

Contents

Acknowledgements	xix
Preface	xxi
References	xxxix
Part I Preliminaries	1
1 Tasks	3
1.1 Introduction	3
1.1.1 Knowledge	4
1.1.2 Inference	4
1.2 Inductive learning tasks	5
1.2.1 Domain	5
1.2.2 Instances	5
1.2.3 Attributes	5
1.2.4 Target attribute	6
1.2.5 Input attributes	6
1.2.6 Training set	6
1.2.7 Model	7
1.2.8 Performance	7
1.2.9 Generalization	8
1.2.10 Overfitting	8
1.2.11 Algorithms	8
1.2.12 Inductive learning as search	9
1.3 Classification	9
1.3.1 Concept	10
1.3.2 Training set	10
1.3.3 Model	11
1.3.4 Performance	12
1.3.5 Generalization	13
1.3.6 Overfitting	13
1.3.7 Algorithms	13

1.4	Regression	14
1.4.1	Target function	14
1.4.2	Training set	14
1.4.3	Model	15
1.4.4	Performance	15
1.4.5	Generalization	15
1.4.6	Overfitting	15
1.4.7	Algorithms	16
1.5	Clustering	16
1.5.1	Motivation	16
1.5.2	Training set	17
1.5.3	Model	18
1.5.4	Crisp vs. soft clustering	18
1.5.5	Hierarchical clustering	18
1.5.6	Performance	18
1.5.7	Generalization	19
1.5.8	Algorithms	19
1.5.9	Descriptive vs. predictive clustering	19
1.6	Practical issues	19
1.6.1	Incomplete data	20
1.6.2	Noisy data	20
1.7	Conclusion	20
1.8	Further readings	21
	References	22
2	Basic statistics	23
2.1	Introduction	23
2.2	Notational conventions	24
2.3	Basic statistics as modeling	24
2.4	Distribution description	25
2.4.1	Continuous attributes	25
2.4.2	Discrete attributes	36
2.4.3	Confidence intervals	40
2.4.4	m -Estimation	43
2.5	Relationship detection	47
2.5.1	Significance tests	48
2.5.2	Continuous attributes	50
2.5.3	Discrete attributes	52
2.5.4	Mixed attributes	56
2.5.5	Relationship detection caveats	61
2.6	Visualization	62
2.6.1	Boxplot	62
2.6.2	Histogram	63
2.6.3	Barplot	64
2.7	Conclusion	65
2.8	Further readings	66
	References	67

Part II	Classification	69
3	Decision trees	71
3.1	Introduction	71
3.2	Decision tree model	72
3.2.1	Nodes and branches	72
3.2.2	Leaves	74
3.2.3	Split types	74
3.3	Growing	76
3.3.1	Algorithm outline	76
3.3.2	Class distribution calculation	78
3.3.3	Class label assignment	79
3.3.4	Stop criteria	80
3.3.5	Split selection	82
3.3.6	Split application	86
3.3.7	Complete process	86
3.4	Pruning	90
3.4.1	Pruning operators	91
3.4.2	Pruning criterion	91
3.4.3	Pruning control strategy	100
3.4.4	Conversion to rule sets	101
3.5	Prediction	103
3.5.1	Class label prediction	104
3.5.2	Class probability prediction	104
3.6	Weighted instances	105
3.7	Missing value handling	106
3.7.1	Fractional instances	106
3.7.2	Surrogate splits	113
3.8	Conclusion	114
3.9	Further readings	114
	References	116
4	Naïve Bayes classifier	118
4.1	Introduction	118
4.2	Bayes rule	118
4.3	Classification by Bayesian inference	120
4.3.1	Conditional class probability	120
4.3.2	Prior class probability	121
4.3.3	Independence assumption	122
4.3.4	Conditional attribute value probabilities	122
4.3.5	Model construction	123
4.3.6	Prediction	124
4.4	Practical issues	125
4.4.1	Zero and small probabilities	125
4.4.2	Linear classification	126
4.4.3	Continuous attributes	127

