

Contents

Acknowledgments	xxiii
List of Figures	xxv
List of Algorithms	xxxi
List of Boxes	xxxiii
1 Introduction	1
1.1 Motivation	1
1.2 Structured Probabilistic Models	2
1.2.1 Probabilistic Graphical Models	3
1.2.2 Representation, Inference, Learning	5
1.3 Overview and Roadmap	6
1.3.1 Overview of Chapters	6
1.3.2 Reader's Guide	9
1.3.3 Connection to Other Disciplines	11
1.4 Historical Notes	12
2 Foundations	15
2.1 Probability Theory	15
2.1.1 Probability Distributions	15
2.1.2 Basic Concepts in Probability	18
2.1.3 Random Variables and Joint Distributions	19
2.1.4 Independence and Conditional Independence	23
2.1.5 Querying a Distribution	25
2.1.6 Continuous Spaces	27
2.1.7 Expectation and Variance	31
2.2 Graphs	34
2.2.1 Nodes and Edges	34
2.2.2 Subgraphs	35
2.2.3 Paths and Trails	36

- 2.2.4 Cycles and Loops 36
- 2.3 Relevant Literature 39
- 2.4 Exercises 39

Representation 43

The Bayesian Network Representation 45

- 3.1 Exploiting Independence Properties 45
 - 3.1.1 Independent Random Variables 45
 - 3.1.2 The Conditional Parameterization 46
 - 3.1.3 The Naive Bayes Model 48
- 3.2 Bayesian Networks 51
 - 3.2.1 The Student Example Revisited 52
 - 3.2.2 Basic Independencies in Bayesian Networks 56
 - 3.2.3 Graphs and Distributions 60
- 3.3 Independencies in Graphs 68
 - 3.3.1 D-separation 69
 - 3.3.2 Soundness and Completeness 72
 - 3.3.3 An Algorithm for d-Separation 74
 - 3.3.4 I-Equivalence 76
- 3.4 From Distributions to Graphs 78
 - 3.4.1 Minimal I-Maps 79
 - 3.4.2 Perfect Maps 81
 - 3.4.3 Finding Perfect Maps * 83
- 3.5 Summary 92
- 3.6 Relevant Literature 93
- 3.7 Exercises 96

Undirected Graphical Models 103

- 4.1 The Misconception Example 103
- 4.2 Parameterization 106
 - 4.2.1 Factors 106
 - 4.2.2 Gibbs Distributions and Markov Networks 108
 - 4.2.3 Reduced Markov Networks 110
- 4.3 Markov Network Independencies 114
 - 4.3.1 Basic Independencies 114
 - 4.3.2 Independencies Revisited 117
 - 4.3.3 From Distributions to Graphs 120
- 4.4 Parameterization Revisited 122
 - 4.4.1 Finer-Grained Parameterization 123
 - 4.4.2 Overparameterization 128
- 4.5 Bayesian Networks and Markov Networks 134
 - 4.5.1 From Bayesian Networks to Markov Networks 134
 - 4.5.2 From Markov Networks to Bayesian Networks 137

- 4.5.3 Chordal Graphs 139
- 4.6 Partially Directed Models 142
 - 4.6.1 Conditional Random Fields 142
 - 4.6.2 Chain Graph Models * 148
- 4.7 Summary and Discussion 151
- 4.8 Relevant Literature 152
- 4.9 Exercises 153

5 *Local Probabilistic Models* 157

- 5.1 Tabular CPDs 157
- 5.2 Deterministic CPDs 158
 - 5.2.1 Representation 158
 - 5.2.2 Independencies 159
- 5.3 Context-Specific CPDs 162
 - 5.3.1 Representation 162
 - 5.3.2 Independencies 171
- 5.4 Independence of Causal Influence 175
 - 5.4.1 The Noisy-Or Model 175
 - 5.4.2 Generalized Linear Models 178
 - 5.4.3 The General Formulation 182
 - 5.4.4 Independencies 184
- 5.5 Continuous Variables 185
 - 5.5.1 Hybrid Models 189
- 5.6 Conditional Bayesian Networks 191
- 5.7 Summary 193
- 5.8 Relevant Literature 194
- 5.9 Exercises 195

