel₂“

$$\neg \Im(\mathbb{X}_1 | \mathbb{X}_2 | \mathbb{X}_3)$$

Modelle kausaler Beziehungen: $\left\{ \begin{array}{l} \text{erklärende} \\ \text{vermittelnde} \\ \text{diagnostische} \end{array} \right\}$ Variablen.

Lemma (Erinnerung)

Ein gerichteter Graph $\mathcal{G} = (V, \mathcal{E})$, $\mathcal{E} \subseteq V \times V$, ist **azyklisch** genau dann, wenn es eine **kantenverträgliche Knotenordnung** $V = \{v_1, \dots, v_N\}$ gibt:

$$(v_i, v_j) \in \mathcal{E} \Leftrightarrow i < j \quad \text{für alle } i, j \in \{1, \dots, N\}$$

M.a.W.: Ein DAG („directed acyclic graph“) besitzt keine gerichteten Zyklen (**Pfade**); ungerichtete Zyklen (**Ketten**) sind hingegen erlaubt.

Ein Trennungskriterium

Satz

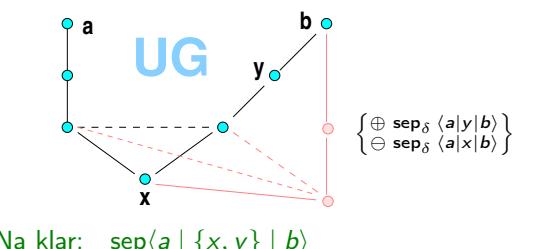
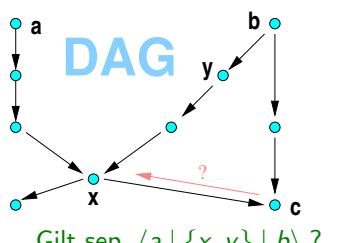
In einem DAG $\mathcal{G} = (V, \mathcal{E})$ gilt für alle disjunkten Mengen $A, B, Z \subset V$:

$$\text{sep}_{\delta}(A | Z | B)_{\mathcal{G}} \Leftrightarrow \text{sep}(A | Z | B)_{\mathcal{G}^*}$$

Dabei bezeichne W die Vorgängerhülle von $A \cup Z \cup B$ und \mathcal{G}^* sei der **moralische Graph** $(\mathcal{G}_W)^m$ von \mathcal{G}_W .

Beispiel

- obere Kette: y blockiert, aber x blockiert nicht!
- untere Kette: x blockiert und c blockiert.



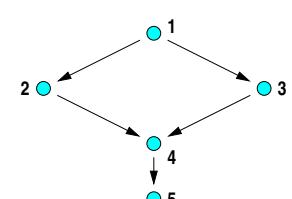
δ -Trennungsrelation

für gerichtete azyklische Graphen

Definition

Es sei $\mathcal{G} = (V, \mathcal{E})$ ein gerichteter azyklischer Graph und $A, B, Z \subset V$ disjunkte Knotenmengen. Eine **Kette** zwischen den Knoten a und b heißt **blockiert von Z**

- wenn sie einen nichtkonvergierenden Knoten $c \in Z$ enthält
- oder wenn sie einen konvergierenden Knoten $c \notin Z$ enthält, der auch keinen Nachfolger in Z besitzt.



Beispiel

Es gilt $\text{sep}_{\delta}(2 | 1 | 3)$, denn:

Die Kette $2 \leftarrow 1 \rightarrow 3$ ist von \mathbb{X}_1 blockiert wegen $1 \in Z = \{1\}$

Die Kette $2 \rightarrow 4 \leftarrow 3$ ist von \mathbb{X}_4 blockiert wg. $4, 5 \notin Z = \{1\}$

Kausale Verteilungen und Dependenzmodelle

Überrepräsentation & Unterrepräsentation von $\Im(\cdot | \cdot | \cdot)$ durch $\text{sep}_{\delta}(\cdot | \cdot | \cdot)$

Definition

Es sei $P(\cdot)$ eine Wahrscheinlichkeitsverteilung auf V und \Im ihr Dependenzmodell. Der gerichtete azyklische Graph $\mathcal{G} = (V, \mathcal{E})$ heißt

- **Abhängigkeitsbild** von P gdw.

$$\Im(A | Z | B) \Rightarrow \text{sep}_{\delta}(A | Z | B)$$

- **Unabhängigkeitsbild** von P gdw.

$$\Im(A | Z | B) \Leftarrow \text{sep}_{\delta}(A | Z | B)$$

- **perfektes Bild** von P gdw.

$$\Im(A | Z | B) \Leftrightarrow \text{sep}_{\delta}(A | Z | B)$$

Die Verteilung $P(\cdot)$ (und das Modell \Im) heißen **kausal**, wenn ein gerichteter azyklischer Graph existiert, der \Im perfekt abbildet.

Rekursive Faktorisierung

Definition

Die Wahrscheinlichkeitsverteilung $P(\cdot)$ zerfällt rekursiv über dem gerichteten azyklischen Graphen $\mathcal{G} = (V, \mathcal{E})$, wenn es für jede Variable $a \in V$ eine nichtnegative **Kernfunktion**

$$\phi_a : \mathcal{X}_a \times \bigotimes_{v \in \text{pa}(a)} \mathcal{X}_v \rightarrow \mathbb{R}_0^+$$

gibt mit

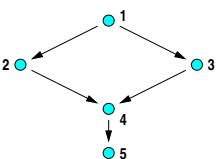
$$P(\mathbf{x}) = \prod_{a \in V} \phi_a(x_a, \mathbf{x}_{\text{pa}(a)})$$

Es bezeichnet $\text{pa}(a)$ die **Eltermenge** $\{v \mid (v, a) \in \mathcal{E}\}$ von a .

Beispiel

Im Rasensprengergraphen zerfällt die Verteilung, falls es Potentialfunktionen gibt mit:

$$P(\mathbf{x}) = \phi_1(x_1) \cdot \phi_2(x_1, x_2) \cdot \phi_3(x_1, x_3) \cdot \phi_4(x_2, x_3, x_4) \cdot \phi_5(x_4, x_5)$$



Beweis.

Wir vereinbaren eine verträgliche Ordnung $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$ und wir wissen, daß nun die Kausalitätsbeziehung gilt:

$$\mathbb{X}_j \in \text{pa}(\mathbb{X}_i) \Rightarrow j < i$$

Wir berechnen nun die Randverteilung der ersten n Variablen:

$$\begin{aligned} P(\mathbf{x}_{1..n}) &= \sum_{x_{n+1}} \dots \sum_{x_N} P(\mathbf{x}) = \sum_{x_{n+1}} \dots \sum_{x_N} \prod_{i=1}^N \phi_i(x_i, \mathbf{x}_{\text{pa}(\mathbb{X}_i)}) \\ &= \prod_{i=1}^n \phi_i(x_i, \mathbf{x}_{\text{pa}(\mathbb{X}_i)}) \cdot \prod_{i=n+1}^N \left(\underbrace{\sum_{x_i \in \mathcal{X}_i} \phi_i(x_i, \mathbf{x}_{\text{pa}(\mathbb{X}_i)})}_{\sigma_i} \right) \end{aligned}$$

Daraus folgt für die bedingte Wahrscheinlichkeit $P(x_n \mid \mathbf{x}_{1..n-1})$:

$$\dots = \frac{P(\mathbf{x}_{1..n})}{P(\mathbf{x}_{1..n-1})} = \frac{\prod_{i=1}^n \phi_i(x_i, \mathbf{x}_{\text{pa}(\mathbb{X}_i)}) \cdot \prod_{i=n+1}^N \sigma_i}{\prod_{i=1}^{n-1} \phi_i(x_i, \mathbf{x}_{\text{pa}(\mathbb{X}_i)}) \cdot \prod_{i=n}^N \sigma_i} = \frac{\phi_n(x_n, \mathbf{x}_{\text{pa}(\mathbb{X}_n)})}{\sigma_n}$$

Wenn wir also normierte Faktoren verwenden ($\sigma_n \equiv 1$), entsprechen die ϕ_n gerade den klassischen Kettenregelgliedern $P(x_n \mid \cdot)$. Daß diese tatsächlich nur von \mathbb{X}_n und deren Elternteilvariablen abhängen, ergibt sich aus der Argumentstruktur von ϕ_n . \square

Die reduzierte Kettenregel

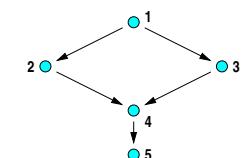
Lemma

Wenn $P(\cdot)$ über \mathcal{G} rekursiv zerfällt, können die Kernfunktionen ϕ_a o.B.d.A. gemäß

$$\phi_a(x_a, \mathbf{x}_{\text{pa}(a)}) = P_{a|\text{pa}(a)}(x_a \mid \mathbf{x}_{\text{pa}(a)})$$

als bedingte Einzelwertwahrscheinlichkeiten gestaltet werden und es gilt — bei kantenverträglicher Variablenordnung — die **reduzierte Kettenregel**:

$$P(\mathbf{x}) = \prod_{i=1}^N P(x_i \mid \mathbf{x}_{\text{pa}(\mathbb{X}_i)})$$



Beispiel

Im Rasensprengergraphen kann die Faktorisierung wie folgt gewählt werden:

$$P(\mathbf{x}) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_1) \cdot P(x_4 \mid x_2, x_3) \cdot P(x_5 \mid x_4)$$

Faktorisierung und Markoveigenschaft

Satz

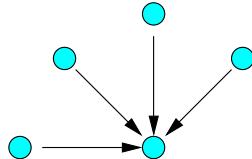
Wenn die Verteilung $P(\cdot)$ über dem DAG \mathcal{G} rekursiv zerfällt, dann zerfällt $P(\cdot)$ auch über dem moralischen Graphen $(\mathcal{G})^m$ von \mathcal{G} .

\mathcal{G} ist dann sicherlich ein Unabhängigkeitsbild von $P(\cdot)$, d.h. es gilt:

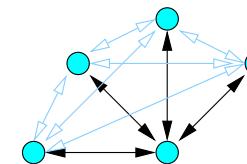
FAK* GME*

FAK GME

Beweisidee
Moralgrapherzeugung



Knoten mit allen Eltern in \mathcal{G}



„just married“ (Clique in $(\mathcal{G})^m$)

Die drei Markoveigenschaften

Beweis.

- FME* \Rightarrow FME

(moralische Faktorisierung)

Für jede Variable $a \in V$ ist die Menge $\{a\} \cup pa(a)$ im moralischen Graphen $(\mathcal{G})^m$ von \mathcal{G} vollständig, denn a ist mit jedem Elter adjazent und alle Eltern wurden miteinander verheiratet.

Damit bilden die Potentialfunktionen ϕ_a auch eine Cliquenfaktorisierung auf $(\mathcal{G})^m$.

- FME \Rightarrow GME

(für $(\mathcal{G})^m$; gilt immer)

- GME \Rightarrow GME*

Gilt nun $\text{sep}_\delta(A | Z | B)$ in \mathcal{G} , so ist auch $\text{sep}(A | Z | B)$ in $(\mathcal{G})^m$.

Es besitzt $(\mathcal{G})^m$ die globale ME für $P(\cdot)$, also ist auch $\mathfrak{S}(A | Z | B)$.

□

Definition

Es sei $P(\cdot)$ eine Wahrscheinlichkeitsverteilung auf V und \mathfrak{S} ihr Dependenzmodell. Der gerichtete azyklische Graph $\mathcal{G} = (V, \mathcal{E})$ erfüllt die

- paarweise Markoveigenschaft**

gdw. für alle $a \not\rightarrow b \in V$ mit $b \notin \text{off}(a)$ gilt:

$$\mathfrak{S}(a | V \setminus \text{off}(a) \setminus \{b\} | b)$$

- lokale Markoveigenschaft**

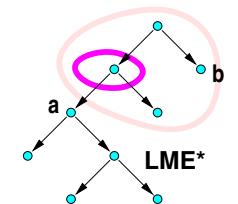
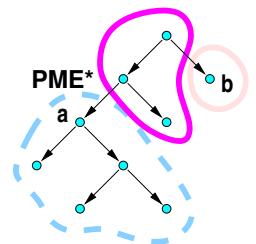
gdw. für jede Variable $a \in V$ gilt:

$$\mathfrak{S}(a | pa(a) | V \setminus \text{off}(a))$$

- globale Markoveigenschaft**

gdw. für alle $A, B, Z \subset V$ mit $\text{sep}_\delta(A | Z | B)$ gilt:

$$\mathfrak{S}(A | Z | B)$$



Beweis.

$$\text{PME*} \not\Rightarrow \text{LME*}$$

(alle anderen Richtungen nur im alten Vorlesungsskriptum)

Als Gegenbeispiel betrachte die vier binärwertigen, uniform verteilten Zufallsvariablen $\mathbb{X} = \mathbb{Y} = \mathbb{Z}$ und \mathbb{W} und den DAG mit den Kanten

$$\mathbb{Z} \rightarrow \mathbb{W} \rightarrow \mathbb{X} \quad \text{und} \quad \mathbb{Z} \rightarrow \mathbb{Y} \rightarrow \mathbb{W}.$$

Der Graph besitzt die paarweise ME, denn von den insgesamt vier nichtadjazenten Variablenpaaren erfüllen nur (\mathbb{X}, \mathbb{Y}) und (\mathbb{X}, \mathbb{Z}) die Nachkommenbedingung. Damit sind

$$\mathfrak{S}(\mathbb{X} | \mathbb{W}, \mathbb{Z} | \mathbb{Y}) \quad \text{und} \quad \mathfrak{S}(\mathbb{X} | \mathbb{W}, \mathbb{Y} | \mathbb{Z})$$

zu überprüfen — die Faktorzerlegung ergibt sich aber wie folgt:

$$P(x, y | z, w) = \begin{cases} 1 & x = y = z \\ 0 & \text{sonst} \end{cases} = \delta_{xz} \cdot \delta_{yz}$$

Ganz analog ergibt sich auch $P(x, z | y, w) = \delta_{xy} \cdot \delta_{zy}$. Der Graph besitzt aber nicht die lokale ME, denn die Unabhängigkeit

$$\mathfrak{S}(\mathbb{X} | \underbrace{\text{pa}(\mathbb{X})}_{\mathbb{W}} | \underbrace{V \setminus \text{off}(\mathbb{X})}_{\{\mathbb{W}, \mathbb{Y}, \mathbb{Z}\}})$$

bedingt nach Axiom DEC auch $\mathfrak{S}(\mathbb{X} | \mathbb{W} | \mathbb{Y}, \mathbb{Z})$, was die Verteilung $P(\cdot)$ offensichtlich nicht hergibt.

□

Die Markoveigenschaften für Semi-/Graphoide

Bayesnetze $\hat{=}$ minimale Unabhängigkeitsbilder

Definition

Der Graph \mathcal{G} heißt **Bayesnetz** von $P(\cdot)$, wenn er minimal mit der globalen Markoveigenschaft für $P(\cdot)$ ist.

