

Inhaltsverzeichnis

Inhaltsverzeichnis	v
Abbildungsverzeichnis	x
1 Einleitung	1
2 Datentypen und Datenrepräsentationen	9
2.1 Merkmalsdaten (Datentyp F1)	9
2.1.1 Abbildungen auf Standardskalen	11
2.1.2 Abbildung von fehlenden Werten (<i>missing values</i>)	12
2.2 Distanz- und Ähnlichkeitsmaße (Fall F2)	13
2.3 Erzeugung von Distanz- und Unähnlichkeitsmaßen	14
2.3.1 Kontinuierliche oder diskrete Daten	14
2.3.2 Binärdaten	16
2.3.3 Nominal- und Ordinaldaten	16
2.3.4 Mischdaten und fehlende Einträge	17
2.4 Spezielle Datenrepräsentationen und Distanzmaße	17
2.4.1 Editdistanzen auf Zeichenketten	18
2.4.2 Vektorraummodell für Text	19
2.4.3 Bildrepräsentation	20
2.4.4 Repräsentationen von Zeitserien	24

3	Datenpräsentation und -exploration	27
3.1	Einige multivariate Visualisierungstechniken	31
3.1.1	Fehler- und Boxplots	34
3.1.2	<i>Scatterplots</i> -Matrix	35
3.1.3	Parallele Koordinaten und <i>Andrews-Plots</i>	36
3.1.4	Ikonographische Darstellungen	38
3.2	Pixelorientierte Visualisierungen	41
3.3	Die interaktive „Tabellenlupe“	41
3.4	<i>Missing-Value</i> -Statistik und Assoziationsanalyse	44
3.5	Integrierte Assoziationsanalyse	45
3.6	Trellis-Darstellung	46
4	Statistische Grundlagen	49
4.1	Zufallsexperimente, Wahrscheinlichkeiten und Verteilungen	50
4.2	Zufallsvariablen und Wahrscheinlichkeitsverteilungen	52
4.2.1	Mehrdimensionale Verallgemeinerung	53
4.2.2	Deskriptive Statistik für metrische Variablen	54
4.2.3	Quantile, Median und Ordnungsstatistiken	57
4.3	Die Gauß'sche Normalverteilung	59
4.4	Konfidenzintervalle und Signifikanz	61
4.5	Nullhypothesen und p-Wert	62
4.6	Hypothesentest und Fehlerarten	63
4.7	Ausgewählte statistische Tests	64
4.7.1	Mittelwert einer Stichprobe	65
4.7.2	Kleine Stichproben und die <i>emphStudent</i> -t-Verteilung	65
4.7.3	Ein- und zweiseitige Fragestellungen	66

4.7.4	t-Test: Mittelwertvergleich zweier Stichproben	67
4.7.5	F-Test: Varianzgleichheit zweier Stichproben	69
4.8	Vergleich mehrerer Stichproben: ANOVA-Test	69
4.9	χ^2 -Verteilung und <i>Goodness-of-fit</i> -Test	70
4.10	Kolmogorov-Smirnov-Test	72
4.11	Bernoulli-Experiment, Binomialverteilung und Zahlverhältnisse	73
4.12	Kontingenztabellen und Assoziation	74
4.12.1	Kontingenztabellen und χ^2 -Test	74
4.12.2	Nichtparametrische Tests	79
4.12.3	Nichtparametrische Tests für abhängige Stichproben	80
4.13	Weitere Assoziationsmaße für zwei Verteilungen	81
4.13.1	Entropie-basierte Assoziationsmaße	81
4.13.2	Lineare Korrelation	84
4.13.3	Nichtparametrische Korrelationsmaße	85
5	Modellbildung	87
5.1	Bayes'sche Modelle und Methoden	89
5.2	Approximationsmodelle	93
5.3	Klassifikation	97
5.4	Clustermodelle	100
5.4.1	Partitionierende Verfahren	101
5.4.2	Hierarchische Verfahren	102
5.4.3	Probabilistische modellbasierte Clusterverfahren	104
5.4.4	Weitere neuronale Verfahren: CLM	104
5.5	Assoziationsregeln	105
5.5.1	Der Apriori-Algorithmus	106