6 *Template-Based Representations* 199

- 6.1 Introduction 199
- 6.2 Temporal Models 200
 - 6.2.1 Basic Assumptions 201
 - 6.2.2 Dynamic Bayesian Networks 202
 - 6.2.3 State-Observation Models 207
- 6.3 Template Variables and Template Factors 212
- 6.4 Directed Probabilistic Models for Object-Relational Domains 216
 - 6.4.1 Plate Models 216
 - 6.4.2 Probabilistic Relational Models 222
- 6.5 Undirected Representation 228
- 6.6 Structural Uncertainty * 232
 - 6.6.1 Relational Uncertainty 233
 - 6.6.2 Object Uncertainty 235
- 6.7 Summary 240
- 6.8 Relevant Literature 242
- 6.9 Exercises 243

7	Gaussian Network Models	247
7.1	Multivariate Gaussians	247
7.1.1	Basic Parameterization	247
7.1.2	Operations on Gaussians	249
7.1.3	Independencies in Gaussians	250
7.2	Gaussian Bayesian Networks	251
7.3	Gaussian Markov Random Fields	254
7.4	Summary	257
7.5	Relevant Literature	258
7.6	Exercises	258
8	The Exponential Family	261
8.1	Introduction	261
8.2	Exponential Families	261
8.2.1	Linear Exponential Families	263
8.3	Factored Exponential Families	266
8.3.1	Product Distributions	266
8.3.2	Bayesian Networks	267
8.4	Entropy and Relative Entropy	269
8.4.1	Entropy	269
8.4.2	Relative Entropy	272
8.5	Projections	273
8.5.1	Comparison	274
8.5.2	M-Projections	277
8.5.3	I-Projections	282
8.6	Summary	282
8.7	Relevant Literature	283
8.8	Exercises	283
II	Inference	285
9	Exact Inference: Variable Elimination	287
9.1	Analysis of Complexity	288
9.1.1	Analysis of Exact Inference	288
9.1.2	Analysis of Approximate Inference	290
9.2	Variable Elimination: The Basic Ideas	292
9.3	Variable Elimination	296
9.3.1	Basic Elimination	297
9.3.2	Dealing with Evidence	303
9.4	Complexity and Graph Structure: Variable Elimination	305
9.4.1	Simple Analysis	306
9.4.2	Graph-Theoretic Analysis	306
9.4.3	Finding Elimination Orderings *	310
9.5	Conditioning *	315

9.5.1	The Conditioning Algorithm	315
9.5.2	Conditioning and Variable Elimination	318
9.5.3	Graph-Theoretic Analysis	322
9.5.4	Improved Conditioning	323
9.6	Inference with Structured CPDs *	325
9.6.1	Independence of Causal Influence	325
9.6.2	Context-Specific Independence	329
9.6.3	Discussion	335
9.7	Summary and Discussion	336
9.8	Relevant Literature	337
9.9	Exercises	338
10	Exact Inference: Clique Trees	345
10.1	Variable Elimination and Clique Trees	345
10.1.1	Cluster Graphs	346
10.1.2	Clique Trees	346
10.2	Message Passing: Sum Product	348
10.2.1	Variable Elimination in a Clique Tree	349
10.2.2	Clique Tree Calibration	355
10.2.3	A Calibrated Clique Tree as a Distribution	361
10.3	Message Passing: Belief Update	364
10.3.1	Message Passing with Division	364
10.3.2	Equivalence of Sum-Product and Belief Update Messages	
10.3.3	Answering Queries	369
10.4	Constructing a Clique Tree	372
10.4.1	Clique Trees from Variable Elimination	372
10.4.2	Clique Trees from Chordal Graphs	374
10.5	Summary	376
10.6	Relevant Literature	377
10.7	Exercises	378
II	Inference as Optimization	381
11.1	Introduction	381
11.1.1	Exact Inference Revisited *	382
11.1.2	The Energy Functional	384
11.1.3	Optimizing the Energy Functional	386
11.2	Exact Inference as Optimization	386
11.2.1	Fixed-Point Characterization	388
11.2.2	Inference as Optimization	390
11.3	Propagation-Based Approximation	391
11.3.1	A Simple Example	391
11.3.2	Cluster-Graph Belief Propagation	396
11.3.3	Properties of Cluster-Graph Belief Propagation	399
11.3.4	Analyzing Convergence *	401
11.3.5	Constructing Cluster Graphs	404