Das Bayesnetz \mathcal{G} ignoriert keine Abhängigkeiten, höchstens Unabhängigkeiten, aber davon so wenige wie möglich.

Satz

Sei $\mathcal{G} = (V, \mathcal{E})$ und $P(\cdot)$ auf V gegeben. Dann gilt

$$\text{FAK*} \iff \text{GME*} \iff \text{LME*} \iff \text{paarweise ME ,}$$

aber es gilt im allgemeinen nicht die Umkehrrichtung

$$\text{PME*} \Rightarrow \text{LME*}.$$

Für streng positive Verteilungen $P(\cdot)$ gilt sogar die Äquivalenz

$$\text{LME*} \iff \text{PME*}.$$

Axiomatisierung kausaler Dependenzmodelle ?

Satz

Ist das Dependenzmodell \mathfrak{S} kausal, so gelten die folgenden sieben unabhängigen Axiome:

SYM Symmetrie

$$\mathfrak{S}(A | Z | B) \Leftrightarrow \mathfrak{S}(B | Z | A)$$

C/D Komposition/Dekomposition

$$\mathfrak{S}(A | Z | B \cup C) \Leftrightarrow \mathfrak{S}(A | Z | B) \wedge \mathfrak{S}(A | Z | C)$$

INT Durchschnitt

$$\mathfrak{S}(A | Z \cup C | B) \wedge \mathfrak{S}(A | Z \cup B | C) \Rightarrow \mathfrak{S}(A | Z | B \cup C)$$

WUN Schwache Vereinigung

$$\mathfrak{S}(A | Z | B \cup C) \Rightarrow \mathfrak{S}(A | Z \cup C | B)$$

CON Kontraktion

$$\mathfrak{S}(A | Z | B) \wedge \mathfrak{S}(A | Z \cup B | C) \Rightarrow \mathfrak{S}(A | Z | B \cup C)$$

WTR Schwache Transitivität

$$\mathfrak{S}(A | Z | B) \wedge \mathfrak{S}(A | Z \cup \{x\} | B) \Rightarrow \mathfrak{S}(A | Z | \{x\}) \vee \mathfrak{S}(\{x\} | Z | B)$$

CHO Kordalität

$$\mathfrak{S}(a | c, d | b) \wedge \mathfrak{S}(c | a, b | d) \Rightarrow \mathfrak{S}(a | c | b) \vee \mathfrak{S}(a | d | b)$$

Bayesnetzkonstruktion

Lemma (Verma 1986)

Ist \mathfrak{S} ein Semigraphoid, so ist jeder Grenzengraph von \mathfrak{S} ein Bayesnetz von \mathfrak{S} .

Ist \mathfrak{S} ein Graphoid, so ist der Grenzengraph von \mathfrak{S} bei gegebener **Variablenordnung** eindeutig.

\mathcal{G} ist ein Bayesnetz für die Verteilung $P(\cdot)$ genau dann, wenn er die LME* für $\mathfrak{S} = \mathfrak{S}_P$ besitzt und die Eltermengen $pa(\mathbb{X}_n)$ minimal mit dieser Eigenschaft sind (Markovgrenzen von \mathbb{X}_n bzgl. $V \setminus \text{off}(\mathbb{X}_n)$).

(Algorithmus)

- 1 Wähl eine Variablenordnung $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$ aus.
- 2 Wähle \mathbb{X}_1 als Wurzel und ordne die Randverteilung $P_1(x_1)$ zu.
- 3 Für alle $i \geq 2$ berechne ein minimales B_i mit

$$B_i \subseteq \{\mathbb{X}_1, \dots, \mathbb{X}_{i-1}\} \quad \text{und} \quad P(x_i | x_1, \dots, x_{i-1}) = P(x_i | x_{B_i})$$

und kreiere Knoten \mathbb{X}_i mit der Vorgängermenge $pa(\mathbb{X}_i) = B_i$ und der lokalen Verteilung $P_i(x_i | x_{B_i})$.

(Endalgorithmus)

Markovdecken und Markovgrenzen

Definition

Sei \mathfrak{S} ein Dependenzmodell auf V und $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$ eine **Variablenordnung**.

- Eine Menge $B \subset V$ heißt **Markovdecke** von $c \in V$ bezüglich $A \subset V$ genau dann, wenn gilt:

$$B \subseteq A \wedge \mathfrak{S}(\{c\} | B | A \setminus B)$$

- Ist B minimal mit dieser Eigenschaft, so heißt B eine **Markovgrenze**.
- Die Folge B_1, \dots, B_N heißt **Grenzensystem** von \mathfrak{S} bezüglich Variablenordnung $\mathbb{X}_1, \dots, \mathbb{X}_N$ genau dann, wenn jede Menge B_n eine Markovgrenze von \mathbb{X}_n bezüglich $V_n = \{\mathbb{X}_1, \dots, \mathbb{X}_{n-1}\}$ ist.
- Ein gerichteter azyklischer Graph \mathcal{G} , dessen Eltermengen $pa(\mathbb{X}_n)$ ein Grenzensystem von \mathfrak{S} bilden, heißt **Grenzengraph** von \mathfrak{S} .

Markovdecken einer Markovkette: $\mathfrak{S}(\mathbb{X}_n | \{\mathbb{X}_{n-1}, \mathbb{X}_{n+1}\} | V \setminus \{\mathbb{X}_{n-1}, \mathbb{X}_n, \mathbb{X}_{n+1}\})$

Markovdecken gegen den Rest der Welt

Fragestellung

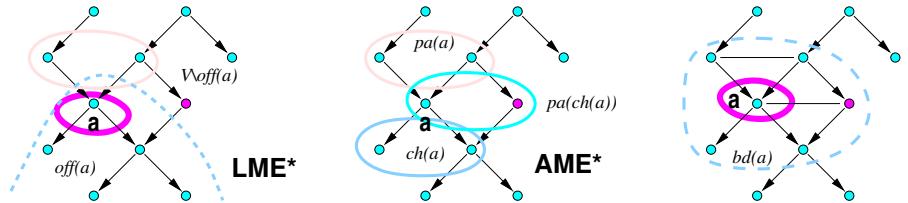
In ungerichteten Graphen fallen die beiden folgenden Fragestellungen zusammen:

LME Lokale Markoveigenschaft: Gegen welche Variablen wird $a \in V$ durch seine unmittelbaren Nachbarn $bd(a)$ abgeschirmt?

AME Allgemeine Markoveigenschaft: Durch welche Menge wird $a \in V$ gegen den „Rest der Welt“ abgeschirmt?



Allgemeine Markoveigenschaft



Lemma (AME*)

Es sei \mathcal{G} ein Bayesnetz für \mathfrak{S} . Für jedes $a \in V$ bildet die Vereinigung der folgenden Variablenmengen eine Markovdecke bzgl. V :

1. die Menge $pa(a)$ der direkten Vorfahren von a ,
2. die Menge $ch(a)$ der direkten Nachkommen von a ,
3. die Menge der direkten Vorfahren der direkten Nachkommen von a .

Mit anderen Worten:

$$\mathfrak{S}(a | pa(a) \cup ch(a) \cup pa(ch(a)) \setminus \{a\} | \text{"Rest"})$$

Graphische versus kausale Verteilungen

Lemma

1. Es gibt graphische Verteilungen, die nicht kausal sind.
2. Es gibt kausale Verteilungen, die nicht graphisch sind.
3. Es gibt Verteilungen, die weder graphisch noch kausal sind.
4. Ist $P : \mathcal{X} \rightarrow \mathbb{R}$ sowohl graphisch als auch kausal, so ist jedes Markovnetz von \mathfrak{S}_P kordal/trianguliert.
5. Ist $P : \mathcal{X} \rightarrow \mathbb{R}$ sowohl graphisch als auch kausal, so ist jedes Bayesnetz von \mathfrak{S}_P moralisch.

Beweisidee

$$\mathfrak{S}(\{a\} | \text{"Rest"} | \{b\})_{P(\cdot)}$$



$$\text{sep}\langle A|Z|B \rangle_{\mathcal{G}_{UG}}$$



$$\mathfrak{S}(A|Z|B)_{P(\cdot)}$$



$$\text{sep}_{\delta} \langle A|Z|B \rangle_{\mathcal{G}_{DAG}}$$



$$\text{sep}\langle A|Z|B \rangle_{(\mathcal{G}_{DAG})^m}$$

Beweis.

- \mathcal{G} ist ein Bayesnetz von \mathfrak{S} , also insbesondere ein Unabhängigkeitsbild; folglich gilt die GME*.
- Wir haben also nur die Trennungseigenschaft

$$\text{sep}_{\delta} \langle a | B_a | \text{"Rest"} \rangle_{\mathcal{G}}, \quad B_a \stackrel{\text{def}}{=} \text{pa}(a) \cup \text{ch}(a) \cup \text{pa}(\text{ch}(a)) \setminus \{a\}$$

zu zeigen.

- Die Trennungseigenschaft beweisen wir im Moralgraphen $(\mathcal{G})^m$. Dort hat Knoten $a \in V$ als Nachbarn genau alle ehemaligen Eltern und Kinder des DAG sowie zusätzlich all jene Knoten, zu denen gemeinsame Kinder in \mathcal{G} existieren, mit anderen Worten gilt:

$$\text{bd}_{(\mathcal{G})^m}(a) = B_a$$

- Selbstverständlich wird a im Moralgraphen $(\mathcal{G})^m$ — wie in jedem UG wegen der LME — durch seinen Rand $\text{bd}_{(\mathcal{G})^m}(a)$ von allen Restknoten getrennt:

$$\text{sep}_{\delta} \langle a | B_a | \text{"Rest"} \rangle_{(\mathcal{G})^m}$$

Damit ist die Behauptung gezeigt. □

Beweis.

1. Jedes P mit dem Diamant-UG $\overset{m_1}{f_1} \diamond \overset{m_2}{f_2}$ als perfektem Bild.
 2. Jedes P mit dem Konvergenz-DAG $w_1 \rightarrow g \leftarrow w_2$ als perfektem Bild.
 3. Jede nichtkausale loglineare Verteilung mit der Modellformel
- $$P(a, b, c) = \phi_1(b, c) \cdot \phi_2(a, c) \cdot \phi_3(a, b)$$
- denn der **vollständige UG** ist das eindeutige Markovnetz zu P, enthält aber die $\{b, c\}, \{a, c\}, \{a, b\}$ nicht als Cliques.
4. Wegen des Spezialfalls partieller Unabhängigkeiten besitzen \mathcal{G}_{UG} und $(\mathcal{G}_{DAG})^m$ identische Kanten, das Markovnetz ist also der Moralgraph des Bayesnetzes. \mathcal{G}_{UG} muß dann aber auch kordal sein, denn jeder Kreis ≥ 4 muß im (azyklischen) Bayesnetz einen konvergierenden Knoten besitzen, folglich (aus Gründen der Moral) auch eine Sehne.
 5. Im Falle der Unmoral gäbe es $a \rightarrow z \leftarrow b$, aber weder $a \rightarrow b$ noch $a \leftarrow b$. Für die „historischen Abschlüsse“ Z von $\{z\}$ und W von $\{a, b\}$ gilt dann aber

$$\text{sep}\langle \{a\} | W \setminus \{a, b\} | \{b\} \rangle_{(\mathcal{G}_W)^m},$$

aber nicht

$$\text{sep}\langle \{a\} | Z \setminus \{a, b\} | \{b\} \rangle_{(\mathcal{G}_Z)^m},$$

ein eklatanter **⚡** zum SUN-Axiom (P graphisch!), da $W \setminus \{a, b\} \subset Z \setminus \{a, b\}$ gilt. □

Beispiele

Markovnetze mit 3, 4, 5 oder 6 Variablen

• • •	$P(x) \cdot P(y) \cdot P(z)$	3-diskret	\oplus
•—• •	$P(x, y) \cdot P(z)$	2+1-diskret	\oplus
•—•—•	$P(x, y) \cdot P(y, z) / P(y)$	kaskadiert	\oplus
Δ	$P(x, y, z)$	saturiert	\oplus
\diamond	$\phi(x, y) \cdot \phi(y, z) \cdot \phi(z, w) \cdot \phi(w, x)$	Diamant	\ominus
$\triangleleft\triangleright$	$P(x, y, z) \cdot P(y, z, w) / P(y, z)$	3/3-Cliquen	\oplus
$\triangleright\triangleleft$	$P(x, y, z) \cdot P(v, w, z) / P(z)$	3/3-Cliquen	\oplus
$\Delta \equiv \Delta$	$\phi(x_1, x_2, x_3) \cdot \phi(y_1, y_2, y_3) \cdot \phi(x_1, y_1) \cdot \phi(x_2, y_2) \cdot \phi(x_3, y_3)$	Toblerone	\ominus

Beispiele

Bayesnetze mit 3 oder 4 Variablen

• • •	$P(x) \cdot P(y) \cdot P(z)$	3-diskret	\oplus
• \rightarrow • •	$P(x) \cdot P(y x) \cdot P(z)$	2+1-diskret	\oplus
• \rightarrow • \rightarrow •	$P(x) \cdot P(y x) \cdot P(z y)$	kaskadiert	\oplus
• \leftarrow • \rightarrow •	$P(x y) \cdot P(y) \cdot P(z y)$	divergent	\oplus
• \rightarrow • \leftarrow •	$P(x) \cdot P(y x, z) \cdot P(z)$	konvergent	\ominus
Δ	$P(x) \cdot P(y x) \cdot P(z x, y)$	saturiert	\oplus
$\triangleleft\triangleright$	$P(x) \cdot P(y x) \cdot P(z x, y) \cdot P(w y, z)$	3-3-Cliquen	\oplus
$\triangleleft\triangleright$	$P(x) \cdot P(y x) \cdot P(w y) \cdot P(z x, y, w)$	unmoralisch!	\ominus

Korrelation, Regression und Transformation

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Berechnung bedingter Wahrscheinlichkeiten

Verbundverteilung

Gemeinsame Verteilung $P(x_1, \dots, x_n)$ aller Variablen in **Produktform**.

Randverteilungen

Gemeinsame Verteilung für eine Teilmenge $A \subset V$:

$$P(V \setminus \{x_i\}) = \sum_{x_i} P(x_1, \dots, x_n)$$

$$P(V \setminus \{x_i, x_j\}) = \sum_{x_i} \sum_{x_j} P(x_1, \dots, x_n)$$

$$P(V \setminus \{x_{i_1}, \dots, x_{i_m}\}) = \sum_{x_{i_1}} \dots \sum_{x_{i_m}} P(x_1, \dots, x_n)$$

Bedingte Verteilungen

Einfluß einer Zufallsvariablen \mathbb{X}_j auf eine andere \mathbb{X}_i :

$$P(x_i|x_j) = \frac{P(x_i, x_j)}{P(x_j)} = \frac{\sum \dots \sum P(x_1, \dots, x_n)}{\sum \sum \dots \sum P(x_1, \dots, x_n)}$$

Warum Bayesnetze ?

Weil sie in Wahrscheinlichkeiten faktorisieren !

Was ist Inferenz ?