4.4.4	Missing attribute values	128
4.4.5	Reducing naïvety	129
4.5	Conclusion	131
4.6	Further readings	131
	References	132
5	Linear classification	134
5.1	Introduction	134
5.2	Linear representation	136
5.2.1	Inner representation function	137
5.2.2	Outer representation function	138
5.2.3	Threshold representation	139
5.2.4	Logit representation	142
5.3	Parameter estimation	145
5.3.1	Delta rule	145
5.3.2	Gradient descent	149
5.3.3	Distance to decision boundary	152
5.3.4	Least squares	153
5.4	Discrete attributes	154
5.5	Conclusion	155
5.6	Further readings	156
	References	157
6	Misclassification costs	159
6.1	Introduction	159
6.2	Cost representation	161
6.2.1	Cost matrix	161
6.2.2	Per-class cost vector	162
6.2.3	Instance-specific costs	163
6.3	Incorporating misclassification costs	164
6.3.1	Instance weighting	164
6.3.2	Instance resampling	167
6.3.3	Minimum-cost rule	169
6.3.4	Instance relabeling	174
6.4	Effects of cost incorporation	176
6.5	Experimental procedure	180
6.6	Conclusion	184
6.7	Further readings	185
	References	187
7	Classification model evaluation	189
7.1	Introduction	189
7.1.1	Dataset performance	189
7.1.2	Training performance	189
7.1.3	True performance	189
7.2	Performance measures	190
7.2.1	Misclassification error	191

7.2.2	Weighted misclassification error	191
7.2.3	Mean misclassification cost	192
7.2.4	Confusion matrix	194
7.2.5	ROC analysis	200
7.2.6	Probabilistic performance measures	210
7.3	Evaluation procedures	213
7.3.1	Model evaluation vs. modeling procedure evaluation	213
7.3.2	Evaluation caveats	214
7.3.3	Hold-out	217
7.3.4	Cross-validation	219
7.3.5	Leave-one-out	221
7.3.6	Bootstrapping	223
7.3.7	Choosing the right procedure	227
7.3.8	Evaluation procedures for temporal data	230
7.4	Conclusion	231
7.5	Further readings	232
	References	233

Part III Regression 235

8	Linear regression	237
8.1	Introduction	237
8.2	Linear representation	238
8.2.1	Parametric representation	239
8.2.2	Linear representation function	240
8.2.3	Nonlinear representation functions	241
8.3	Parameter estimation	242
8.3.1	Mean square error minimization	242
8.3.2	Delta rule	243
8.3.3	Gradient descent	245
8.3.4	Least squares	248
8.4	Discrete attributes	250
8.5	Advantages of linear models	251
8.6	Beyond linearity	252
8.6.1	Generalized linear representation	252
8.6.2	Enhanced representation	255
8.6.3	Polynomial regression	256
8.6.4	Piecewise-linear regression	257
8.7	Conclusion	258
8.8	Further readings	258
	References	259
9	Regression trees	261
9.1	Introduction	261
9.2	Regression tree model	262
9.2.1	Nodes and branches	262

9.2.2	Leaves	262
9.2.3	Split types	262
9.2.4	Piecewise-constant regression	262
9.3	Growing	263
9.3.1	Algorithm outline	264
9.3.2	Target function summary statistics	265
9.3.3	Target value assignment	266
9.3.4	Stop criteria	267
9.3.5	Split selection	268
9.3.6	Split application	271
9.3.7	Complete process	272
9.4	Pruning	274
9.4.1	Pruning operators	275
9.4.2	Pruning criterion	275
9.4.3	Pruning control strategy	277
9.5	Prediction	277
9.6	Weighted instances	278
9.7	Missing value handling	279
9.7.1	Fractional instances	279
9.7.2	Surrogate splits	284
9.8	Piecewise linear regression	284
9.8.1	Growing	285
9.8.2	Pruning	289
9.8.3	Prediction	290
9.9	Conclusion	292
9.10	Further readings	292
	References	293
10	Regression model evaluation	295
10.1	Introduction	295
10.1.1	Dataset performance	295
10.1.2	Training performance	295
10.1.3	True performance	295
10.2	Performance measures	296
10.2.1	Residuals	296
10.2.2	Mean absolute error	297
10.2.3	Mean square error	297
10.2.4	Root mean square error	299
10.2.5	Relative absolute error	299
10.2.6	Coefficient of determination	300
10.2.7	Correlation	301
10.2.8	Weighted performance measures	301
10.2.9	Loss functions	302
10.3	Evaluation procedures	303
10.3.1	Hold-out	304
10.3.2	Cross-validation	304

10.3.3	Leave-one-out	305
10.3.4	Bootstrapping	305
10.3.5	Choosing the right procedure	307
10.4	Conclusion	309
10.5	Further readings	309
	References	310