5.5.2	Verallgemeinerte und quantitative Assoziationsregeln	106
5.5.3	<i>Contrast Set Mining</i>	107
5.6	Merkmals- und Modellselektion	108
5.6.1	Merkmalsselektion	108
5.6.2	Modellselektion	108
5.7	Wichtige Modelle zur Regression	110
5.7.1	Lineare Regression	110
5.7.2	Logistische Regression	120
5.7.3	Lokale logistische Regression mittels Maximum Likelihood	121
5.7.4	ROC Analyse	126
5.8	Neuronale-Netz-Modelle: MLP	133
5.9	Selbstorganisierende Karten	136
5.10	Multidimensionale Skalierung (MDS)	140
5.10.1	Klassische multidimensionale Skalierung	141
5.10.2	Least-Square- oder Kruskal-Scaling	143
5.10.3	MDS nach Sammon	143
5.10.4	Dimensionsreduktion mit FastMap	145
6	Grundlagen hyperbolischer Geometrie	151
6.1	Geschichte	152
6.2	Abbilder des hyperbolischen Raumes	154
6.3	Metriken für die fünf hyperbolischen Modelle	157
6.3.1	Ein weitere \mathbb{H}^2 Einbettungen in den \mathbb{R}^6	158
6.4	Eigenschaften des \mathbb{H}^2 : Geodäten, Flächen etc.	160
6.4.1	Längenmessung im Poincaré-Modell	169
6.4.2	Generator einer isotropen Datenverteilung im \mathbb{H}^2 :	170

6.5	Die Isometrien des Poincaré-Modelles	171
6.6	Mensch-Maschine-Interaktion im Poincaré-Modell	172
6.6.1	Animation	175
6.6.2	Zeichnen von \mathbb{H}^2 -Verbindungen	175
6.6.3	Nonkonforme Vergrößerungsabbildung: Zooming	175
7	Datenvisualisierung im hyperbolischen Raum	177
7.1	HTL – Hyperbolic Tree Layout	178
7.2	HSOM – Hyperbolic Self-Organizing Map	180
7.2.1	Interpolationsansätze	186
7.2.2	Unebenheiten bei hochdimensionalen Gittereinbettungen	190
7.3	HMDS – Hyperbolic Multi-Dimensional Scaling	193
7.3.1	Vorverarbeitung der Unähnlichkeiten	194
7.3.2	Beispiel: der <i>Iris</i> -Datensatz	195
7.3.3	Beispiel: der <i>Animals</i> -Datensatz	195
7.3.4	Beispiel: Zufallsbäume in 200 Dimensionen	195
7.4	Verteilungen in hochdimensionalen Räumen	198
8	\mathbb{H}^2-Navigation in Dokumentkolektionen mit hybrider Architektur	205
8.1	Anwendungsbeispiele: <i>Space of Movies</i>	206
8.1.1	Repräsentation der Filme	206
8.1.2	Modulation des Ähnlichkeitskontrastes	207
8.1.3	Ist die hyperbolische Einbettung letztlich vorteilhaft?	210
8.2	Anwendungsbeispiele: Navigation in Bildsammlungen	211
8.3	Eigenschaftsvergleich der Layouttechniken	214
8.3.1	Zulässige Typen von Eingabedaten	214

8.3.2	Skalierverhalten bezüglich der Datenanzahl N	217
8.3.3	Layoutresultat	217
8.3.4	Neue Objekte	217
8.4	Ein hybrider Ansatz zum Navigieren in großen Datenkollektionen	218
8.5	Anwendungsbeispiele: Reuters-Nachrichten	219
8.5.1	Textkategorisierung	219
8.5.2	Suchanfragen und ähnliche Dokumente	226
8.5.3	Weitere Schritte in der Ergebnispräsentation	228
8.5.4	Wahl der HSOM-Gittergröße	230
8.5.5	Auswahloptimierung für die Ähnlichkeitssuche	230
8.6	Jumpstarting	231
9	Fallbeispiel: Datamining in der Herzchirurgie	233
9.1	Anwendungsdomäne Herzchirurgie in Lahr	233
9.1.1	Tätigkeitsspektrum	234
9.1.2	Risikoadjustierung und EuroSCORE	235
9.2	Probleme und Herausforderungen	237
9.3	Aufbau eines Data-Marts	238
9.4	Intranet Auswertungs-Portal	242
9.5	Risikoadjustierte Hypothesentests	246
9.6	Interaktive Präsentation von Merkmalsähnlichkeiten	250
	Literatur	253