Logik Axiome, Schlußregeln \Rightarrow neue Sätze

Arithmetik Parameterwerte, Operationen \Rightarrow Funktionswerte

Stochastik Observablen, W-Modell \Rightarrow bedingte W'keiten

A posteriori Verteilungen

Evidenz $E = \{e_1, \dots, e_m\}$ („instanzierte“ Variablen)

$$P(x_i|E) = P(x_i = \xi | e_1 = \eta_1, \dots, e_m = \eta_m)$$

Rand- und Rückschlußverteilungen sind aufwendig zu berechnen!

- Eliminiere Variablen in ökonomischer Reihenfolge — gemäß Dependenzstruktur bzw. Modellformel.
- Propagationsalgorithmen, Marker-Passing, Sampling ...

Notation der Rechengrößen

für baumförmige Bayesnetze

Wahrscheinlichkeitsparametermatrix

Jeder Knoten y im DAG hat **genau einen** Elterknoten x .

$$\begin{aligned} \mathbf{M}_{y|x} &= P(y|x) = [P(y = \eta_j | x = \xi_i)]_{ij} \\ &= \begin{Bmatrix} P(y = \eta_1 | x = \xi_1) & \cdots & P(y = \eta_k | x = \xi_1) \\ \vdots & & \vdots \\ P(y = \eta_1 | x = \xi_m) & \cdots & P(y = \eta_k | x = \xi_m) \end{Bmatrix} \end{aligned}$$

Evidenz

Instanzierte Variablen $e \in V$ bzw. $E \subseteq V$.

Belief-Funktion

Subjektive Einschätzung von x auf Grundlage von E (Wahr'keitsfeld):

$$\text{bel}(x) \stackrel{\text{def}}{=} P(x|E)$$

$$\text{bel}(x) = P(x | z = \zeta) = (P(x = \xi_1 | z = \zeta), \dots, P(x = \xi_\ell | z = \zeta))^\top$$

Unidirektionale Fortpflanzung in Ketten

Zwei Variablen

Beispiel: $x \rightarrow y$, Evidenz $\{y = \eta\}$

Nach der Bayesformel gilt:

$$\text{bel}(x) = P(x | y = \eta) = \frac{P(x) \cdot P(y = \eta | x)}{P(y = \eta)} \propto P(x) \cdot \lambda(x)$$

mit der a priori Wahrsch'keit $P(x)$ und dem **diagnostischen Vektor**

$$\lambda(x) = P(y = \eta | x) \quad (\eta\text{-te Spalte der Matrix } \mathbf{M}_{y|x}).$$

$P(x) \cdot \lambda(x)$ bezeichnet das komponentenweise Produkt.

Unidirektionale Fortpflanzung in Ketten

Drei Variablen

Beispiel: $x \rightarrow y \rightarrow z$, Evidenz $\{z = \zeta\}$

Nach der Bayesformel gilt wiederum:

$$\text{bel}(x) = P(x | z = \zeta) = \frac{P(x) \cdot P(z = \zeta | x)}{P(z = \zeta)} \propto P(x) \cdot \lambda(x)$$

Der **diagnostische Vektor** lautet nunmehr

$$\begin{aligned} \lambda(x) &= P(z = \zeta | x) = \sum_y P(z = \zeta, y | x) \\ &= \sum_y P(z = \zeta | y) \cdot P(y | x) \\ &= \mathbf{M}_{y|x} \bullet \lambda(y) \end{aligned}$$

$\mathbf{M}_{y|x} \bullet \lambda(y)$ bezeichnet das Vektor-Matrix-Produkt über die Variable y .

Unidirektionale Fortpflanzung in Ketten

Mehr als drei Variablen

Beispiel: $x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_n$, Evidenz $\{x_n = \xi\}$

Nach der Bayesformel gilt wiederum:

$$\text{bel}(x_1) = P(x_1 | x_n = \xi) = \frac{P(x_1) \cdot P(x_n = \xi | x_1)}{P(x_n = \xi)} \propto P(x_1) \cdot \lambda(x_1)$$

Der **diagnostische Vektor** gehorcht der Rekursion:

$$\begin{aligned} \lambda(x_1) &= M_{x_2|x_1} \bullet \lambda(x_2) \\ &= M_{x_2|x_1} \bullet M_{x_3|x_2} \bullet \lambda(x_3) \\ &= M_{x_2|x_1} \bullet M_{x_3|x_2} \bullet M_{x_4|x_3} \bullet \lambda(x_4) \\ &= M_{x_2|x_1} \bullet M_{x_3|x_2} \bullet \dots \bullet M_{x_{n-1}|x_{n-2}} \bullet \underbrace{P(x_n = \xi | x_{n-1})}_{M_{\xi|x_{n-1}}} \end{aligned}$$

Bidirektionale Fortpflanzung in Ketten

Beispiel: $e^+ \rightarrow v \rightarrow w \rightarrow x \rightarrow y \rightarrow z \rightarrow e^-$

A posteriori Wahrscheinlichkeiten nach Bayesformel:

$$\begin{aligned} \text{bel}(x) &= P(x | e^+, e^-) \propto P(e^- | x, e^+) \cdot P(x | e^+) \\ &= P(e^- | x) \cdot P(x | e^+) = \lambda(x) \cdot \pi(x) \end{aligned}$$

Diagnostische Evidenz

Kausale Evidenz

$$\lambda(x) = P(e^- | x)$$

$$\pi(x) = P(x | e^+)$$

Fortpflanzung rückwärts

$$\begin{aligned} \pi(x) &= P(x | e^+) \\ &= \sum_w P(x | w, e^+) \cdot P(w | e^+) \\ &= \sum_w P(x | w) \cdot P(w | e^+) \\ &= \pi(w) \bullet M_{x|w} \end{aligned}$$

Fortpflanzung vorwärts

$$\begin{aligned} \lambda(x) &= P(e^- | x) \\ &= \sum_y P(e^- | y, x) \\ &= \sum_y P(e^- | y) \cdot P(y | x) \\ &= M_{y|x} \bullet \lambda(y) \end{aligned}$$

Bidirektionale Fortpflanzung in Bäumen

Zerlegung der Belief-Funktion

Zerlegung der Evidenz

Für $x \in V$ unterscheiden wir zwei Quellgebiete:

$$E = E_x^+ \uplus E_x^- \quad \text{mit} \quad \begin{cases} E_x^- \subset \text{off}(x) & \text{"flussabwärts"} \\ E_x^+ \subset V \setminus \text{off}(x) & \text{"flussaufwärts"} \end{cases}$$

Belief-Funktion

Nach Kettenregel und $\text{sep}_\delta(E_x^- | \{x\} | E_x^+)$ folgt:

$$\begin{aligned} \text{bel}(x) &= P(x | E_x^+, E_x^-) \\ &\propto P(E_x^-, x | E_x^+) \\ &= P(E_x^- | x, E_x^+) \cdot P(x | E_x^+) = \lambda(x) \cdot \pi(x) \end{aligned}$$

$\pi(x) = \text{kausale}$ Unterstützung von x durch die Vorgänger

$\lambda(x) = \text{diagnostische}$ Unterstützung von x durch die Nachfolger

Bidirektionale Fortpflanzung in Bäumen

Zerlegung der Evidenz

Vertikale Zerlegung
kausal/diagnostisch:

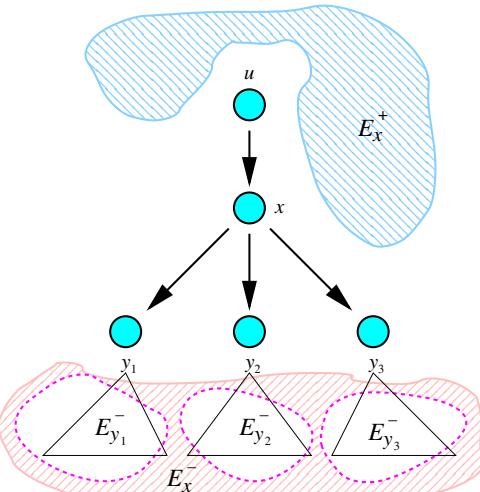
$$E_x = E_x^+ \uplus E_x^-$$

Horizontale Zerlegung
des diagnostischen Teils

$$E_x^- = \biguplus_{y_\ell \in \text{ch}(x)} E_{y_\ell}^-$$

Horizontale Zerlegung
des kausalen Teils

$$E_{y_\ell}^+ = E_x^+ \uplus \biguplus_{\kappa \neq \ell} E_{y_\kappa}^-$$



Bidirektionale Fortpflanzung in Bäumen

Diagnostische und prädiktive Wahrscheinlichkeiten

Diagnostische Komponente

Seien u_1, \dots, u_r die Nachfolger von x :

$$\lambda(x) = P(E_x^- | x) = P(E_{u_1}^-, \dots, E_{u_r}^- | x) = \prod_{s=1}^r \underbrace{P(E_{u_s}^- | x)}_{\lambda_{u_s}(x)}$$

Falls $\{x = \xi\}$ selbst instanziert, so erzeuge Dummyknoten d mit $\lambda_d(x) = \mathbf{I}_{x=\xi}$.

Prädiktive Komponente

Sei $u \in V$ der Vater (die Mutter) von x :

$$\begin{aligned} \pi(x) &= P(x | E_x^+) = \sum_u P(x, u | E_x^+) \\ &= \sum_u P(x|u) \cdot P(u|E_x^+) =: M_{x|u} \bullet \pi_x(u) \end{aligned}$$

Bidirektionale Fortpflanzung in Bäumen

Variablenversetzte diagnostische und prädiktive Komponenten

Berechnung von $\lambda_x(u)$

für $u \rightarrow x$

$$\begin{aligned} \lambda_x(u) &= \sum_x P(E_x^- | u, x) \cdot P(x|u) \\ &= \sum_x P(E_x^- | x) \cdot P(x|u) \\ &= \sum_x \lambda(x) \cdot P(x|u) \\ &= M_{x|u} \bullet \lambda(x) \end{aligned}$$

Berechnung von $\pi_y(x)$

für $u \rightarrow x$ und $y \leftarrow x \rightarrow z$

$$\begin{aligned} \pi_y(x) &= P(x | E_y^+) \\ &= P(x | E_x^+, E_z^-) \\ &\propto P(E_z^- | x, E_x^+) \cdot P(x | E_x^+) \\ &= \lambda_z(x) \cdot \pi(x) \\ &= \lambda_z(x) \cdot M_{x|u} \bullet \pi_x(u) \end{aligned}$$

Spezialfall: $x = \xi$ evident

$$\begin{aligned} \lambda_x(u) &= P(x = \xi | u) \\ (\xi\text{-te Spalte von Matrix } M_{x|u}) \end{aligned}$$

Allgemeinfall: ≥ 3 Kinder

$$\pi_y(x) = \pi(x) \cdot \sum_{z \neq y} \lambda_z(x)$$

Vorwärts-Rückwärts-Algorithmus

in baumförmigen Bayesnetzen

Start

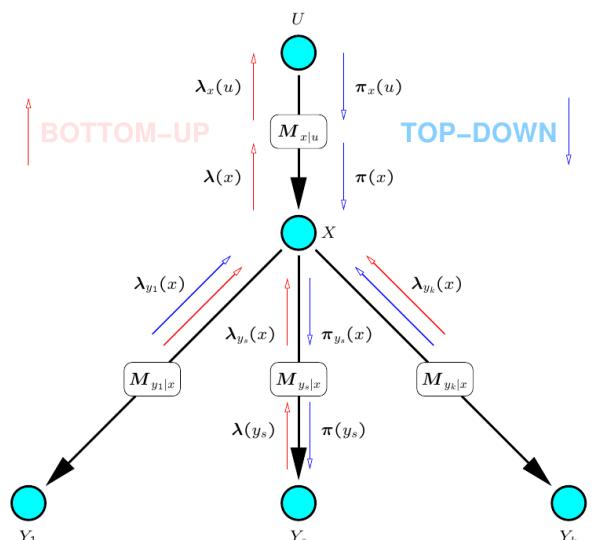
$$\begin{aligned} \pi(x_0) &= M_{x_0} \cdot && \text{(Wurzel)} \\ \lambda(x_\ell) &= 1 && \text{(Blatt)} \\ \lambda(x_e) &= e^{(\zeta)} && \text{(Evidenz)} \end{aligned}$$

Bottom-up

$$\begin{aligned} \lambda(y_k) &&& \text{(I.V.)} \\ \lambda_{y_k}(x) &= M_{y_k|x} \bullet \lambda(y_k) \\ \lambda(x) &= \prod_k \lambda_{y_k}(x) \end{aligned}$$

Top-down

$$\begin{aligned} \pi(x) &&& \text{(I.V.)} \\ \pi_{y_k}(x) &\propto \pi(x) \cdot \prod_{\ell \neq k} \lambda_{y_\ell}(x) \\ \pi(y_k) &= M_{y_k|x} \bullet \pi_{y_k}(x) \end{aligned}$$



Inferenz in moralischen Bayesnetzen

Vorwärts-Rückwärts-Algorithmus über Variablenkomplexen

Algorithmus

- ENTFERNE ALLE KANTENRICHTUNGEN
↔ äquivalentes kordales Markovnetz

Verbundbaum:

- BILDE VETRÄGLICHEN VERBUNDBAUM mit Cliquensequenz $C = \{C_1, \dots, C_K\}$

Moralische Bayesnetze

- KONSTRUIERE VARIABLENkomplexe $\mathbb{Y}_k := \bigotimes_{x_j \in B_k} \mathbb{X}_j$ mit $B_k := C_k \setminus C_{k-1}$

Markovnetz:

- EXPANDIERE VERTEILUNGSPARAMETER $M_{k|\ell} = (P(\mathbb{Y}_k = \eta | \mathbb{Y}_\ell = \zeta) | \eta \in \mathcal{Y}_k, \zeta \in \mathcal{Y}_\ell)$

Unmoralische Bayesnetze

- EXEKUTIERE VR-ALGORITHMUS AUF VB

Schummeln:

nur Imputation!

Monte Carlo:

Iteratives Auswürfeln und Neuschätzen

Spezialfall Imputation

Vorhersage eines Attributwertes aus allen anderen

Belief-Funktion

mit Zielvariable x_k und Evidenzvariablen $E = V \setminus \{x_k\}$:

$$\text{bel}(x_k)_{\xi} = P(x_k = \xi | \mathbf{x}_E) = \frac{P(x_k = \xi, \mathbf{x}_E)}{P(\mathbf{x}_E)} = \frac{P(\mathbf{x}_{|x_k=\xi})}{P(\mathbf{x}_E)}$$

\mathbb{X}_k ist diskretes Attribut

1. Für alle $\xi_\ell \in \mathcal{X}_k$ berechne $q_\ell = P(\mathbf{x}_{|x_k=\xi_\ell})$ mit

$$\mathbf{x}_{|x_k=\xi_\ell} = (x_1, \dots, x_{k-1}, \mathbf{x}_k = \xi_\ell, x_{k+1}, \dots, x_n)^\top \in \mathbb{R}^n.$$

2. Setze $\text{bel}(x_k)_\ell = q_\ell / \sum_i q_i$.

\mathbb{X}_k ist stetiges Attribut

Effiziente Lösungsmöglichkeit trotz $|\mathcal{X}_k| = \infty$?

