

Contents

List of Tables	xv
List of Figures	xvii
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	4
1.3 Overview and Roadmap	5
1.3.1 Overview of Chapters	5
1.3.2 Reader's Guide	7
1.3.3 Connection to Other Disciplines	8
1.4 Historical Notes	9
2 Foundations	11
2.1 Probability Theory	11
2.1.1 Probability Distributions	11
2.1.2 Basic Concepts in Probability	13
2.1.3 Random Variables	15
2.1.4 Continuous Spaces	19
2.1.5 Expectation and Variance	23
2.2 Information Theory	25
2.2.1 Compression and Entropy	25
2.2.2 Conditional Entropy and Information	26
2.2.3 Relative Entropy & Distances Between Distributions	27
2.3 Graphs	30
2.3.1 Nodes and Edges	30
2.3.2 Paths	31
2.3.3 Cycles and Loops	32
2.3.4 Subgraphs	33
2.3.5 Basic Graph Algorithms	33
2.4 References	34
3 The Bayesian Network Representation	39
3.1 Exploiting Independence Properties	39
3.1.1 Independent Random Variables	39
3.1.2 The Conditional Parameterization	40
3.1.3 The Naive Bayes Model	41
3.2 Bayesian Networks	44

3.2.1	The Student Example Revisited	45
3.2.2	Basic Independences in Bayesian Networks	48
3.2.3	Graphs and Distributions	53
3.3	Independencies in Graphs	60
3.3.1	d-separation	60
3.3.2	Soundness and Completeness	63
3.3.3	An Algorithm for d-Separation	64
3.3.4	I-Equivalence	66
3.4	From Distributions to Graphs	68
3.4.1	Minimal I-Maps	68
3.4.2	Perfect Maps	70
3.4.3	Finding Perfect Maps	72
3.5	Summary	79
4	<i>Undirected Graphical Models</i>	87
4.1	The Misconception Example	87
4.2	Parameterization	88
4.2.1	Factors	89
4.2.2	Gibbs Distributions and Markov Networks	90
4.2.3	Reduced Markov Networks	92
4.3	Markov Network Independencies	94
4.3.1	Basic Independencies	94
4.3.2	Independencies Revisited	97
4.3.3	From Distributions to Graphs	99
4.4	Parameterization Revisited	101
4.4.1	Finer-Grained Parameterization	101
4.4.2	Over-Parameterization	105
4.5	Bayesian Networks and Markov Networks	111
4.5.1	From Bayesian Networks to Markov Networks	111
4.5.2	From Markov Networks to Bayesian Networks	114
4.5.3	Chordal Graphs	116
4.6	Partially Directed Models	118
4.6.1	Conditional Random Fields	118
4.6.2	Chain Graphs	121
4.7	Summary	124
4.8	Historical Notes	124
4.9	Exercises	126
5	<i>Local probabilistic models</i>	129
5.1	Tabular CPDs	129
5.2	Deterministic CPDs	130
5.2.1	Representation	130
5.2.2	Independencies	131
5.3	Context-Specific CPDs	133
5.3.1	Representation	133
5.3.2	Independencies	140
5.4	Independence of Causal Influence	143
5.4.1	The Noisy-Or Model	143
5.4.2	Generalized Linear Models	146
5.4.3	The General Formulation	150
5.4.4	Independencies	151
5.5	Continuous Variables	152

5.5.1	Hybrid Models	155
5.6	Encapsulated Bayesian Networks	156
5.7	Summary	158
6	<i>Gaussian Network Models</i>	163
6.1	Multi-Variate Gaussians	163
6.1.1	Basic Parameterization	163
6.1.2	Operations on Gaussians	164
6.1.3	Independencies in Gaussians	166
6.2	Gaussian Bayesian Networks	166
6.3	Gaussian Markov Random Fields	169
6.4	Summary	171
6.5	Historical Notes	172
7	<i>Template-Based Representations</i>	175
7.1	Motivation	175
7.2	Temporal Models	176
7.2.1	Basic Assumptions	176
7.2.2	Dynamic Bayesian Networks	177
7.2.3	State-Observation Models	179
7.3	Object-Relational Models	183
7	<i>The Exponential Family</i>	175
7.1	Introduction	175
7.2	Exponential Families	175
7.2.1	Linear Exponential Families	177
7.3	Factored Exponential Families	179
7.3.1	Product Distributions	179
7.3.2	Bayesian Networks	180
7.4	Entropy and Relative Entropy	182
7.4.1	Entropy	182
7.4.2	Relative Entropy	185
7.5	Projections	185
7.5.1	Comparison	186
7.5.2	M-Projections	188
7.5.3	I-Projections	192
7.6	Summary	193
8	<i>Exact Inference: Variable Elimination</i>	195
8.1	Analysis of Complexity	200
8.2	Variable Elimination: The Basic Ideas	201
8.2.1	A Simple Chain	202
8.2.2	Summary	206
8.3	Variable Elimination	206
8.3.1	Basic Elimination	206
8.3.2	Dealing with evidence	212
8.4	Complexity and Graph Structure: Variable Elimination	214
8.4.1	Simple Analysis	214
8.4.2	Graph-Theoretic Analysis	214
8.4.3	Finding Elimination Orderings	218
8.5	Conditioning	221
8.5.1	The Conditioning Algorithm	223
8.5.2	Conditioning and Variable Elimination	224