11.3.6	Variational Analysis	411
11.3.7	Other Entropy Approximations *	414
11.3.8	Discussion	428
11.4	Propagation with Approximate Messages *	430
11.4.1	Factorized Messages	431
11.4.2	Approximate Message Computation	433
11.4.3	Inference with Approximate Messages	436
11.4.4	Expectation Propagation	442
11.4.5	Variational Analysis	445
11.4.6	Discussion	448
11.5	Structured Variational Approximations	448
11.5.1	The Mean Field Approximation	449
11.5.2	Structured Approximations	456
11.5.3	Local Variational Methods *	469
11.6	Summary and Discussion	473
11.7	Relevant Literature	475
11.8	Exercises	477

12 Particle-Based Approximate Inference 487

12.1	Forward Sampling	488
12.1.1	Sampling from a Bayesian Network	488
12.1.2	Analysis of Error	490
12.1.3	Conditional Probability Queries	491
12.2	Likelihood Weighting and Importance Sampling	492
12.2.1	Likelihood Weighting: Intuition	492
12.2.2	Importance Sampling	494
12.2.3	Importance Sampling for Bayesian Networks	498
12.2.4	Importance Sampling Revisited	504
12.3	Markov Chain Monte Carlo Methods	505
12.3.1	Gibbs Sampling Algorithm	505
12.3.2	Markov Chains	507
12.3.3	Gibbs Sampling Revisited	512
12.3.4	A Broader Class of Markov Chains *	515
12.3.5	Using a Markov Chain	518
12.4	Collapsed Particles	526
12.4.1	Collapsed Likelihood Weighting *	527
12.4.2	Collapsed MCMC	531
12.5	Deterministic Search Methods *	536
12.6	Summary	540
12.7	Relevant Literature	541
12.8	Exercises	544

13 MAP Inference 551

13.1	Overview	551
13.1.1	Computational Complexity	551

13.1.2	Overview of Solution Methods	552
13.2	Variable Elimination for (Marginal) MAP	554
13.2.1	Max-Product Variable Elimination	554
13.2.2	Finding the Most Probable Assignment	556
13.2.3	Variable Elimination for Marginal MAP *	559
13.3	Max-Product in Clique Trees	562
13.3.1	Computing Max-Marginals	562
13.3.2	Message Passing as Reparameterization	564
13.3.3	Decoding Max-Marginals	565
13.4	Max-Product Belief Propagation in Loopy Cluster Graphs	567
13.4.1	Standard Max-Product Message Passing	567
13.4.2	Max-Product BP with Counting Numbers *	572
13.4.3	Discussion	575
13.5	MAP as a Linear Optimization Problem *	577
13.5.1	The Integer Program Formulation	577
13.5.2	Linear Programming Relaxation	579
13.5.3	Low-Temperature Limits	581
13.6	Using Graph Cuts for MAP	588
13.6.1	Inference Using Graph Cuts	588
13.6.2	Nonbinary Variables	592
13.7	Local Search Algorithms *	595
13.8	Summary	597
13.9	Relevant Literature	598
13.10	Exercises	601