Beweis.

Wir definieren die *lograt*-Funktion $\ell(x, y) = -2 \cdot \log(g(x)/g(y))$ und folgern die Identität

$$\ell(x, y) = \frac{1}{\sigma^2} \cdot (x^2 - y^2 - 2\mu \cdot (x - y)).$$

Wir definieren nun die Differentiale

$$\ell_h^-(x) \stackrel{\text{def}}{=} \ell(x, x - h) = \frac{1}{\sigma^2} \cdot (+2hx - h^2 - 2h\mu)$$

$$\ell_h^+(x) \stackrel{\text{def}}{=} \ell(x, x + h) = \frac{1}{\sigma^2} \cdot (-2hx - h^2 + 2h\mu)$$

für $h > 0$ und finden nach deren Addition einen Lösungsausdruck

$$\hat{\sigma}^2 = -2 \cdot \frac{h^2}{\ell_h^+(x) + \ell_h^-(x)}$$

für die gesuchte Varianz. Anschließend können wir aus jeder der Differentialformeln den Erwartungswert berechnen, z.B.:

$$\hat{\mu} = \frac{\hat{\sigma}^2 \cdot \ell_h^+(x) + 2hx + h^2}{2h} = \frac{\hat{\sigma}^2}{2h} \cdot \ell_h^+(x) + x + \frac{h}{2}$$

Schließlich bestimmen wir noch den Skalierungsfaktor c ; die numerisch stabilste Methode besteht in einer weiteren Auswertung der Geheimfunktion $g(\cdot)$, und zwar am Dichtegipfel:

$$\hat{c} = \frac{g(x)}{\mathcal{N}(x | \hat{\mu}, \hat{\sigma}^2)} = \frac{g(\hat{\mu})}{\mathcal{N}(\hat{\mu} | \hat{\mu}, \hat{\sigma}^2)} = \frac{g(\hat{\mu})}{\mathcal{N}(0 | 0, \hat{\sigma}^2)} = \sqrt{2\pi} \cdot \hat{\sigma} \cdot g(\hat{\mu})$$

Spezialfall Imputation

Vorhersage eines normalverteilten Attributwertes

Die Geheimfunktion

Es ist $P(x_k = \xi | \mathbf{x}_E) = \mathcal{N}(\xi | \mu, \sigma^2)$, also erhalten wir Resultate der Form

$$P(\mathbf{x}_{|x_k=\xi}) = P(\mathbf{x}_E) \cdot P(\xi | \mathbf{x}_E) = \underbrace{c \cdot \mathcal{N}(\xi | \mu, \sigma^2)}_{g_{c,\mu,\sigma}(\xi)}$$

durch Auswertung des Bayesnetzes an der Stelle $\xi \in \mathbb{R}$.

Lemma

Die unbekannten Parameter $c > 0$, $\mu \in \mathbb{R}$ und $\sigma > 0$ der skalierten univariaten Gaußdichte

$$g_{c,\mu,\sigma}(\xi) \stackrel{\text{def}}{=} c \cdot \mathcal{N}(\xi | \mu, \sigma^2)$$

können aus den Funktionswerten von $g(\cdot)$ an vier reellen Stützstellen bestimmt werden.

Diese Entschlüsselungstechnik lässt sich auf **multivariate** Gaußdichten verallgemeinern.

Imputation in (nichtkordalen) Markovnetzen

Effiziente Berechnung als $\text{bd}(x_k)$ -ausgedünntes Cliquenprodukt

Bedingte Wahrscheinlichkeit nach Faktorisierung

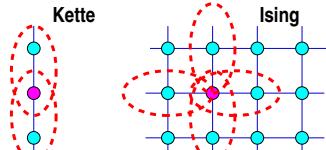
$$P(\mathbb{X}_k = x_k | \mathbb{X}_{V \setminus k} = \mathbf{x}') = \frac{P(x_k, \mathbf{x}')}{P(\mathbf{x}')} = \frac{\prod_{C \in \mathcal{C}} \phi_C(x_k, \mathbf{x}')}{\sum_{\xi \in \mathcal{X}_k} \prod_{C \in \mathcal{C}} \phi_C(x_k, \mathbf{x}')}$$

Ausklammern & Kürzen aller Gibbspotenziale ϕ_C mit $x_k \notin C$

Reduzierte Faktorisierung über $\mathcal{C}_{(k)} := \{C \mid x_k \in C\} = \{C \mid C \subseteq \text{cl}(x_k)\}$ (wegen $\mathfrak{S}(x_k | \text{bd}(x_k) | \text{"Rest"})$ nicht ganz unerwartet!)

Binäres Zielattribut $|\mathcal{X}_k| = 2$

$$\log \text{odds}(\mathbf{x}') = \log \frac{P(1 | \mathbf{x}')}{P(0 | \mathbf{x}')} = \sum_{C \in \mathcal{C}_{(k)}} \log \frac{\phi_C(x_{C \setminus k}, 1)}{\phi_C(x_{C \setminus k}, 0)}$$



Korrelation, Regression und Transformation

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Diskrete loglineare Modelle

Spezialfall: drei Variablen ($N = 3$)

Dreiwegetabellen

Drei diskrete Zufallsvariablen $V = \{\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3\} = \{a, b, c\}$

- Endliche Wertebereiche $\mathcal{X}_a, \mathcal{X}_b, \mathcal{X}_c$
- Endlich viele Zellen $(j, k, l) \in \mathcal{X}_a \times \mathcal{X}_b \times \mathcal{X}_c$
- Würfel $\{p_{jkl}\}$ von Wahrscheinlichkeiten
- Würfel $\{n_{jkl}\}$ von (absoluten) Häufigkeiten

$$\sum_j \sum_k \sum_l p_{jkl} = 1$$

$$\sum_j \sum_k \sum_l n_{jkl} = T$$

Loglineares Verteilungsmodell

Produktform

$$p_{jkl} = \prod_{A \in \Delta} \underbrace{\phi_A(\mathbf{x}_A)}_{z_{jkl}^A}$$

Summenform

$$\log p_{jkl} = \sum_{A \in \Delta} \underbrace{\log \phi_A(\mathbf{x}_A)}_{u_{jkl}^A}$$

$$\Delta \subset \mathfrak{P}V$$

MLS $\hat{=}$ relative Ereignishäufigkeiten

Happy End — für alle kausalen Verteilungen

Zerlegbare Loglinearmodelle

Cliquen $\mathcal{C} = \{C_1, \dots, C_M\}$

$$\begin{aligned} P(\mathbf{x}) &= \prod_{C \in \mathcal{C}} z_x^C \\ &= \prod_{C \in \mathcal{C}} \frac{P(\mathbf{x}_C)}{P(\mathbf{x}_{C \cap \pi(C)})} \end{aligned}$$

Bayesnetze

Ordnung $V = \{V_1, \dots, V_N\}$

$$\begin{aligned} P(\mathbf{x}) &= \prod_{n=1}^N P(x_n | \mathbf{x}_{\text{pa}(n)}) \\ &= \prod_{n=1}^N M_{x_n | \text{pa}(n)}(\mathbf{x}) \end{aligned}$$

Maximum-Likelihood

$$\hat{z}_x^C = \frac{n_{\mathbf{x}_C}}{n_{\mathbf{x}_{C \cap \pi(C)}}}$$

$$\hat{M}_{x_n | \text{pa}(n)}(\mathbf{x}) = \frac{n_{\mathbf{x}_{\{x_n\} \cup \text{pa}(n)}}}{n_{\mathbf{x}_{\text{pa}(n)}}}$$

Maximum-Likelihood

Kovarianzselektion

Beispiele — Dreiwegemodelle

Menge der (maximalen) Interaktionsterme · „Generatoren“

Unabhängiges Modell

 a, b, c

$$\log p_{jkl} = u + u_j^a + u_k^b + u_l^c$$

$$p_{jkl} = P(a = \alpha_j) \cdot P(b = \beta_k) \cdot P(c = \gamma_l) = p_{j..} \cdot p_{k..} \cdot p_{l..}$$

Kettenförmiges Modell

 ab, ac

$$\log p_{jkl} = u + u_j^a + u_k^b + u_l^c + u_{jk}^{ab} + u_{jl}^{ac}$$

$$p_{jkl} = \frac{p_{jk} \cdot p_{j..}}{p_{j..}} \quad \text{bzw.} \quad \frac{p_{jk} \cdot p_{j..}}{p_{j..}} = \frac{p_{jk}}{p_{j..}} \cdot \frac{p_{j..}}{p_{j..}}$$

Saturiertes Modell

 abc

$$\log p_{jkl} = u + u_j^a + u_k^b + u_l^c + u_{jk}^{ab} + u_{jl}^{ac} + u_{kl}^{bc} + u_{jkl}^{abc}$$

Schätzung der kanonischen Modellparameter

Normierungseigenschaft

$$1 \stackrel{!}{=} \sum_{jkl} p_{jkl} = \sum_{jkl} \exp \left\{ \sum_{A \in \Delta} u_{jkl}^A \right\} = e^u \cdot \sum_{jkl} \exp \left\{ \sum_{A \neq \emptyset} u_{jkl}^A \right\}$$

Multinomial gezogene Stichprobe

$$P(\mathbf{n}|\mathbf{p}) = P(\{n_{jkl}\} | \{p_{jkl}\}) = \frac{T!}{\prod_{j,k,l} n_{jkl}!} \cdot \prod_{j,k,l} p_{jkl}^{n_{jkl}}$$

Logarithmierte Likelihood-Funktion

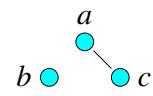
$$\ell_{\text{ML}}(\mathbf{n}|\mathbf{p}) = \log \frac{T!}{\prod_{j,k,l} n_{jkl}!} + \sum_{j,k,l} n_{jkl} \log p_{jkl}$$

Maximum-Likelihood-Schätzwerte

Kanonische Verteilungsparameter für das saturierte Modell

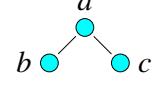
$$\hat{\mathbf{p}} = \underset{\mathbf{p}}{\operatorname{argmax}} \ell_{\text{ML}}(\mathbf{n}|\mathbf{p}) \Rightarrow \hat{p}_{jkl} = \frac{n_{jkl}}{T}$$

Beispiele — Loglinearmodelle I



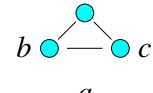
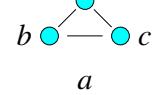
GRAPHISCH

ZERLEGBAR

 b, ac 

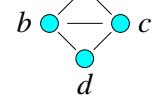
GRAPHISCH

ZERLEGBAR

 ab, ac  ab, ac, bc 

GRAPHISCH

ZERLEGBAR

 abc 

GRAPHISCH

ZERLEGBAR

 abc, bcd

Diskrete Loglinearmodelle

Definition

Die Familie diskreter Wahrscheinlichkeitsfunktionen $\{p_x\}_{x \in \Omega}$ der Gestalt

$$\log p_x = \sum_{A \in \Delta} u_x^A, \quad x \in \Omega, \quad \Delta \subseteq \mathcal{P}V$$

heißt **Loglinearmodell** mit der Menge Δ von **Interaktionstermen**.

1. Ein Loglinearmodell heißt **hierarchisch**, falls gilt:

$$A \subseteq B \quad \text{und} \quad B \in \Delta \quad \Rightarrow \quad A \in \Delta$$

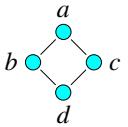
2. Ein (hierarchisches) Loglinearmodell heißt **graphisch**, wenn gilt:

$$C \in \Delta \quad \Leftrightarrow \quad \forall a, b \in C : \{a, b\} \in \Delta$$

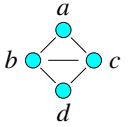
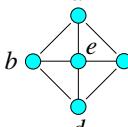
3. Ein graphisches LLM heißt **zerlegbar**, wenn sein Graph kordal ist.

Die maximalen Interaktionsterme eines hierarchischen Loglinearmodells heißen **Generatoren**. Die Generatorenmenge wird auch als **Modellformel** bezeichnet.

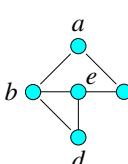
Beispiele — Loglinearmodelle II



GRAPHISCH

 ab, ac, bd, cd  ab, bcd, ac 

GRAPHISCH

 abe, ace, bde, cde 

GRAPHISCH

 ab, ac, bde, ce

Elementare und marginale Ereignisse

Häufigkeit und charakteristische Funktion

Definition

Es sei $\{n_x\}_{x \in \Omega}$ die Tafel elementarer Ereignishäufigkeiten über V . Das Zahlenfeld $\{n_{x_A}\}_{x_A \in \Omega_A}$ für eine Variablenmenge $A \subseteq V$ mit Einträgen

$$n_{x_A} \stackrel{\text{def}}{=} \sum_{x_{A'} \in \Omega_{A'}} n_x, \quad A' = V \setminus A$$

heißt **marginale Tafel** oder Tabelle für A .

Definition

Sei $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$, $A \subseteq V$ und $x_A \in \Omega_A$ ein marginales Ereignis.

Die zweiwertige Abbildung

$$\varphi_{x_A} : \begin{cases} \Omega & \rightarrow \{1, 0\} \\ \xi & \mapsto \begin{cases} 1 & \xi_j = x_j \text{ für alle } j \in A \\ 0 & \text{sonst} \end{cases} \end{cases}$$

heißt **charakteristische Funktion** von x_A .

Lernen der Loglinearparameter

GIS-Algorithmus — Generalized Iterative Scaling

Satz (Deming & Stephan, 1940)

Mit der abkürzenden Schreibweise $z_x^A = \exp(u_x^A)$ gilt:

Das Iterationsverfahren

$$z_x^A \leftarrow z_x^A \cdot \left(\frac{n_{x_A}/T}{\mathcal{E}[\varphi_{x_A}(\mathbb{X})]} \right)^{1/|\Delta|} = z_x^A \cdot \left(\frac{\sum_{y \in \Omega} \varphi_{x_A}(y) \cdot \frac{n_y}{T}}{\sum_{y \in \Omega} \varphi_{x_A}(y) \cdot \prod_{B \in \Delta} z_y^B} \right)^{1/|\Delta|}$$

mit den Startwerten $z_x^A \equiv 1$ konvergiert gegen die Maximum-Likelihood-Schätzwerte des loglinearen Modells.

Bemerkung

Die Gleichung für $\emptyset \in \Delta$ garantiert $\sum p_x = 1$.

Das Bedingungssystem ist konsistent: alle $C_y = \sum \varphi_{x_A}(y)$ sind gleich $|\Delta|$.

Beweis \rightsquigarrow Skriptum „Stochastische Grammatikmodelle“

Maximum-Entropie-Prinzip

Edwin Thompson Jaynes, 1957

Satz ($ML \hat{=} ME$)

Es sei ein hierarchisches loglineares Modell

$$\log p_x = \sum_{A \in \Delta} u_x^A$$

gegeben sowie die Häufigkeitstafel $\{n_x\}_{x \in \Omega}$ der Daten $\omega \subset \Omega$.