8.5.3	Graph-Theoretic Analysis	226	
8.5.4	Improved Conditioning	228	
8.6	Inference with Structured CPDs	229	
8.6.1	Independence of Causal Influence	229	
8.6.2	Context-Specific Independence	232	
8.7	Summary	240	
8.8	References	240	
9	<i>Exact Inference: Clique Trees</i>	245	
9.1	Variable Elimination and Clique Trees	245	
9.1.1	Cluster Graphs	245	
9.1.2	Clique Trees	246	
9.2	Message Passing: Sum Product	248	
9.2.1	Variable Elimination in a Clique Tree	248	
9.2.2	Clique Tree Calibration	253	
9.2.3	A Calibrated Clique Tree as a Distribution	255	
9.3	Message Passing: Belief Update	257	
9.3.1	Message Passing with Division	257	
9.3.2	Clique Tree Invariant	260	
9.3.3	Equivalence of Sum-Product and Belief-Update Messages	262	
9.3.4	Answering Queries	263	
9.4	Constructing a Clique Tree	266	
9.4.1	Clique Trees from Variable Elimination	266	
9.4.2	Clique Trees from Chordal Graphs	267	
9.5	Summary	268	
9.6	References	270	
10	<i>Global Approximate Inference: Inference as Optimization</i>	273	
10.1	Inference as Optimization	273	
10.1.1	Exact Inference as Optimization	274	
10.1.2	The Energy Functional	276	
10.1.3	Optimizing the Energy Functional	276	
10.2	Exact Inference as Optimization	277	
10.2.1	Fixed-Point Characterization	283	
10.2.2	Inference as Optimization	285	
10.3	Propagation-Based Approximation	286	
10.3.1	A Simple Example	286	
10.3.2	Generalized Belief Propagation	288	
10.3.3	Properties of Generalized Belief Propagation	290	
10.3.4	Analyzing Convergence	292	
10.3.5	Constructing Cluster Graphs	294	
10.3.6	Variational Analysis	299	
10.3.7	Other Free Energy Approximations	301	
10.3.8	Discussion	310	
10.4	Propagation with Approximate Messages	311	
10.4.1	Factorized Messages	312	
10.4.2	Approximate Message Computation	314	
10.4.3	Inference with Approximate Messages	316	
10.4.4	Expectation Propagation	321	
10.4.5	Variational Analysis	324	
10.4.6	Discussion	326	
10.5	Structured Variational Approximations	327	

10.5.1	The Mean Field Approximation	327
10.5.2	Markov Network Approximations	333
10.5.3	Local Variational Methods	343
10.6	Summary	346
10.7	Historical Notes	347
11	<i>Particle-Based Approximate Inference</i>	355
11.1	Approximate Inference and its Computational Complexity	356
11.1.1	Evaluation Metrics for Approximate Inference	356
11.2	Forward Sampling	357
11.2.1	Sampling from a Bayesian Network	357
11.2.2	Analysis of Error	361
11.2.3	Conditional Probability Queries	362
11.3	Likelihood Weighting and Importance Sampling	363
11.3.1	Likelihood Weighting: Intuition	363
11.3.2	Importance Sampling	364
11.3.3	Importance Sampling for Bayesian Networks	367
11.3.4	Importance Sampling Revisited	373
11.4	Markov Chain Monte Carlo Methods	373
11.4.1	Gibbs Sampling Algorithm	373
11.4.2	Markov Chains	375
11.4.3	Gibbs Sampling Revisited	378
11.4.4	A Broader Class of Markov Chains	381
11.4.5	Using a Markov Chain	384
11.5	Deterministic Search Methods	388
11.5.1	Bounds	389
11.5.2	Selecting Instantiations	391
11.6	Distributional Particles	392
11.6.1	Distributional Likelihood Weighting	392
11.6.2	Distributional MCMC	395
11.6.3	Distributional Deterministic Search	397
11.7	Summary	397
11.8	Historical Notes	398
12	<i>MAP Inference</i>	403
12.1	Overview	403
12.1.1	MAP and Partial MAP Queries	403
12.1.2	Computational Complexity	405
12.1.3	Overview of Solution Methods	406
12.2	Variable Elimination for MAP Queries	407
12.2.1	Variable Elimination	407
12.2.2	Max-Product Variable Elimination	408
12.2.3	Finding the Most Probable Assignment	410
12.3	Max-Product Clique Trees	411
12.3.1	Computing Max Marginals	412
12.3.2	Message Passing as Recalibration	413
12.3.3	MAP via Max-Marginals	414
12.4	Max-Product Belief Propagation	416
12.4.1	Standard Max-Product Message Passing	416
12.4.2	Generalized Max-Product Message-Passing	419
12.4.3	Discussion	421
12.5	MAP as a Linear Optimization Problem	422