Part IV Clustering 311

11	(Dis)similarity measures	313
11.1	Introduction	313
11.2	Measuring dissimilarity and similarity	313
11.3	Difference-based dissimilarity	314
11.3.1	Euclidean distance	314
11.3.2	Minkowski distance	315
11.3.3	Manhattan distance	316
11.3.4	Canberra distance	316
11.3.5	Chebyshev distance	317
11.3.6	Hamming distance	318
11.3.7	Gower's coefficient	318
11.3.8	Attribute weighting	320
11.3.9	Attribute transformation	320
11.4	Correlation-based similarity	321
11.4.1	Discrete attributes	322
11.4.2	Pearson's correlation similarity	322
11.4.3	Spearman's correlation similarity	323
11.4.4	Cosine similarity	323
11.5	Missing attribute values	324
11.6	Conclusion	325
11.7	Further readings	325
	References	326
12	<i>k</i> -Centers clustering	328
12.1	Introduction	328
12.1.1	Basic principle	328
12.1.2	(Dis)similarity measures	329
12.2	Algorithm scheme	330
12.2.1	Initialization	331
12.2.2	Stop criteria	331
12.2.3	Cluster formation	331
12.2.4	Implicit cluster modeling	332
12.2.5	Instantiations	332
12.3	<i>k</i> -Means	334
12.3.1	Center adjustment	335
12.3.2	Minimizing dissimilarity to centers	336

12.4	Beyond means	338
12.4.1	k -Medians	338
12.4.2	k -Medoids	339
12.5	Beyond (fixed) k	342
12.5.1	Multiple runs	343
12.5.2	Adaptive k -centers	343
12.6	Explicit cluster modeling	343
12.7	Conclusion	345
12.8	Further readings	345
	References	347
13	Hierarchical clustering	349
13.1	Introduction	349
13.1.1	Basic approaches	349
13.1.2	(Dis)similarity measures	349
13.2	Cluster hierarchies	351
13.2.1	Motivation	351
13.2.2	Model representation	352
13.3	Agglomerative clustering	353
13.3.1	Algorithm scheme	353
13.3.2	Cluster linkage	356
13.4	Divisive clustering	361
13.4.1	Algorithm scheme	361
13.4.2	Wrapping a flat clustering algorithm	361
13.4.3	Stop criteria	362
13.5	Hierarchical clustering visualization	364
13.6	Hierarchical clustering prediction	366
13.6.1	Cutting cluster hierarchies	366
13.6.2	Cluster membership assignment	368
13.7	Conclusion	369
13.8	Further readings	370
	References	371
14	Clustering model evaluation	373
14.1	Introduction	373
14.1.1	Dataset performance	373
14.1.2	Training performance	374
14.1.3	True performance	374
14.2	Per-cluster quality measures	376
14.2.1	Diameter	376
14.2.2	Separation	377
14.2.3	Isolation	378
14.2.4	Silhouette width	379
14.2.5	Davies–Bouldin index	382
14.3	Overall quality measures	385
14.3.1	Dunn index	386
14.3.2	Average Davies–Bouldin index	387

14.3.3	C index	388
14.3.4	Average silhouette width	389
14.3.5	Loglikelihood	390
14.4	External quality measures	393
14.4.1	Misclassification error	393
14.4.2	Rand index	394
14.4.3	General relationship detection measures	396
14.5	Using quality measures	397
14.6	Conclusion	398
14.7	Further readings	398
	References	399
 Part V Getting Better Models		 401
15	Model ensembles	403
15.1	Introduction	403
15.2	Model committees	404
15.3	Base models	406
15.3.1	Different training sets	406
15.3.2	Different algorithms	412
15.3.3	Different parameter setups	412
15.3.4	Algorithm randomization	412
15.3.5	Base model diversity	418
15.4	Model aggregation	420
15.4.1	Voting/Averaging	420
15.4.2	Probability averaging	422
15.4.3	Weighted voting/averaging	424
15.4.4	Using as attributes	427
15.5	Specific ensemble modeling algorithms	431
15.5.1	Bagging	431
15.5.2	Stacking	433
15.5.3	Boosting	433
15.5.4	Random forest	443
15.5.5	Random Naïve Bayes	446
15.6	Quality of ensemble predictions	448
15.7	Conclusion	449
15.8	Further readings	450
	References	451
16	Kernel methods	454
16.1	Introduction	454
16.2	Support vector machines	457
16.2.1	Classification margin	457
16.2.2	Maximum-margin hyperplane	460
16.2.3	Primal form	460
16.2.4	Dual form	464