1. Die **Maximum-Likelihood-Parameter** $\{u_x^A\}_{x_A}$ des Modells erfüllen die Bedingungsgleichungen (*)

$$\mathcal{E}[\varphi_{x_A}(\mathbb{X}) | \mathbf{u}] = \frac{n_{x_A}}{T}, \quad A \in \Delta, x \in \Omega.$$

2. Unter allem Wahrscheinlichkeitsverteilungen, die das Gleichungssystem (*) erfüllen, hat obiges loglineare Modell mit Parametern $\{u_x^A\}_{x_A}$ die **maximale Entropie**.

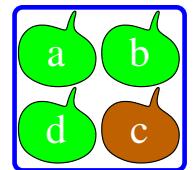
„Unter allen Zuständen eines physikalischen Systems, die kompatibel mit dem vorhandenen Wissen sind, ist der zu wählen, welcher die Entropie maximiert.“

Beispiel — Tafelobst im Tetrapack

Diamantenes Verteilungsmodell

für die vier frischen/faulen Äpfel:

$$P : \begin{cases} \{0, 1\}^4 & \rightarrow [0, 1] \\ (\alpha, \beta, \zeta, \delta) & \mapsto z \cdot z_{\alpha\beta}^{ab} \cdot z_{\beta\zeta}^{bc} \cdot z_{\zeta\delta}^{cd} \cdot z_{\delta\alpha}^{da} \end{cases}$$



Datensammlung und Statistiken

Absolute Häufigkeiten

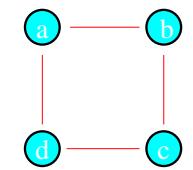
$$\{n_{\alpha\beta\zeta\delta} \mid \alpha, \beta, \zeta, \delta \in \{0, 1\}\}$$

Minimale suffiziente Statistiken, z.B. für $ab \in \Delta$:

$$n_{\alpha\beta}^{ab} = \sum_{a,b,c,d} \delta_{a=\alpha} \cdot \delta_{b=\beta} \cdot n_{abcd}$$

Symmetrie I

$$n_{\xi\eta}^{ab} = n_{\xi\eta}^{bc} = n_{\xi\eta}^{cd} = n_{\xi\eta}^{da} = n_{\xi\eta}$$



Symmetrie II

$$n_{01} = n_{10}$$

Beispiel — Tafelobst im Tetrapack

Daten = 100 Obstkörbe

	0:4	1:3	2:2	3:1	4:0	
n_{00}	60	2	1	0	0	63
n_{01}	0	1	2	1	0	4
n_{10}	0	1	2	1	0	4
n_{11}	0	0	1	2	26	29
	60	4	6	4	26	100

Iterationsanfang

$$z_{00} = z_{01} = z_{11} = 1$$

Iterationsschritt

$$z_{\xi\eta} \leftarrow z_{\xi\eta} \cdot \left(\frac{n_{\xi\eta}/100}{\mathcal{E}[\varphi_{\xi\eta}(\mathbb{X})]} \right)^{1/5}$$

Generalized Iterative Scaling

i	Loglinearparameter				Wahrscheinlichkeiten in Promille				
	z_{00}	z_{01}	z_{11}	$1/z$	(00) (00)	(10) (00)	(11) (00)	(11) (10)	(11) (11)
0	1	1	1	16	62.5	62.5	62.5	62.5	62.5
1	1.2	0.693	1.03	10.9	192	63.9	54.7	46.8	103
2	1.33	0.509	1.06	9.05	351	51.1	40.5	32.2	139
3	1.4	0.402	1.1	8.39	460	37.8	29.5	23.1	172
4	1.43	0.339	1.13	8.15	517	28.9	22.8	18	201
6	1.46	0.279	1.17	8.03	561	20.6	16.6	13.4	235
9	1.47	0.252	1.19	8.01	579	17.1	13.9	11.3	251
12	1.47	0.245	1.19	8	584	16.3	13.2	10.7	255
16	1.47	0.244	1.2	8	585	16	13	10.6	256
20	1.47	0.243	1.2	8	585	16	13	10.6	256
saturiertes Modell:				600	10	10	10	260	

Welches ist die beste Modellstruktur ?

Wahrscheinlichste Kombination aus Struktur und Parametern

Gegeben

Datenprobe ω aus der Objektmenge Ω über den Variablen V

$$\Omega = \bigotimes_{a \in V} \mathcal{X}_a$$

Gesucht

Das bestpassende graphische/kausale/kordale/loglineare Modell

$$\hat{\Delta} = \underset{\Delta \subset \mathfrak{P}V}{\operatorname{argmax}} J_\omega(\Delta) = \underset{\Delta \subset \mathfrak{P}V}{\operatorname{argmax}} \frac{f_{\text{prior}}(\Delta) \cdot P(\omega | \Delta)}{P(\omega)}$$

$$P(\omega | \Delta) = \sum_{\theta \in \mathcal{M}(\Delta)} f_{\text{prior}}(\theta | \Delta) \cdot P(\omega | \Delta, \theta)$$

Markovnetze

$\binom{N}{2}$ Kanten & insgesamt
 $2^{\binom{N}{2}}$ ungerichtete Graphen

Bayesnetze

$N!$ Ordnungen & jeweils
 $2^{\binom{N}{2}}$ zyklenfreie Graphen

Loglinear

2^N Terme
 2^{2^N} Modelle

Korrelation, Regression und Transformation

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Gütemaße für Modellstrukturen

Definition

Mit den **Maximum-Likelihood-Parametern**

$$\hat{\theta}_\Delta(\omega) = \underset{\theta}{\operatorname{argmax}} \ell_\omega(\Delta, \theta) = \underset{\theta}{\operatorname{argmax}} \sum_{x \in \omega} \log P(x | \Delta, \theta)$$

und der ML-bezogenen Bewertung $\hat{\ell}_\omega(\Delta) = \ell_\omega(\Delta, \hat{\theta}_\Delta(\omega))$ heißt

$$\text{dev}(\Delta) \stackrel{\text{def}}{=} 2 \cdot (\hat{\ell}(\mathfrak{P}V) - \hat{\ell}(\Delta)), \quad \mathfrak{P}V = \{V\} = \text{saturiertes Modell}$$

die **Devianz** der Modellstruktur Δ für die Daten ω .

Lemma

Das Devianzmaß besitzt die folgenden Eigenschaften:

1. Gilt $\omega \sim P(\cdot | \Delta)$, so ist die Devianz asymptotisch χ_d^2 -verteilt, wobei d die Differenz der Freiheitsgrade von Δ und saturiertem Modell bezeichnet.
2. Es gilt $\text{dev}(\mathfrak{P}V) = 0$ und $\mathcal{E}[\text{dev}(\Delta)] = d$.

Einige regularisierte Gütemaße

Kreuzvalidierung

Datenpartition $\omega = \omega_a \uplus \omega_b$

$$J(\Delta) = \ell_{\omega_b}(\Delta, \hat{\theta}_\Delta(\omega_a))$$

ML-Bewertung + Strafterm

$$J(\Delta) = \hat{\ell}_\omega(\Delta) - \psi(N) \cdot |\theta_\Delta|$$

Entropie

Bedingte Entropien $H(\mathbb{X}_n | \mathbf{x}) = - \sum_{\xi \in \mathcal{X}_n} P(\xi | \mathbf{x}) \cdot \log P(\xi | \mathbf{x})$

$$J(\Delta) = H(\Delta) = \sum_{n=1}^N \sum_{x \in \mathcal{X}_{\text{pa}(n)}} P(x) \cdot H(\mathbb{X}_n | \mathbb{X}_{\text{pa}(n)} = x)$$

Rotationsvalidierung ($L^1 O$)

„leave-one-out“ $\omega^{(x)} = \omega \setminus \{\mathbf{x}\}$

$$J(\Delta) = \sum_{x \in \omega} \ell_{\{\mathbf{x}\}}(\Delta, \hat{\theta}_\Delta(\omega^{(x)}))$$

AIC $\Rightarrow \psi(N) \equiv 1$

„Akaike Information Criterion“

BIC $\Rightarrow \psi(N) = \frac{1}{2} \log N$

„Bayesian Information Criterion“

Die K2-Metrik für Bayesnetze

Cooper & Herskovitz, 1991

Fakt

Eine perfekte Gütfunktion wäre die *a posteriori* Wahrscheinlichkeit $P(\Delta | \omega)$ der Modellstruktur auf Basis der Datenprobe.

Gleich- und Dirichletverteilungsannahme

für Bayesnetzstruktur Δ und -parameter $M_{n|\text{pa}(n)}$:

$$P(\Delta | \omega) \propto P(\omega | \Delta) = \underbrace{\int \mathcal{D}(\theta | \Psi) \cdot P(\omega | \theta, \Delta) d\theta}_{P(\omega^{(\Psi)} | \theta, \Delta)}$$

K2-Metrik

Geschlossene Darstellung der *a posteriori* Wahrscheinlichkeit:

$$J(\Delta) = \prod_{n=1}^N \prod_{x \in \mathcal{X}_{\text{pa}(n)}} \frac{(L_n - 1)!}{(n_x^{\text{pa}(n)} + L_n - 1)!} \cdot \prod_{\xi \in \mathcal{X}_n} n_{x, \xi}^{\text{pa}(n), \{\mathbf{n}\}}$$

Suchverfahren

Wer findet die Stecknadel im Heuhaufen vor Anbruch des jüngsten Tages?

Ungerichtete Graphen — Markovnetze

Gesucht ist eine J-optimale Teilmenge von

$$\mathfrak{E}_V = \{\{a, b\} \mid a, b \in V, a \neq b\}$$

\Rightarrow Jedes $\mathcal{E} \subseteq \mathfrak{E}_V$ ist „erlaubt“!

Kombinatorische Merkmalauswahl

Alle „wrapper“-Verfahren sind sinngemäß anwendbar:

- $\begin{cases} \text{backward} \\ \text{forward} \end{cases}$ selection: sukzessiv Kanten $\begin{cases} \text{entfernen} \\ \text{einfügen} \end{cases}$
- $\begin{cases} \text{pulsierende Suche} \\ (\text{Gedächtnis}) \end{cases}$ $\begin{cases} \text{geordnete Suche} \\ (> 1 \text{ Kandidaten}) \end{cases}$ $\begin{cases} \text{evolutionäre Suche} \\ (\text{Populationen}) \end{cases}$

SFS — Sequential Forward Selection

Gierige bottom-up Suche (Whitney 1971 · Buntine 1991)

1 INITIALISIERUNG

$$\mathcal{G} = (V, \emptyset)$$

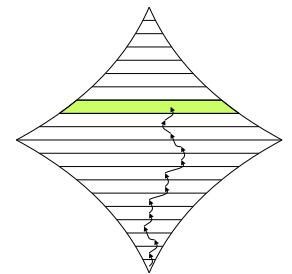
2 AUSWAHL

einer nützlichsten neuen Kante

$$\mathfrak{e}^* = \operatorname{argmax} \{J(E, \mathfrak{e}) \mid \mathfrak{e} \in \mathfrak{E}_V \setminus E\}$$

3 TERMINIERUNG

Wenn $J(E, \mathfrak{e}^*) \leq J(E)$
dann \rightsquigarrow ENDE
sonst $E \leftarrow E \cup \{\mathfrak{e}^*\}$ und \rightsquigarrow 2.



Bemerkung
SFS trifft voreilige Entscheidungen (Horizont=1) und verfehlt i.a. die Optimallösung.

$E^{(1)} \subset E^{(2)} \subset E^{(3)} \subset \dots$

SBE — Sequential Backward Elimination

Gierige top-down Suche (Marill & Green 1963 · Edwards/MIM 1995)

1 INITIALISIERUNG

$$\mathcal{G} = (V, \mathfrak{E}_V)$$

2 AUSWAHL einer nutzlosen alten Kante

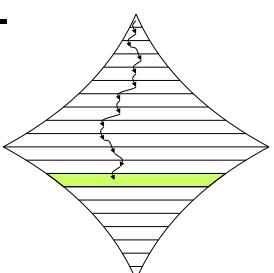
$$\epsilon^* = \operatorname{argmax} \{J(E \setminus \epsilon) | \epsilon \in E\}$$

3 TERMINIERUNG

Wenn $J(E \setminus \epsilon^*) \leq J(E)$

dann \rightsquigarrow ENDE

sonst $E \leftarrow E \setminus \{\epsilon^*\}$ und \rightsquigarrow 2.



Bemerkung
SBE findet synergistische Paare
SBE löscht redundante Kanten
SBE aufwändiger als SFS:
Start mit umfangreicheren E
Längerer Weg zum Ziel

FBS — Forward/Backward Selection

Gierige bidirektionale Suche (Wahba 1988)

1 INITIALISIERUNG

$$\mathcal{G} = (V, \emptyset)$$

2 KANTENAUSWAHL

$$\epsilon^F = \operatorname{argmax} \{J(E, \epsilon) | \epsilon \in \mathfrak{E}_V \setminus E\}$$

$$\epsilon^B = \operatorname{argmax} \{J(E \setminus \epsilon) | \epsilon \in E\}$$

3 WENN MÖGLICH LÖSCHEN

Wenn $J(E \setminus \epsilon^B) > J(E)$

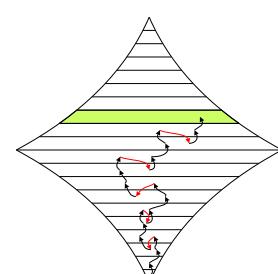
dann $E \leftarrow E \setminus \{\epsilon^B\}$ und \rightsquigarrow 2.

4 TERMINIERUNG

Wenn $J(E, \epsilon^F) > J(E)$

dann $E \leftarrow E \cup \{\epsilon^F\}$ und \rightsquigarrow 2.

Sonst \rightsquigarrow ENDE.



Bemerkung
Redundant gewordene Kanten können jetzt wieder eliminiert werden.
Bewertungsgesteuert oder immer löschen.
Extrem riskantes „Hillclimbing“

SFFS — Sequential Forward Floating Search

Pulsierende Suche mit $q = \binom{N}{2}$ Schubladen (Pudil 1994)

1 INITIALISIERUNG

$$\mathcal{G} = (V, \emptyset), \quad n = 0, \quad \iota_0 = J(\emptyset)$$

2 VORWÄRTSAUSWAHL

$$\epsilon^* = \operatorname{argmax} \{J(E, \epsilon) | \epsilon \in \mathfrak{E}_V \setminus E\}$$

Setze $E \leftarrow E \cup \{\epsilon^*\}$, $n \leftarrow n + 1$ und $\iota_n = J(E)$ und \rightsquigarrow 3.

3 RÜCKWÄRTSAUSWAHL

$$\epsilon^* = \operatorname{argmax} \{J(E \setminus \epsilon) | \epsilon \in \mathfrak{E}_V \setminus E\}$$

Wenn $J(E \setminus \epsilon^*) \leq \iota_{n-1}$ dann \rightsquigarrow 2.