12.5.1	The Integer Program Formulation	422
12.5.2	Linear Programming Relaxation	424
12.5.3	Low-Temperature Limits	426
12.6	Graph Cuts	430
12.6.1	Inference Using Graph Cuts	430
12.6.2	Non-Binary Variables	432
12.6.3	Discussion	434
12.7	Partial MAP Queries	435
12.7.1	Variable Elimination for Partial MAP	435
12.7.2	Search Algorithms	437
12.8	Summary	443
12.9	Historical Notes	444

13 Inference in Temporal Models 451

13.1	Inference Tasks	451
13.2	Exact inference	452
13.2.1	Tracking in State-Observation Models	453
13.2.2	Tracking as Clique Tree Propagation	453
13.2.3	Clique Tree Inference in DBNs	454
13.2.4	Entanglement	455
13.3	Approximate Inference	458
13.3.1	Key Ideas	459
13.3.2	Factored Belief State Methods	460
13.3.3	Particle Filtering	461
13.4	Hybrid DBNs	469
13.4.1	Continuous Models	469
13.4.2	Hybrid Models	471
13.4.3	Particle Filtering for Hybrid DBNs	474
13.5	Summary	474
13.6	Historical Notes	475

14 Inference in Hybrid Networks 477

14.1	Introduction	477
14.1.1	The Basic Task	477
14.1.2	Difficulties	478
14.1.3	Discretization	480
14.1.4	Overview	481
14.2	Linear Gaussians	481
14.2.1	Canonical Forms	482
14.2.2	Exact Inference Algorithms	484
14.2.3	Gaussian Belief Propagation	485
14.3	Non-Linear Dependencies	485
14.3.1	Gaussian Approximation	485
14.3.2	Inference Algorithms	490
14.3.3	Context-Restricted Integration	492
14.3.4	Multiple Functions	492
14.3.5	Constraints on Elimination Ordering	493
14.3.6	Computing Messages	493
14.3.7	Ordering Constraints on Message Passing	494
14.4	CLG Models	494
14.4.1	Complexity Analysis	495
14.4.2	Lauritzen's CLG Algorithm	497

14.4.3	Non-linear CPDs	506
14.5	Particle-Based Approximation Methods	507
14.5.1	Full Particles	508
14.5.2	MCMC Methods	508
14.5.3	Distributional Particles	509
14.6	Global Approximation Methods	511
14.6.1	Loopy Propagation in Linear Gaussians	511
14.6.2	Expectation Propagation	511
14.7	Summary	511
14.8	Historical Notes	511
15	<i>Learning Graphical Models: Overview</i>	513
15.1	Goals of Learning	514
15.1.1	Density Estimation	514
15.1.2	Specific Prediction Tasks	515
15.1.3	Knowledge Discovery	516
15.2	Learning as Optimization	517
15.2.1	Empirical Loss and Overfitting	517
15.2.2	Discriminative versus Generative Training	523
15.3	Learning Tasks	524
15.3.1	Model constraints	525
15.3.2	Data observability	525
15.3.3	Taxonomy of Learning Tasks	526
15.4	Additional Readings	527
16	<i>Parameter Estimation</i>	529
16.1	Maximum Likelihood Estimation	529
16.1.1	The Thumbtack Example	529
16.1.2	The Maximum Likelihood Principle	531
16.2	MLE for Bayesian Networks	533
16.2.1	A Simple Example	533
16.2.2	Likelihood Decomposition	535
16.2.3	Table CPDs	536
16.2.4	Gaussian Bayesian Networks	537
16.2.5	Maximum Likelihood Estimation as M-Projection	540
16.3	Bayesian Parameter Estimation	541
16.3.1	The Thumbtack Example Revisited	541
16.3.2	Priors and Posteriors	545
16.4	Bayesian Estimation in Bayesian Networks	549
16.4.1	Parameter Independence and Global Decomposition	549
16.4.2	Local Decomposition	552
16.4.3	Priors for Bayesian Network Learning	554
16.4.4	MAP Estimation	556
16.5	Learning Models with Shared Parameters	559
16.6	Generalization Analysis	561
16.6.1	Asymptotic analysis	562
16.6.2	PAC-Bounds	562
16.7	Summary	567
16.8	References and Additional Readings	568
17	<i>Structure Learning in Bayesian Networks</i>	573
17.1	Introduction	573
17.2	Constraint Based Approaches	575