14 Inference in Hybrid Networks 605

14.1	Introduction	605
14.1.1	Challenges	605
14.1.2	Discretization	606
14.1.3	Overview	607
14.2	Variable Elimination in Gaussian Networks	608
14.2.1	Canonical Forms	609
14.2.2	Sum-Product Algorithms	611
14.2.3	Gaussian Belief Propagation	612
14.3	Hybrid Networks	615
14.3.1	The Difficulties	615
14.3.2	Factor Operations for Hybrid Gaussian Networks	618
14.3.3	EP for CLG Networks	621
14.3.4	An "Exact" CLG Algorithm *	626
14.4	Nonlinear Dependencies	630
14.4.1	Linearization	631
14.4.2	Expectation Propagation with Gaussian Approximation	637
14.5	Particle-Based Approximation Methods	642
14.5.1	Sampling in Continuous Spaces	642
14.5.2	Forward Sampling in Bayesian Networks	643

14.5.3	MCMC Methods	644
14.5.4	Collapsed Particles	645
14.5.5	Nonparametric Message Passing	646
14.6	Summary and Discussion	646
14.7	Relevant Literature	647
14.8	Exercises	649
15	<i>Inference in Temporal Models</i>	651
15.1	Inference Tasks	652
15.2	Exact Inference	653
15.2.1	Filtering in State-Observation Models	653
15.2.2	Filtering as Clique Tree Propagation	654
15.2.3	Clique Tree Inference in DBNs	655
15.2.4	Entanglement	656
15.3	Approximate Inference	661
15.3.1	Key Ideas	661
15.3.2	Factored Belief State Methods	663
15.3.3	Particle Filtering	665
15.3.4	Deterministic Search Techniques	675
15.4	Hybrid DBNs	675
15.4.1	Continuous Models	676
15.4.2	Hybrid Models	683
15.5	Summary	688
15.6	Relevant Literature	690
15.7	Exercises	692

III Learning 695

16	<i>Learning Graphical Models: Overview</i>	697
16.1	Motivation	697
16.2	Goals of Learning	698
16.2.1	Density Estimation	698
16.2.2	Specific Prediction Tasks	700
16.2.3	Knowledge Discovery	701
16.3	Learning as Optimization	702
16.3.1	Empirical Risk and Overfitting	703
16.3.2	Discriminative versus Generative Training	709
16.4	Learning Tasks	711
16.4.1	Model Constraints	712
16.4.2	Data Observability	712
16.4.3	Taxonomy of Learning Tasks	714
16.5	Relevant Literature	715
17	<i>Parameter Estimation</i>	717
17.1	Maximum Likelihood Estimation	717

17.1.1	The Thumbtack Example	717
17.1.2	The Maximum Likelihood Principle	720
17.2	MLE for Bayesian Networks	722
17.2.1	A Simple Example	723
17.2.2	Global Likelihood Decomposition	724
17.2.3	Table-CPDs	725
17.2.4	Gaussian Bayesian Networks *	728
17.2.5	Maximum Likelihood Estimation as M-Projection *	731
17.3	Bayesian Parameter Estimation	733
17.3.1	The Thumbtack Example Revisited	733
17.3.2	Priors and Posteriors	737
17.4	Bayesian Parameter Estimation in Bayesian Networks	741
17.4.1	Parameter Independence and Global Decomposition	742
17.4.2	Local Decomposition	746
17.4.3	Priors for Bayesian Network Learning	748
17.4.4	MAP Estimation *	751
17.5	Learning Models with Shared Parameters	754
17.5.1	Global Parameter Sharing	755
17.5.2	Local Parameter Sharing	760
17.5.3	Bayesian Inference with Shared Parameters	762
17.5.4	Hierarchical Priors *	763
17.6	Generalization Analysis *	769
17.6.1	Asymptotic Analysis	769
17.6.2	PAC-Bounds	770
17.7	Summary	776
17.8	Relevant Literature	777
17.9	Exercises	778