Sonst setze $E \leftarrow E \setminus \{\epsilon^*\}$, $n \leftarrow n - 1$ und $\iota_n = J(E)$ und \rightsquigarrow 3.

4 TERMINIERUNG

Wenn Ziellkardinalität $n = n_0$ erreicht dann \rightsquigarrow ENDE.

Geordnete Suche

Branch & Bound (Narendra/Fukunaga 1977)

Suchraum

Zustände

Startzustand

Übergänge

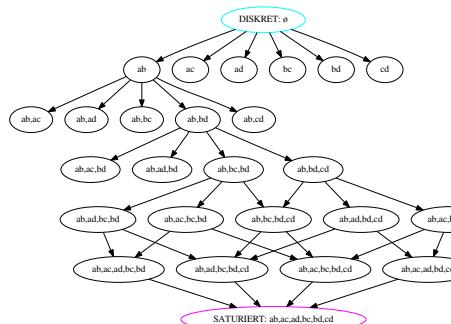
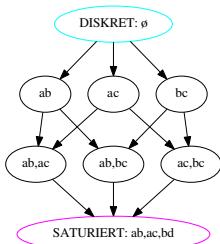
Kosten

Kantenmengen $E \subset \mathfrak{E}_V$

$$E = \emptyset$$

Zusatzkante $\{a, b\}$

Devianzabbau $\frac{\partial}{\partial a} \text{dev}(E)$



Geordnete Suche

$\text{dev}(E)$ monoton \Rightarrow Kosten ≥ 0

Branch&Bound-Algorithmus

A* mit trivialer Restschätzung

$$f(E) = g(E) + h(E)$$

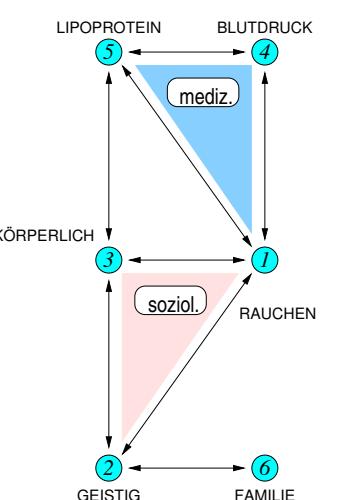
$\underbrace{g(E)}_{\text{dev}(E)} + \underbrace{h(E)}_{\equiv 0}$

Beispiel — Koronare Herzschwäche

P. Stuyvesant (1978)

Sechs binäre Attribute

- X_1 Versuchsperson ist Raucher ?
- X_2 Streß durch geistige Arbeit
- X_3 Streß durch körperliche Arbeit
- X_4 systolischer Blutdruck $\leq 140\text{mm}$
- X_5 Lipoproteinquotient $\beta/\alpha \leq 3$
- X_6 famili. Befund koronarer Schwäche



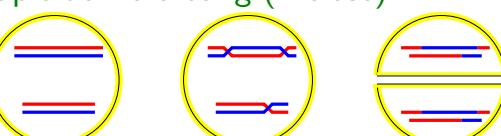
Datenerhebung

- Befundung bei $T = 1841$ Automobilarbeitern in Detroit
- Statistik mit $2^6 = 64$ Zellen
- Inkrem. Löschen partieller Abhängigkeiten
- χ^2 -Kriterium: Abweichung vom saturierten Modell
- Kordale Graphen (zerlegbare Modelle!)

Beispiel — Chromosomensequenzierung

Wir sortieren das Erbgut von „barley powdery mildew fungus“

Haploide Vererbung (Meiose)



nach der Verschmelzung Crossover-Operation nach der Zellteilung

Erhebung

- 70 Geschwisterindividuen
- 6 binäre phänotypische Attribute
- $\mathbb{X}_j^{1:1}$ unbekannter Genlocus

Hypothese über Genloci $\mathbb{X}_1, \dots, \mathbb{X}_6$

- unterschiedliche Chromosomen \rightsquigarrow unabhängig
- gleiches Chromosom \rightsquigarrow distanzabhängig korreliert
- Sequenz von Genen $g_1, g_2, g_3 \Rightarrow \mathcal{S}(g_1 | g_2 | g_3)$

Resultat

$$d \longleftrightarrow a \longleftrightarrow b \longleftrightarrow e \longleftrightarrow c \longleftrightarrow f$$

Suchverfahren

Die Stecknadel piekt jetzt nur noch auf einer Seite !

Gerichtete azyklische Graphen — Bayesnetze

Optimale Teilmenge von $\mathfrak{E}_{\{x_i\}} = \{(x_i, x_j) \mid 1 \leq i, j \leq N \text{ und } i \neq j\}$

- UG-Kantenselektion — Test auf Kordalität
- DAG-Kantenselektion — Test auf Zyklen
DAG-Kantenselektion — Test auf Zyklen und Moralität

Optimale Teilmengen von $\mathfrak{E}_{(x_i)} = \{(x_i, x_j) \mid 1 \leq i < j \leq N\}$

- Lineare Variablenordnung $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$ vorlegen
Optimale Eltermenge $B_n \subseteq V_n$ für jedes \mathbb{X}_n berechnen
(zulässiges Verfahren sofern $J(\cdot)$ in „Familienterme“ zerfällt)

Exakte Suche für eingeschränkte Netzstrukturen

- Bäume & Fallschirme
Minimaler Spannbaum (Cormen, Leiserson, Rivest 1990)

K2-Algorithmus

Elternsuchverfahren (Cooper & Herskovits, 1992)

1 INITIALISIERUNG

Eine Variablenordnung ist a priori vorzugeben:

$$V = \{x_1, \dots, x_N\}, \quad n = 2$$

2 ELTERNAUSWAHL

Triff eine Vorwärtsauswahl (SFS) bezüglich K2-Bewertung:

$$\text{pa}(x_n) = \text{argmax} \{J(A) \mid A \subseteq \{x_1, \dots, x_{n-1}\}\}$$

3 TERMINIERUNG

Wenn $n < N$ dann $n \leftarrow n + 1$ und \rightsquigarrow 2 sonst \rightsquigarrow ENDE.

Tetrad III Algorithmus

UG-Kantenselektion (Scheines 1996)

(Algorithmus)

Bayesian Network SFFS

Pulsierende DAG-Kantenselektion für Bayesnetze (Blanco & Inza, 2002)

(Algorithmus)

1 INITIALISIERUNG

$$\mathcal{G} = (V, \emptyset), \quad n = 0, \quad \iota_0 = J(\emptyset)$$

2 VORWÄRTSAUSWAHL

$$\epsilon^* = \operatorname{argmax} \{J(E, \epsilon) \mid \epsilon \in \mathfrak{E}_{\{x_i\}} \setminus E \wedge \text{DAG}(E, \epsilon)\}$$

Setze $E \leftarrow E \cup \{\epsilon^*\}$, $n \leftarrow |n| + 1$ und $\iota_n = J(E)$ und dann \rightsquigarrow 3.

3 RÜCKWÄRTSAUSWAHL

$$\epsilon^* = \operatorname{argmax} \{J(E \setminus \epsilon) \mid \epsilon \in \mathfrak{E}_{\{x_i\}} \setminus E\}$$

Wenn $J(E \setminus \epsilon^*) \leq \iota_{n-1}$ dann \rightsquigarrow 2.

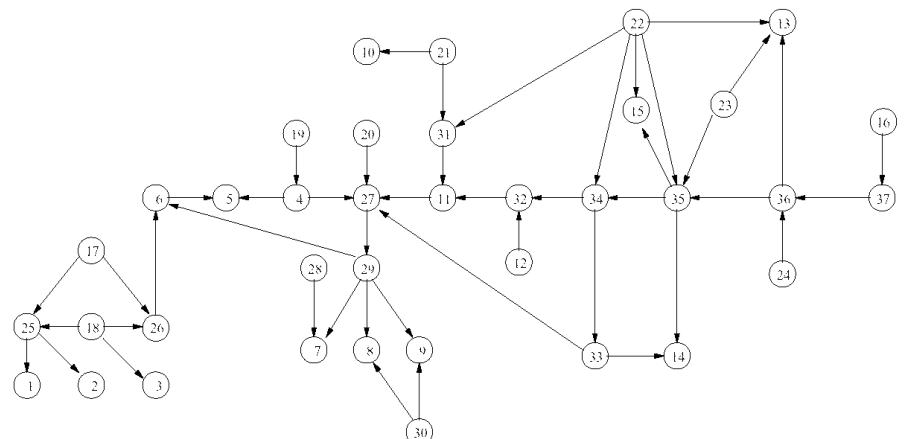
Sonst setze $E \leftarrow E \setminus \{\epsilon^*\}$, $n \leftarrow n - 1$ und $\iota_n = J(E)$ und \rightsquigarrow 3.

4 TERMINIERUNG

Wenn Zielkardinalität $n = n_0$ erreicht dann \rightsquigarrow ENDE.

Beispiel — Alarmkette

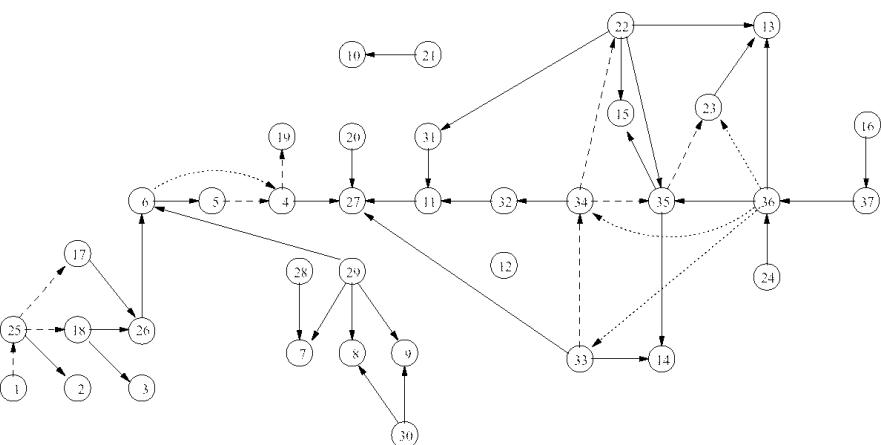
37 Attribute · 46 Kanten



Aus diesem DAG wurden 370 Lernbeispiele ausgewürfelt

Beispiel — Alarmkette

37 Attribute · nunmehr 45 statt 46 Kanten



Der gelernte DAG — eingefügte und gelöschte Kanten gestrichelt

TBN — Baumförmige Bayesnetze

Erinnerung: moralische Bayesnetze $\hat{=}$ zerlegbare Markovnetze

Modellformel für ein TBN mit Wurzel \mathbb{X}_{i_0}

$V = \{x_1, \dots, x_N\}$ und $\pi : V \setminus \{i_0\} \rightarrow V$ mit $\text{pa}(x_j) = \{x_{\pi_j}\}$ für $j \neq i_0$:

$$P(\mathbf{x}) = P(x_{i_0}) \cdot \prod_{j \neq i_0} P(x_j | x_{\pi_j}) = \prod_{i=1}^n P(x_i) \cdot \prod_{j \neq i_0} \underbrace{\frac{P(x_j, x_{\pi_j})}{P(x_j) \cdot P(x_{\pi_j})}}_{\exp(\mathfrak{S}(x_j; x_{\pi_j}))}$$

Nur die **punktweisen Transformationen** sind abhängig von der Baumstruktur!

Relevanter Anteil der logarithmierten Likelihood-Zielgröße für Lerndatenprobe ω und Baumkantenmenge $E = \{(j, \pi_j) \mid j \neq i_0\}$:

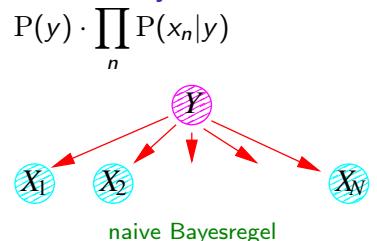
$$\ell_{\text{ML}}(\omega | E) = \sum_{(i,j) \in E} \underbrace{\{\mathcal{H}(\omega, P_{\mathbb{X}_i}) + \mathcal{H}(\omega, P_{\mathbb{X}_j}) - \mathcal{H}(\omega, P_{\mathbb{X}_i, \mathbb{X}_j})\}}_{\mathfrak{S}_{\omega}(\mathbb{X}_i; \mathbb{X}_j)}$$

↳ Berechnung aller **empirischen Transformationen**

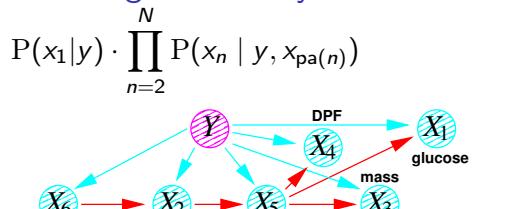
Klassifizieren mit Bayesnetzen

$\mathbb{X}_1, \dots, \mathbb{X}_N \quad \Rightarrow \quad \mathbb{Y} \in \{1, 2, \dots, K\}$

Naives Bayesnetz



Tree Augmented Bayesnet



Bayes-Multinetz



Kausalpfadanalyse mit baumförmigen Bayesnetzen

Suche nach dem minimalen Spannbaum (Chow & Liu 1968)

(Algorithmus)

1 INITIALISIERUNG

Berechne alle Transformationswerte ($i, j = 1, \dots, N$):

$$\text{TI}(\mathbb{X}_i, \mathbb{X}_j) \stackrel{\text{def}}{=} \sum_{x_i \in \mathcal{X}_i} \sum_{x_j \in \mathcal{X}_j} P(x_i, x_j) \cdot \log \frac{P(x_i, x_j)}{P(x_i) \cdot P(x_j)}$$

2 BEWERTETER GRAPH

Erzeuge $\tilde{\mathcal{G}} = (V, V^2, \beta)$ mit der Kantengewichtung

$$\beta : \begin{cases} V^2 & \rightarrow \mathbb{IR} \\ \{x_i, x_j\} & \mapsto -\text{TI}(\mathbb{X}_i, \mathbb{X}_j) \end{cases}$$

3 SPANNBAUM ($O(N^2 \log N)$) SLAC — „single-linkage agglomerative clustering“

Konstruiere den minimalen spannenden Baum $\mathcal{G} \subset \tilde{\mathcal{G}}$.

4 ORIENTIERUNG VON \mathcal{G}

Wähle eine beliebige Wurzelvariable $v_0 \in V$.

Alle Kanten von \mathcal{G} werden „wurzelauswärts“ gerichtet.