16.2.5	Soft margin	468
16.3	Support vector regression	473
16.3.1	Regression tube	474
16.3.2	Primal form	475
16.3.3	Dual form	475
16.4	Kernel trick	482
16.5	Kernel functions	484
16.5.1	Linear kernel	485
16.5.2	Polynomial kernel	485
16.5.3	Radial kernel	485
16.5.4	Sigmoid kernel	486
16.6	Kernel prediction	487
16.7	Kernel-based algorithms	489
16.7.1	Kernel-based SVM	489
16.7.2	Kernel-based SVR	492
16.8	Conclusion	494
16.9	Further readings	495
	References	496
17	Attribute transformation	498
17.1	Introduction	498
17.2	Attribute transformation task	499
17.2.1	Target task	499
17.2.2	Target attribute	500
17.2.3	Transformed attribute	500
17.2.4	Training set	500
17.2.5	Modeling transformations	500
17.2.6	Nonmodeling transformations	503
17.3	Simple transformations	504
17.3.1	Standardization	504
17.3.2	Normalization	505
17.3.3	Aggregation	506
17.3.4	Imputation	507
17.3.5	Binary encoding	508
17.4	Multiclass encoding	510
17.4.1	Encoding and decoding functions	511
17.4.2	1-ok- k encoding	514
17.4.3	Error-correcting encoding	515
17.4.4	Effects of multiclass encoding	519
17.5	Conclusion	521
17.6	Further readings	521
	References	522
18	Discretization	524
18.1	Introduction	524
18.2	Discretization task	525

18.2.1	Motivation	525
18.2.2	Task definition	526
18.2.3	Discretization as modeling	527
18.2.4	Discretization quality	529
18.3	Unsupervised discretization	530
18.3.1	Equal-width intervals	530
18.3.2	Equal-frequency intervals	531
18.3.3	Nonmodeling discretization	532
18.4	Supervised discretization	533
18.4.1	Pure-class discretization	533
18.4.2	Bottom-up discretization	535
18.4.3	Top-down discretization	546
18.5	Effects of discretization	551
18.6	Conclusion	553
18.7	Further readings	553
	References	556
19	Attribute selection	558
19.1	Introduction	558
19.2	Attribute selection task	559
19.2.1	Motivation	559
19.2.2	Task definition	560
19.2.3	Algorithms	561
19.3	Attribute subset search	562
19.3.1	Search task	562
19.3.2	Initial state	563
19.3.3	Search operators	564
19.3.4	State selection	564
19.3.5	Stop criteria	565
19.4	Attribute selection filters	568
19.4.1	Simple statistical filters	568
19.4.2	Correlation-based filters	571
19.4.3	Consistency-based filters	575
19.4.4	RELIEF	577
19.4.5	Random forest	584
19.4.6	Cutoff criteria	585
19.4.7	Filter-driven search	586
19.5	Attribute selection wrappers	588
19.5.1	Subset evaluation	588
19.5.2	Wrapper attribute selection	591
19.6	Effects of attribute selection	593
19.7	Conclusion	598
19.8	Further readings	599
	References	600

20	Case studies	602
20.1	Introduction	602
20.1.1	Datasets	603
20.1.2	Packages	603
20.1.3	Auxiliary functions	603
20.2	Census income	605
20.2.1	Data loading and preprocessing	606
20.2.2	Default model	608
20.2.3	Incorporating misclassification costs	610
20.2.4	Pruning	616
20.2.5	Attribute selection	624
20.2.6	Final models	628
20.3	Communities and crime	631
20.3.1	Data loading	632
20.3.2	Data quality	632
20.3.3	Regression trees	634
20.3.4	Linear models	635
20.3.5	Attribute selection	636
20.3.6	Piecewise-linear models	639
20.4	Cover type	640
20.4.1	Data loading and preprocessing	640
20.4.2	Class imbalance	641
20.4.3	Decision trees	641
20.4.4	Class rebalancing	644
20.4.5	Multiclass encoding	647
20.4.6	Final classification models	649
20.4.7	Clustering	650
20.5	Conclusion	654
20.6	Further readings	655
	References	655
	Closing	657
A	Notation	659
A.1	Attribute values	659
A.2	Data subsets	659
A.3	Probabilities	660
B	R packages	661
B.1	CRAN packages	661
B.2	DMR packages	662
B.3	Installing packages	663
	References	664
C	Datasets	666
	Index	667