18 Structure Learning in Bayesian Networks 783

18.1	Introduction	783
18.1.1	Problem Definition	783
18.1.2	Overview of Methods	785
18.2	Constraint-Based Approaches	786
18.2.1	General Framework	786
18.2.2	Independence Tests	787
18.3	Structure Scores	790
18.3.1	Likelihood Scores	791
18.3.2	Bayesian Score	794
18.3.3	Marginal Likelihood for a Single Variable	797
18.3.4	Bayesian Score for Bayesian Networks	799
18.3.5	Understanding the Bayesian Score	801
18.3.6	Priors	804
18.3.7	Score Equivalence *	807
18.4	Structure Search	807
18.4.1	Learning Tree-Structured Networks	808

18.4.2	Known Order	809
18.4.3	General Graphs	811
18.4.4	Learning with Equivalence Classes *	821
18.5	Bayesian Model Averaging *	824
18.5.1	Basic Theory	824
18.5.2	Model Averaging Given an Order	826
18.5.3	The General Case	828
18.6	Learning Models with Additional Structure	832
18.6.1	Learning with Local Structure	833
18.6.2	Learning Template Models	837
18.7	Summary and Discussion	838
18.8	Relevant Literature	840
18.9	Exercises	843
19	Partially Observed Data	849
19.1	Foundations	849
19.1.1	Likelihood of Data and Observation Models	849
19.1.2	Decoupling of Observation Mechanism	853
19.1.3	The Likelihood Function	856
19.1.4	Identifiability	860
19.2	Parameter Estimation	862
19.2.1	Gradient Ascent	863
19.2.2	Expectation Maximization (EM)	868
19.2.3	Comparison: Gradient Ascent versus EM	887
19.2.4	Approximate Inference *	893
19.3	Bayesian Learning with Incomplete Data *	897
19.3.1	Overview	897
19.3.2	MCMC Sampling	899
19.3.3	Variational Bayesian Learning	904
19.4	Structure Learning	908
19.4.1	Scoring Structures	909
19.4.2	Structure Search	917
19.4.3	Structural EM	920
19.5	Learning Models with Hidden Variables	925
19.5.1	Information Content of Hidden Variables	926
19.5.2	Determining the Cardinality	928
19.5.3	Introducing Hidden Variables	930
19.6	Summary	933
19.7	Relevant Literature	934
19.8	Exercises	935
20	Learning Undirected Models	943
20.1	Overview	943
20.2	The Likelihood Function	944
20.2.1	An Example	944

20.2.2	Form of the Likelihood Function	946
20.2.3	Properties of the Likelihood Function	947
20.3	Maximum (Conditional) Likelihood Parameter Estimation	949
20.3.1	Maximum Likelihood Estimation	949
20.3.2	Conditionally Trained Models	950
20.3.3	Learning with Missing Data	954
20.3.4	Maximum Entropy and Maximum Likelihood *	956
20.4	Parameter Priors and Regularization	958
20.4.1	Local Priors	958
20.4.2	Global Priors	961
20.5	Learning with Approximate Inference	961
20.5.1	Belief Propagation	962
20.5.2	MAP-Based Learning *	967
20.6	Alternative Objectives	969
20.6.1	Pseudolikelihood and Its Generalizations	970
20.6.2	Contrastive Optimization Criteria	974
20.7	Structure Learning	978
20.7.1	Structure Learning Using Independence Tests	979
20.7.2	Score-Based Learning: Hypothesis Spaces	981
20.7.3	Objective Functions	982
20.7.4	Optimization Task	985
20.7.5	Evaluating Changes to the Model	992
20.8	Summary	996
20.9	Relevant Literature	998
20.10	Exercises	1001

