

Inhalt

Vorwort	XIII
1 Worum geht es?	1
1.1 Data Mining und maschinelles Lernen	2
Beschreibung strukturierter Muster	4
Maschinelles Lernen	5
Data Mining	7
1.2 Einfache Beispiele: Das Wetterproblem und andere	8
Das Wetterproblem	9
Kontaktlinsen: Ein idealisiertes Problem	11
Iris: Eine klassische numerische Datenmenge	14
CPU-Leistung: Einführung in numerische Vorhersagen	15
Tarifverhandlungen: Ein realistischeres Beispiel	16
Sojabohnen-Klassifizierung: Ein Erfolg des klassischen maschinellen Lernens	19
1.3 Anwendungen in der Praxis	21
Entscheidungen durch Beurteilungen	22
Bildanalyse	23
Lastabschätzung	24
Diagnose	25
Marketing und Verkauf	26
1.4 Maschinelles Lernen und Statistik	28
1.5 Generalisierung als Suche	29
Auflistung des Konzeptraums	31
Bias	32
Sprach-Bias	32
Such-Bias	33
Bias zur Vermeidung einer Überanpassung	34
1.6 Data Mining und Ethik	35
1.7 Weiterführende Literatur	37

2 Eingaben: Konzepte, Instanzen, Attribute	41
2.1 Was ist ein Konzept?	42
2.2 Was enthält ein Beispiel?	45
2.3 Was enthält ein Attribut?	49
2.4 Aufbereitung der Eingaben	52
Sammeln der Daten	53
Das ARFF-Format	54
Attributtypen	55
Fehlende Werte	57
Ungenauere Werte	58
Lernen Sie Ihre Daten kennen	59
2.5 Weiterführende Literatur	60
3 Ausgabe: Wissensdarstellung	61
3.1 Entscheidungstabellen	61
3.2 Entscheidungsbäume	62
3.3 Klassifikationsregeln	63
3.4 Assoziationsregeln	67
3.5 Regeln mit Ausnahmen	69
3.6 Regeln mit Relationen	72
3.7 Bäume für numerische Vorhersagen	74
3.8 Instanzbasierte Darstellung	76
3.9 Cluster	80
3.10 Weiterführende Literatur	81
4 Algorithmen: Die grundlegenden Methoden	83
4.1 Ableitung elementarer Regeln	84
Fehlende Werte und numerische Attribute	85
Diskussion	88
4.2 Statistische Modellierung	88
Fehlende Werte und numerische Attribute	92
Diskussion	95

4.3	Teile und Herrsche; Der Aufbau von Entscheidungsbäumen	95
	Berechnung des Informationsmaßes	99
	Attribute mit vielen Verzweigungen	101
	Diskussion	104
4.4	Abdeckungs-Algorithmen: die Konstruktion von Regeln	104
	Regeln oder Bäume?	105
	Ein einfacher Abdeckungs-Algorithmus	106
	Regeln oder Entscheidungslisten?	111
4.5	Erzeugen von Assoziationsregeln	112
	Gegenstandsmengen	112
	Assoziationsregeln	113
	Regeln effizient generieren	117
	Diskussion	119
4.6	Lineare Modelle	120
	Numerische Vorhersagen	120
	Klassifikation	121
	Diskussion	122
4.7	Instanzbasiertes Lernen	123
	Die Distanzfunktion	123
	Diskussion	124
4.8	Weiterführende Literatur	125
5	Glaubwürdigkeit: Auswertung des Gelernten	127
5.1	Trainieren und Testen	128
5.2	Leistungsvorhersage	131
5.3	Kreuzvalidierung	133
5.4	Andere Schätzverfahren	136
	Leave-one-out	136
	Bootstrap	137
5.5	Data Mining-Verfahren im Vergleich	138
5.6	Vorhersage von Wahrscheinlichkeiten	142
	Quadratische Verlustfunktion	143
	Informatorische Verlustfunktion	144
	Diskussion	145