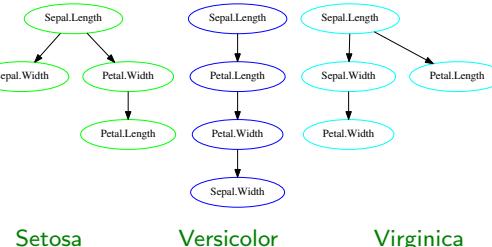
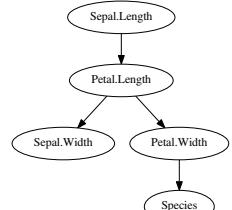
(summingA)

Beispiel — Fishers Irisdatensatz

5 Attribute ($\mathbb{IR}^4 \times \{1, 2, 3\}$) · 150 Objekte (50 je Spezies)

Transformationsmatrix

	\mathbb{X}_1	\mathbb{X}_2	\mathbb{X}_3	\mathbb{X}_4	\mathbb{X}_5
ℓ_{sepal}	\mathbb{X}_1	0	1.33	2.04	1.88
w_{sepal}	\mathbb{X}_2	1.33	0	1.43	1.40
ℓ_{petal}	\mathbb{X}_3	2.04	1.43	0	2.64
w_{petal}	\mathbb{X}_4	1.88	1.40	2.64	0
species	\mathbb{X}_5	0.69	0.38	1.37	1.43

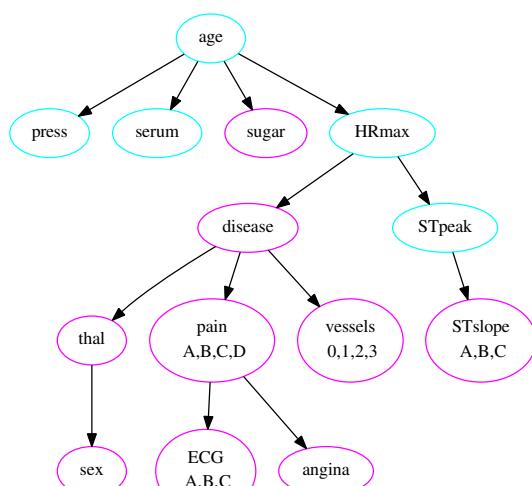


Bayesbaum
für alle fünf Attribute

Bayeswald
ein Baum je Spezies

Beispiel — Statlog Herzdatensammlung

13 Attribute · 270 Objekte · Klassifikation: „disease“



$$\mathbb{S}(\mathbf{X}_{14} \mid \mathbf{X}_3, \mathbf{X}_8, \mathbf{X}_{12}, \mathbf{X}_{13} \mid \text{„Rest“})$$

- 1. age (IR)
- 2. sex {male, female}
- 3. chest pain {A, B, C, D}
- 4. blood pressure (IR)
- 5. serum cholestral (IR)
- 6. fasting blood sugar {T, F}
- 7. resting ECG results [0 : 2]
- 8. maximum heart rate achieved (IR)
- 9. exercise induced angina {T, F}
- 10. ST depression (exercise:rest) (IR)
- 11. slope of peak exercise ST {A, B, C}
- 12. vessels colored by fluroscopy [0 : 3]
- 13. thal {normal, fixed, defect}
- 14. heart disease {T, F}

Korrelation, Regression und Transformation

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Stetige Loglinearmodelle

Motivation: multivariate Normalverteilungsdichte

Definition

Es sei die N -dimensionale multivariate Normalverteilungsdichte

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{S}) = |2\pi\mathbf{S}|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

mit dem Mittelwertvektor $\boldsymbol{\mu}$ und der Kovarianzmatrix \mathbf{S} gegeben.

Die Werte $\alpha, \beta_i, \kappa_{ij}$ der Darstellung

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{S}) = \exp\left(\alpha + \sum_i \beta_i \cdot x_i + \sum_{i,j} \kappa_{ij} \cdot x_i x_j\right)$$

heißen **kanonische Parameter** der exponentiellen Familie; die Matrix $\mathbf{K} = [\kappa_{ij}]$ heißt **Konzentrationsmatrix** oder **Präzisionsmatrix**.

Kanonische Parameter & Standardparameter

$$\alpha = -\frac{1}{2} \cdot (\log |2\pi\mathbf{S}| + \boldsymbol{\mu}^\top \mathbf{S}^{-1} \boldsymbol{\mu}), \quad \boldsymbol{\beta} = \mathbf{S}^{-1} \boldsymbol{\mu}, \quad \mathbf{K} = -\frac{1}{2} \cdot \mathbf{S}^{-1}.$$

Definition

Sei $\Delta \subset \mathbb{N}^N$ eine (endliche) Menge von Exponenten- N -Tupeln.
Die Familie stetiger Wahrscheinlichkeitsdichtefunktionen der Gestalt

$$\log f_\Delta(\mathbf{x}) = \sum_{i \in \Delta} u^i \cdot \prod_{n=1}^N x_n^{i_n}, \quad \mathbf{x} \in \mathbb{R}^N$$

heißt **stetiges Loglinearmodell** über V mit der Menge Δ von **Kovarianztermen**.

Lemma

Für Loglinearmodelle Δ , die Normalverteilungen sind, gilt:

$$\Delta \text{ hieratisch} \Rightarrow \Delta \text{ graphisch}$$

Wir nennen diese Familien **Gaußsche Graphische Modelle** oder **Kovarianzselektionsmodelle**.

Loglinearmodelle

Die Kovarianzterme in Δ sind die nichtnegativen Koeffizienten der Summationsterme

$$u^{(i_1, \dots, i_N)} \cdot x_1^{i_1} x_2^{i_2} x_3^{i_3} \dots x_n^{i_n} \dots x_{N-1}^{i_{N-1}} x_N^{i_N}$$

des Dichtefunktionsexponenten. Insbesondere fällt dem Term

$$u^{(0, \dots, 0)} \cdot x_1^0 x_2^0 \dots x_N^0 = u^{(0, \dots, 0)} \cdot 1 = u^{(0, \dots, 0)}$$

wieder die Rolle des Normierungsfaktors zu.

Die Vektoren i können wir auch als *Multimengen* von Zufallsvariablen auffassen.

Gaußsche graphische Modelle

Hier werden ausschließlich Kovarianzterme $i \in \Delta$ zugelassen mit

$$\sum_{n=1}^N i_n = i_1 + i_2 + i_3 + \dots + i_N \leq 2.$$

Beweis.

Das Modell ist auch graphisch, denn es gibt grundsätzlich keinerlei Interaktion zwischen mehr als zwei Variablen. Es gilt $\Im(A \mid Z \mid B)$ genau dann, wenn Δ ausschließlich Auz-Terme und Buz-Terme enthält, also wenn es keine $\{a, b\}$ -Terme mit $a \in A$ und $b \in B$ gibt. \square

Marginalisierung

Ähnlich wie schon zuvor für Vektoren definieren wir Matrixausschnitte durch

$$M_{A,B} \stackrel{\text{def}}{=} (M_{ab} \mid a \in A, b \in B).$$

Die Matrix S_{CC} insbesondere enthält also alle Varianzen von und Kovarianzen zwischen Variablen aus C .

Die Matrix S_{AB} heißt übrigens auch „Kreuzkovarianzmatrix“ der Variablenmengen A und B .

Konditionierung

Bei geeigneter Variablennummerierung gilt in der Situation $A \uplus B = V$:

$$S = \begin{pmatrix} S_{AA} & S_{AB} \\ S_{BA} & S_{BB} \end{pmatrix} = \begin{pmatrix} S_{AA} & S_{AB} \\ S_{AB}^\top & S_{BB} \end{pmatrix}, \quad K = \begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix}$$

Die Matrixgleichung ergibt sich aus der (unschönen!) Formel zur Blockmatrixinvertierung.

Wissenswertes über multivariate Normalverteilungsdichten

Lemma

Für normalverteilte Zufallsvariablen $V = \{\mathbb{X}_1, \dots, \mathbb{X}_N\}$ gelten die folgenden Aussagen:

1. *Summenbildung:* $\mathbb{X} = \mathbb{X}' + \mathbb{X}'' \sim \mathcal{N}(\mu' + \mu'', S' + S'')$
2. *Affine Abbildung:* $A\mathbb{X} + b \sim \mathcal{N}(A\mu + b, ASA^\top)$
3. *Marginalisierung:* $\mathbb{X}_C \sim \mathcal{N}(\mu_C, S_{CC})$
4. *Konditionierung:* $\mathbb{X}_{A|x_B} \sim \mathcal{N}(\mu_{A|x_B}, S_{A|x_B})$

Dabei gelte $A \uplus B = V$ und es sind definiert:

$$\begin{aligned} \mu_{A|x_B} &= \mu_A + S_{AB} \cdot S_{BB}^{-1} \cdot (\mathbf{x}_B - \mu_B) & \mu &= (\mu_A^\top, \mu_B^\top)^\top \\ S_{A|x_B} &= S_{AA} - S_{AB} \cdot S_{BB}^{-1} \cdot S_{BA} & S &= \begin{pmatrix} S_{AA} & S_{AB} \\ S_{BA} & S_{BB} \end{pmatrix} \end{aligned}$$

5. Für die bedingte Kreuzkovarianzmatrix gilt der Zusammenhang:

$$(S_{A|x_B})^{-1} = K_{AA} = (S^{-1})_{AA}$$

Kovarianz und Konzentration

Nulleinträge $\hat{=}$ marginale & partielle Unabhängigkeiten

Lemma (Wermuth 1976)

Für normalverteilte Variablen $a, b \in V \sim \mathcal{N}(\mu, S)$ mit $a \neq b$ gilt:

- **Marginale Unabhängigkeit:** $\Im(a \mid \emptyset \mid b) \Leftrightarrow s_{ab} = 0$
- **Partielle Unabhängigkeit:** $\Im(a \mid \text{Rest} \mid b) \Leftrightarrow \kappa_{ab} = 0$

Gaußscher Diamant

$$\begin{pmatrix} \kappa_{11} & \kappa_{12} & 0 & \kappa_{14} \\ \kappa_{21} & \kappa_{22} & \kappa_{23} & 0 \\ 0 & \kappa_{32} & \kappa_{33} & \kappa_{34} \\ \kappa_{41} & 0 & \kappa_{43} & \kappa_{44} \end{pmatrix}$$

$$\begin{array}{c} \uparrow \\ \mathbb{X}_1 - \mathbb{X}_2 \\ | \\ \mathbb{X}_4 - \mathbb{X}_3 \end{array}$$

Gaußscher Schlüssel

$$\begin{pmatrix} \kappa_{11} & \kappa_{12} & 0 & 0 & 0 \\ \kappa_{21} & \kappa_{22} & \kappa_{23} & \kappa_{23} & 0 \\ 0 & \kappa_{32} & \kappa_{33} & \kappa_{34} & \kappa_{35} \\ 0 & \kappa_{42} & \kappa_{43} & \kappa_{44} & \kappa_{45} \\ 0 & 0 & \kappa_{53} & \kappa_{54} & \kappa_{55} \end{pmatrix}$$

$$\begin{array}{c} \uparrow \\ \mathbb{X}_1 - \mathbb{X}_2 \triangleleft \mathbb{X}_4 \triangleright \mathbb{X}_5 \\ | \\ \mathbb{X}_3 \end{array}$$

Beweis.

Die partielle Unabhängigkeit, d.h. die Frage nach einer Kante oder keiner Kante zwischen zwei Variablen im Markovnetz, lässt sich ganz einfach aus der inversen Kovarianzmatrix \mathbf{K} ablesen.

• Beweisidee 1:

Betrachte die bedingte Verteilung mit $A = \{a, b\}$ und $B = V \setminus \{a, b\}$.

Das Variablenpaar $(\mathbb{X}_a, \mathbb{X}_b)$ ist, bei gegebenem x_B , mit der Kovarianzmatrix $\mathbf{S}_{ab|}$ normalverteilt.

$\mathbb{X}_a, \mathbb{X}_b$ sind unabhängig genau dann, wenn $\mathbf{S}_{ab|}$ eine Diagonalmatrix ist; dies aber ist genau dann der Fall, wenn ihre Inverse, also $\mathbf{K}_{\{a,b\}}$ diagonal ist, also falls $\kappa_{ab} = \kappa_{ba} = 0$ ist.

• Beweisidee 2:

Die Normalverteilungsdichte ist faktorisierbar in Gibbs-Komponenten mit maximal zwei Variablen.

Sie lässt sich also in zwei Faktoren $g_{V \setminus \{a\}}$ und $h_{V \setminus \{b\}}$ genau dann zerlegen, wenn mindestens die Gibbs-Komponente für $\{a, b\}$ fehlt.

□

Charakterisierung bedingter Unabhängigkeiten

Lemma

Für Variablenmengen A, B, Z mit $V = A \uplus B \uplus Z$ und die bedingte Kreuzkovarianzmatrix

$$\mathbf{S}_{AB|Z} = [s_{ab}]_{a \in A}^{b \in B}, \quad s_{ab} \stackrel{\text{def}}{=} \text{Cov}[\mathbb{X}_a, \mathbb{X}_b | \mathbb{X}_Z = x_Z]$$

gilt die Beziehung:

$$\mathbf{S}_{AB|Z} = \mathbf{S}_{AB} - \mathbf{S}_{AZ} \cdot (\mathbf{S}_{ZZ})^{-1} \cdot \mathbf{S}_{ZB}$$

Satz (Speed & Kiiveri 1986)

Für normalverteilte Variablen $V = A \uplus B \uplus Z$ mit Kovarianzmatrix \mathbf{S} sind äquivalent:

1. $\mathbf{S}_{AB} = \mathbf{S}_{AZ} \cdot (\mathbf{S}_{ZZ})^{-1} \cdot \mathbf{S}_{ZB}$
2. $(\mathbf{S}^{-1})_{AB} = \mathbf{0}$ beziehungsweise $\mathbf{K}_{AB} = \mathbf{0}$
3. $\Im(A | Z | B)$

Maximum-Likelihood-Schätzung

für Gaußsche Graphische Modelle

Satz (Dempster 1972)

Es sei $\mathcal{G} = (V, \mathcal{E})$ ein Gaußsches Graphisches Modell mit der Generatorenmenge $\mathcal{C} \subset \mathfrak{P}V$ und sei $\omega \subset \mathbb{R}^N$ ein Datensatz mit den Statistiken \mathbf{m} und $\mathbf{\Sigma}$. Dann bilden \mathbf{m} und $\{\mathbf{\Sigma}_{cc} \mid C \in \mathcal{C}\}$ eine minimale suffiziente Statistik des Modells für ω .