IV Actions and Decisions 1007

21	Causality	1009
21.1	Motivation and Overview	1009
21.1.1	Conditioning and Intervention	1009
21.1.2	Correlation and Causation	1012
21.2	Causal Models	1014
21.3	Structural Causal Identifiability	1017
21.3.1	Query Simplification Rules	1017
21.3.2	Iterated Query Simplification	1020
21.4	Mechanisms and Response Variables *	1026
21.5	Partial Identifiability in Functional Causal Models *	1031
21.6	Counterfactual Queries *	1034
21.6.1	Twinned Networks	1034
21.6.2	Bounds on Counterfactual Queries	1037
21.7	Learning Causal Models	1040
21.7.1	Learning Causal Models without Confounding Factors	1041
21.7.2	Learning from Interventional Data	1044

21.7.3	Dealing with Latent Variables ★	1048
21.7.4	Learning Functional Causal Models ★	1051
21.8	Summary	1053
21.9	Relevant Literature	1054
21.10	Exercises	1055
22	Utilities and Decisions	1059
22.1	Foundations: Maximizing Expected Utility	1059
22.1.1	Decision Making Under Uncertainty	1059
22.1.2	Theoretical Justification ★	1062
22.2	Utility Curves	1064
22.2.1	Utility of Money	1065
22.2.2	Attitudes Toward Risk	1066
22.2.3	Rationality	1067
22.3	Utility Elicitation	1068
22.3.1	Utility Elicitation Procedures	1068
22.3.2	Utility of Human Life	1069
22.4	Utilities of Complex Outcomes	1071
22.4.1	Preference and Utility Independence ★	1071
22.4.2	Additive Independence Properties	1074
22.5	Summary	1081
22.6	Relevant Literature	1082
22.7	Exercises	1084
23	Structured Decision Problems	1085
23.1	Decision Trees	1085
23.1.1	Representation	1085
23.1.2	Backward Induction Algorithm	1087
23.2	Influence Diagrams	1088
23.2.1	Basic Representation	1089
23.2.2	Decision Rules	1090
23.2.3	Time and Recall	1092
23.2.4	Semantics and Optimality Criterion	1093
23.3	Backward Induction in Influence Diagrams	1095
23.3.1	Decision Trees for Influence Diagrams	1096
23.3.2	Sum-Max-Sum Rule	1098
23.4	Computing Expected Utilities	1100
23.4.1	Simple Variable Elimination	1100
23.4.2	Multiple Utility Variables: Simple Approaches	1102
23.4.3	Generalized Variable Elimination ★	1103
23.5	Optimization in Influence Diagrams	1107
23.5.1	Optimizing a Single Decision Rule	1107
23.5.2	Iterated Optimization Algorithm	1108
23.5.3	Strategic Relevance and Global Optimality ★	1110
23.6	Ignoring Irrelevant Information ★	1119

23.7	Value of Information	1121
23.7.1	Single Observations	1122
23.7.2	Multiple Observations	1124
23.8	Summary	1126
23.9	Relevant Literature	1127
23.10	Exercises	1130

24 Epilogue 1133

A Background Material 1137

A.1	Information Theory	1137
A.1.1	Compression and Entropy	1137
A.1.2	Conditional Entropy and Information	1139
A.1.3	Relative Entropy and Distances Between Distributions	1141
A.2	Convergence Bounds	1143
A.2.1	Central Limit Theorem	1144
A.2.2	Convergence Bounds	1145
A.3	Algorithms and Algorithmic Complexity	1146
A.3.1	Basic Graph Algorithms	1146
A.3.2	Analysis of Algorithmic Complexity	1147
A.3.3	Dynamic Programming	1149
A.3.4	Complexity Theory	1150
A.4	Combinatorial Optimization and Search	1154
A.4.1	Optimization Problems	1154
A.4.2	Local Search	1154
A.4.3	Branch and Bound Search	1160
A.5	Continuous Optimization	1161
A.5.1	Characterizing Optima of a Continuous Function	1161
A.5.2	Gradient Ascent Methods	1163
A.5.3	Constrained Optimization	1167
A.5.4	Convex Duality	1171

Bibliography 1173

Notation Index 1211

Subject Index 1215