5.7	Die Kosten	146
	Steigerungsdiagramme	148
	ROC-Kurven	151
	Berücksichtigung der Lernkosten	154
	Diskussion	155
5.8	Auswertung numerischer Vorhersagen	157
5.9	Das Prinzip der minimalen Beschreibungslänge	161
5.10	Anwendung des MDL-Prinzips auf das Clustering	165
5.11	Weiterführende Literatur	166
6	Implementierungen: Maschinelles Lernen in der Praxis	169
6.1	Entscheidungsbäume	171
	Numerische Attribute	171
	Fehlende Werte	173
	Pruning	174
	Abschätzung der Fehlerrate	177
	Komplexität der Entscheidungsbaum-Induktion	180
	Von Bäumen zu Regeln	181
	C4.5: Auswahlmöglichkeiten und Optionen	182
	Diskussion	183
6.2	Klassifikationsregeln	184
	Kriterien für die Auswahl von Auswertungen	184
	Fehlende Werte, numerische Attribute	186
	Gute Regeln, schlechte Regeln	187
	Gute Regeln erzeugen	188
	Gute Entscheidungslisten erzeugen	190
	Wahrscheinlichkeitswert zur Regelevaluation	191
	Regeln mit einer Testmenge evaluieren	193
	Regeln aus partiellen Bäumen entnehmen	196
	Regeln mit Ausnahmen	200
	Diskussion	203
6.3	Erweiterung der linearen Klassifikation: Support-Vektor-Maschinen	204
	Die maximal diskriminierende Hyperebene	206
	Nichtlineare Klassengrenzen	208
	Diskussion	209

6.4	Instanzbasiertes Lernen	210
	Zahl der Exemplare verringern	210
	Verrauschte Exemplare beschneiden	211
	Attribute gewichten	213
	Exemplare generalisieren	214
	Distanzfunktionen für generalisierte Exemplare	215
	Generalisierte Distanzfunktionen	217
	Diskussion	217
6.5	Numerische Vorhersage	219
	Modellbäume	220
	Den Baum aufbauen	221
	Den Baum beschneiden	221
	Nominale Attribute	222
	Fehlende Werte	223
	Pseudocode für die Modellbaum-Induktion	224
	Lokal gewichtete lineare Regression	227
	Diskussion	228
6.6	Clustering	229
	Iteratives distanzbasiertes Clustering	230
	Inkrementelles Clustering	231
	Kategorienützlichkeit	236
	Wahrscheinlichkeitsbasiertes Clustering	238
	Der EM-Algorithmus	241
	Das Mischungsmodell erweitern	243
	Bayessches Clustering	245
	Diskussion	247
7	Es geht weiter:	
	Aufbereitung der Ein- und Ausgabe	249
7.1	Attributauswahl	252
	Verfahrensunabhängige Auswahl	254
	Durchsuchen des Attributraums	255
	Verfahrensspezifische Auswahl	257
7.2	Diskretisierung numerischer Attribute	259
	Unüberwachte Diskretisierung	260
	Entropie-basierte Diskretisierung	261
	Weitere Methoden der Diskretisierung	265
	Entropiebasierte und fehlerbasierte Diskretisierung im Vergleich	266
	Diskrete in numerische Attribute umwandeln	268

7.3	Automatische Datensäuberung	269
	Entscheidungsbäume verbessern	269
	Robuste Regression	270
	Anomalien entdecken	272
7.4	Kombination mehrerer Modelle	273
	Bagging	274
	Boosting	277
	Stacking	282
	Fehlerkorrigierende Ausgabecodes	284
7.5	Weiterführende Literatur	287
8	Nägel mit Köpfen:	
	Algorithmen des maschinellen Lernens in Java	291
8.1	Die ersten Schritte	293
8.2	Javadoc und die Klassenbibliothek	298
	Klassen, Instanzen und Packages	298
	Das weka.core-Package	299
	Das weka.classifiers-Package	300
	Andere Packages	303
	Indizes	303
8.3	Datenmengen mit maschinellen Lernprogrammen verarbeiten	304
	M5' verwenden	304
	Allgemeine Optionen	306
	Verfahrensspezifische Optionen	309
	Klassifizierer	310
	Metalernverfahren	314
	Filter	317
	Assoziationsregeln	322
	Clustering	324
8.4	Eingebettetes maschinelles Lernen	326
	Ein einfacher Nachrichten-Klassifizierer	326
	Main()	328
	MessageClassifier()	328
	UpdateModel()	333
	ClassifyMessage()	334
8.5	Neue Lernverfahren schreiben	335
	Ein Beispielklassifizierer	336
	BuildClassifier()	336

MakeTree()	336
ComputeInfoGain()	341
ClassifyInstance()	342
Main()	342
Konventionen zur Implementierung von Klassifizierern	343
Das Schreiben von Filtern	344
Ein Beispielfilter	346
Konventionen für das Schreiben von Filtern	350
9 Blick nach vorn	351
9.1 Lernen aus sehr großen Datenmengen	352
9.2 Visualisierung von maschinellem Lernen	355
Visualisierung der Eingabe	355
Visualisierung der Ausgabe	357
9.3 Das Einbinden von Domänenwissen	359
9.4 Text Mining	362
Schlüsselworte in Dokumenten finden	362
Informationen aus Fließtext entnehmen	364
Soft-Parsing	366
9.5 Mining im World Wide Web	367
9.6 Weiterführende Literatur	367
Literatur	369
Stichwortverzeichnis	381