Die Maximum-Likelihood-Parameter μ und \mathbf{S} bzw. \mathbf{K} gehorchen den Bedingungen

$$\mu = \mathbf{m}$$

$$s_{ab} = \sigma_{ab}$$

$$\kappa_{ab} = 0$$

$$\{a, b\} \in \mathcal{E} \vee a = b$$

$$\{a, b\} \notin \mathcal{E} \wedge a \neq b$$

Bemerkung

Da $\mathcal{N}(\mu, \mathbf{\Sigma})$ das saturierte Modell ist, beträgt die Devianz:

$$\text{dev}(\mathcal{C}) = 2 \cdot (\ell(\mu, \mathbf{\Sigma}) - \ell(\mu, \mathbf{S})) = T \cdot \log(\det \mathbf{S} / \det \mathbf{\Sigma})$$

Maximum-Likelihood-Schätzung

Existenz & Eindeutigkeit

Datenkovarianzmatrix

$$\mathbf{\Sigma} = \begin{pmatrix} 3.023 & 1.258 & 1.004 \\ 1.258 & 1.709 & 0.842 \\ 1.004 & 0.842 & 1.116 \end{pmatrix}$$

ML-Kovarianzmatrix

$$\mathbf{S} = \begin{pmatrix} 3.023 & 1.258 & 0.620 \\ 1.258 & 1.709 & 0.842 \\ 0.620 & 0.842 & 1.116 \end{pmatrix}$$

ML-Konzentrationsmatrix

$$\mathbf{K} = \begin{pmatrix} 0.477 & -0.351 & 0.000 \\ -0.351 & 1.190 & -0.703 \\ 0.000 & -0.703 & 1.426 \end{pmatrix}$$

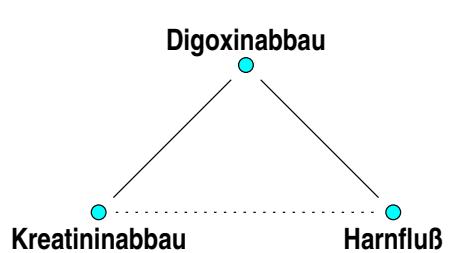
Satz (Dempster 1972)

Es seien \mathbf{A}, \mathbf{B} zwei symmetrische, positiv-definite $(N \times N)$ -Matrizen. Ferner sei $\mathcal{E} \subseteq \{1, \dots, N\} \times \{1, \dots, N\}$ symmetrisch mit $(i, i) \in \mathcal{E}$ für alle i . Dann gibt es eine symmetrische, positiv-definite Matrix \mathbf{S} mit

$$\begin{aligned} s_{ij} &= a_{ij} & \forall (i, j) \in \mathcal{E} \\ (\mathbf{S}^{-1})_{ij} &= b_{ij} & \forall (i, j) \notin \mathcal{E} \end{aligned}$$

und \mathbf{S} ist eindeutig mit diesen Eigenschaften.

Beispiel — Digoxin-Abbau



Datensammlung

$\omega \subset \mathbb{R}^3$

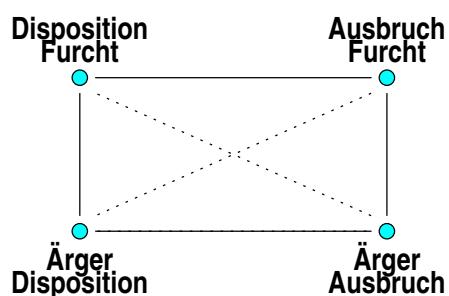
 $|\omega| = 35$ Patienten \mathbb{X} = Abbau von Kreatinin \mathbb{Y} = Abbau von Digoxin \mathbb{Z} = Harnflußrate

$\chi^2\text{-Test} \Rightarrow \{\mathbb{X}, \mathbb{Z}\} \notin \mathcal{E}$

Beispiel — Furcht versus Ärger

Datensammlung

$\omega \subset \mathbb{R}^4$

 $|\omega| = 684$ Versuchspersonen

• Augenblickszustand:

 \mathbb{X} = Furcht \mathbb{W} = Ärger

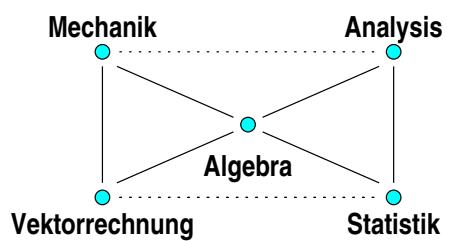
• mentale Prägung:

 \mathbb{Z} = Furcht \mathbb{Y} = Ärger $\chi^2\text{-Test ergibt:}$

$\mathfrak{S}(\mathbb{X} | \mathbb{W}, \mathbb{Z} | \mathbb{Y})$

$\mathfrak{S}(\mathbb{W} | \mathbb{X}, \mathbb{Y} | \mathbb{Z})$

Beispiel — Punktzahl in Übungsserien



Datensammlung

$\omega \subset \{0, 1, 2, \dots, 100\}^5$

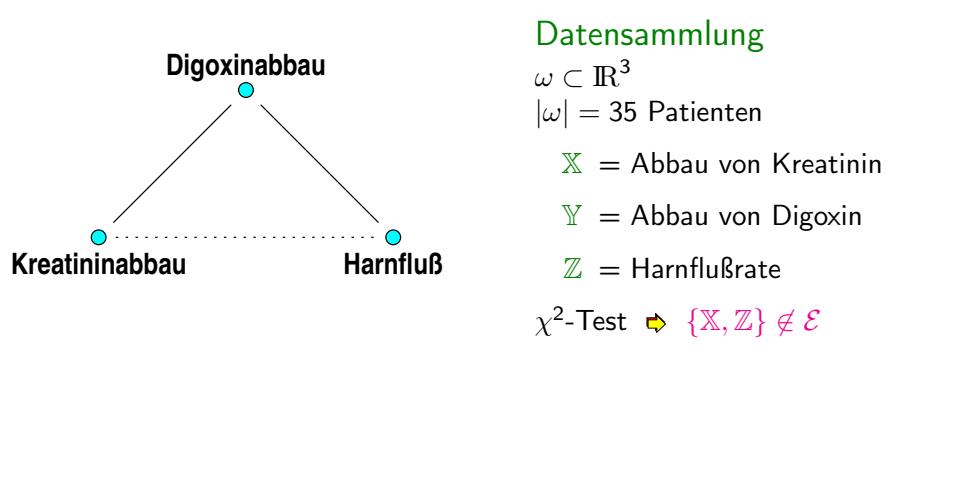
 $|\omega| = 88$ Studierende

- 5 Übungsgruppen in 5 Fächern
- je 100 Punkte erzielbar

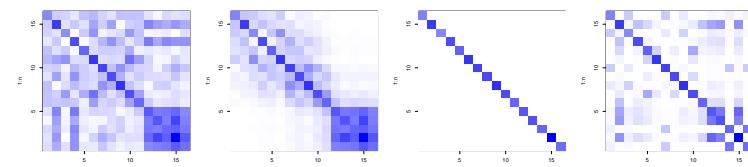
Inferenz

$\mathfrak{S}(\text{Stat} | \text{Alg, Analysis} | \cdot)$

$\mathfrak{S}(\text{Mech} | \text{Alg, Vektor} | \cdot)$

Zentrale Befähigung: **Algebra**

Beispiel — Schriftzeichenklassifikation

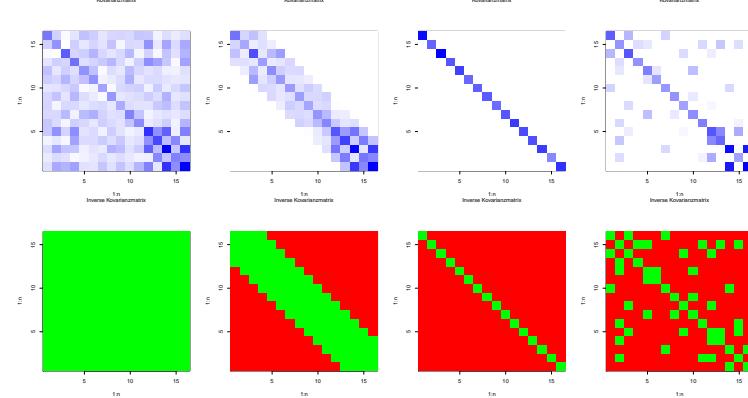


Beispiel

Datensatz letter

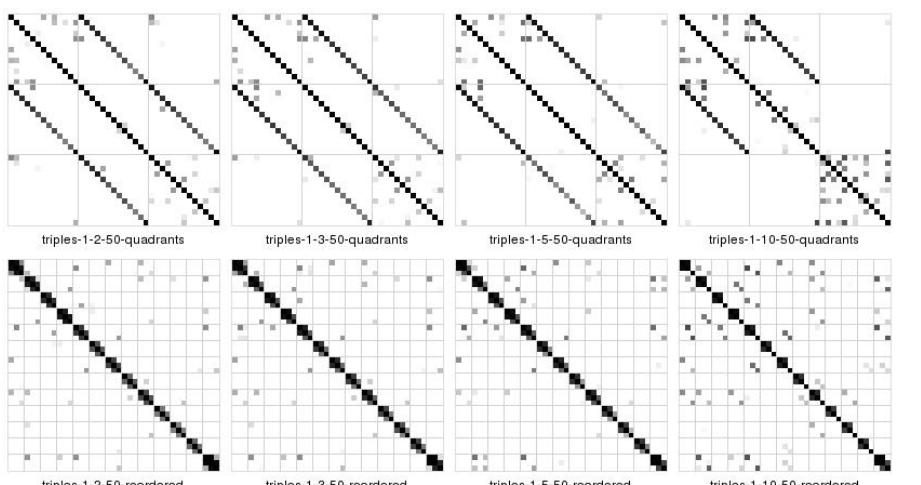
16 Merkmale

alle Klassen

oben:
Kovarianz
 $\hat{S} = K^{-1}$ Mitte:
Konzentration
 K erfüllt A unten:
Adjazenz A
Abhängigkeits-
muster (durch Δ
gegeben)

Beispiel — Sprachsignalparameter

12 MFCC-Parameter · drei Zeitpunkte

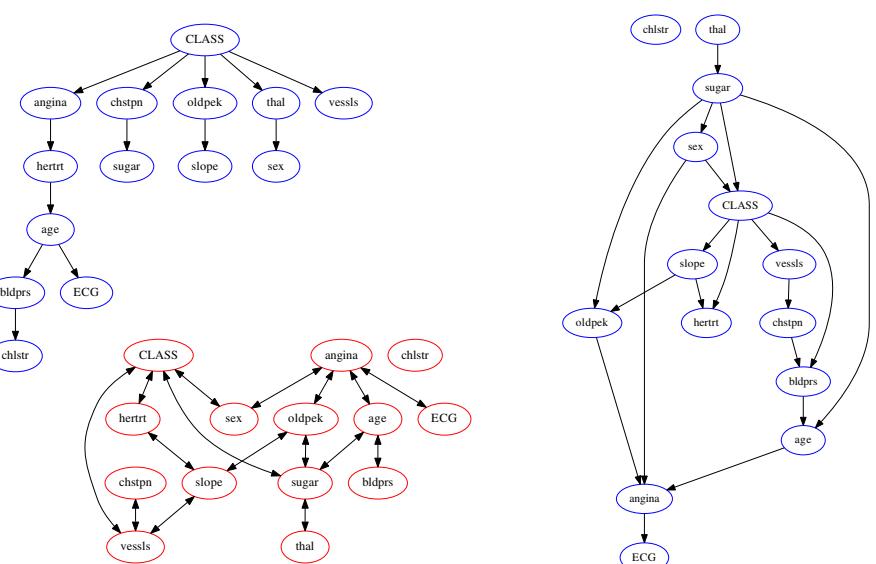


Dünne Abhängigkeitsstruktur

Die Vergangenheit wird durch unmittelbare Vorgänger „maskiert“.

Beispiel — Statlog Herzdatensammlung I

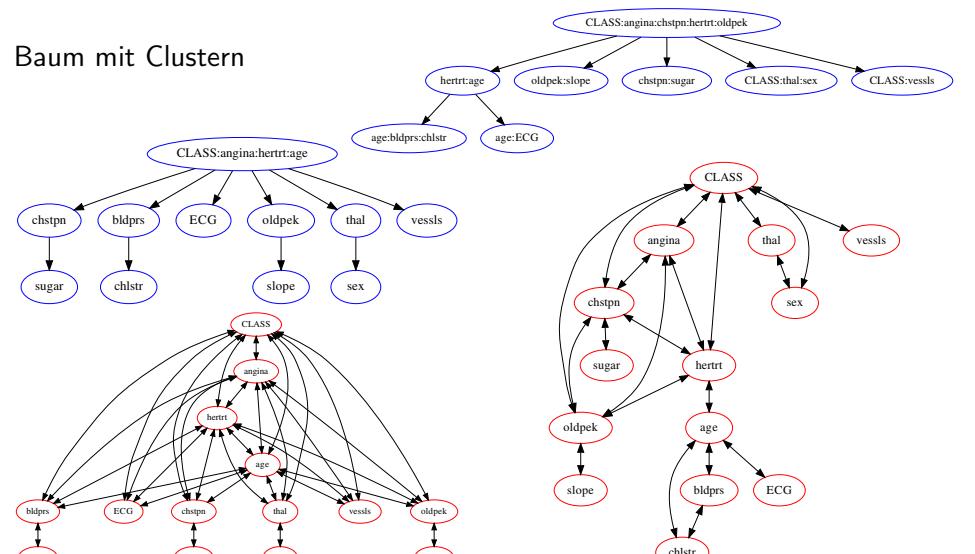
Numerisch kodierter Datensatz · BIC-optimaler Baum/DAG/UG



Beispiel — Statlog Herzdatensammlung II

Numerisch kodierter Datensatz · BIC-optimaler Cluster/Cliquen-Baum

Baum mit Clustern



Baum mit Cliquen

Assoziationsregeln und Netzwerkanalyse

Bedingte statistische Unabhängigkeit

Graphische Modelle: ungerichtete Graphen

Kausale Modelle: gerichtete azyklische Graphen

Berechnen bedingter Wahrscheinlichkeiten

Parameterschätzung in Bayesnetzen und Loglinearmodellen

Aufdeckung der Abhängigkeitsstruktur

Kovarianzselektion

Zusammenfassung (6)

1. Kovarianz und Korrelation sind **quantitative** Charakterisierungen der **linearen** Aspekte („Regression“) statistischer Abhangigkeit.
2. Die **Transformation** quantifiziert statistische Abhangigkeit in allgemeiner Form, setzt aber die Kenntnis der **wahren Verteilung** voraus.
3. Die **Warenkorbanalyse** sucht **Assoziationsregeln** mit gleichermaßen hohen Werten fur **Support**, **Konfidenz** und **Relevanz** (z.B. **Apriori-Algorithmus**).
4. Das **Dependenzmodell** charakterisiert die **bedingten Unabhangigkeiten** $\Im(A \mid Z \mid B)$ je dreier Attributmengen einer Verteilung.
5. Verteilungen heien **graphisch (kausal)**, wenn ihr DM durch die (δ -)**Separation** eines **UG (DAG)** gegeben ist.
6. Kausale Modelle faktorisieren **attributweise** in **bedingte Wahrsch'keiten**, graphische Modelle faktorisieren **cliquenweise** in **Gibbspotenziale**.
7. **Kordale Modelle** besitzen einen **triangulierten UG**, einen **moralischen DAG** und eine **Cliquenschnittfaktorisierung**.
8. Statistische **Inferenz** ist nur fur **Ketten** und (Verbund-)**Baume** effizient.
9. Die **ML-Schatzung** der **Modellparameter** aus Daten beruht auf **relativen Hufigkeiten (DAG)** oder dem **Maximum-Entropie-Prinzip (UG)**.
10. Die Aufdeckung der **Modellstruktur** basiert auf **Unabhangigkeitstests** (Kantenelimination/Grenzengraph) oder **gieriger Suche mit Strafterm**.