17.2.1	General Framework	575
17.2.2	Independence Tests	576
17.3	Structure Scores	579
17.3.1	Likelihood Scores	579
17.3.2	Bayesian Score	582
17.3.3	Marginal Likelihood for a Single Variable	585
17.3.4	Bayesian Score for Bayesian Networks	586
17.3.5	Understanding the Bayesian Score	588
17.3.6	Priors	590
17.3.7	Score Equivalence	593
17.4	Structure Search	593
17.4.1	Learning Tree-Structured Networks	594
17.4.2	Known Order	595
17.4.3	General Graphs	596
17.4.4	Learning with Equivalence Classes	604
17.5	Bayesian Model Averaging	607
17.5.1	Basic Theory	607
17.5.2	Model Averaging Given an Order	608
17.5.3	The General Case	610
17.6	Learning Local Structure	615
17.6.1	Scoring Networks with Local Structure	615
17.6.2	Search with Local Structure	616
17.7	Summary and Discussion	618
17.8	References	619
18 Partially Observed Data		627
18.1	Foundations	627
18.1.1	Likelihood of Data and Observation Models	627
18.1.2	Decoupling of Observation Mechanism	630
18.1.3	The Likelihood Function	632
18.1.4	Identifiability	635
18.2	Parameter Estimation	637
18.2.1	Gradient ascent	637
18.2.2	Expectation Maximization (EM)	645
18.2.3	Comparison	658
18.2.4	Approximate Inference	662
18.3	Bayesian Inference with Incomplete Data	665
18.3.1	Foundations	665
18.3.2	MAP Estimation	666
18.3.3	Learning as Inference	667
18.3.4	MCMC Sampling	667
18.3.5	Variational Bayesian Learning	670
18.4	Structure Learning	674
18.4.1	Scoring Structures	674
18.4.2	Structure search	680
18.4.3	Structural EM	683
18.5	Detecting Hidden variables	687
18.5.1	Learning Hidden Variables	687
18.5.2	Signature Based Approaches	690
18.5.3	Initializing a Hidden Variable	691
18.5.4	Refining Hidden Variables	691
18.6	Summary	693

19 Learning Undirected Models 699

19.1	Overview	699
19.2	The Likelihood Function	700
19.2.1	An Example	700
19.2.2	Form of the Likelihood Function	701
19.2.3	Properties of the Likelihood Function	702
19.3	Maximum (Conditional) Likelihood Parameter Estimation	703
19.3.1	Maximum Likelihood Estimation	703
19.3.2	Conditionally-Trained Models	704
19.3.3	Learning with Missing Data	705
19.3.4	Maximum Entropy and Maximum Likelihood	707
19.4	Parameter Priors and Regularization	709
19.4.1	Local Priors	709
19.4.2	Global Priors	711
19.5	Learning with Approximate Inference	712
19.5.1	Belief Propagation	712
19.5.2	MAP-Based Learning	716
19.6	Alternative Objectives	717
19.6.1	Pseudo-Likelihood and its Generalizations	718
19.6.2	Contrastive Optimization Criteria	721
19.7	Structure Learning	725
19.7.1	Structure Learning Using Independence Tests	725
19.7.2	Score-Based Learning: Hypothesis Spaces	726
19.7.3	Objective Functions	727
19.7.4	Optimization Task	730
19.7.5	Evaluating Changes to the Model	735
19.8	Summary	738
19.9	References	740

20 Causality 747

20.1	Motivation and Overview	747
20.1.1	Conditioning and Intervention	747
20.1.2	Correlation and Causation	749
20.2	Causal Models	751
20.3	Structural Causal Identifiability	753
20.3.1	Query Simplification Rules	754
20.3.2	Iterated Query Simplification	755
20.4	Mechanisms and Response Variables	760
20.5	Partial Identifiability in Functional Causal Models	764
20.6	Counterfactual Queries	767
20.6.1	Twinned Networks	768
20.6.2	Bounds on Counterfactual Queries	770
20.7	Learning Causal Models	771
20.7.1	Learning Causal Models without Confounding Variables	772
20.7.2	Learning from Interventional Data	774
20.7.3	Dealing with Latent Variables	777
20.7.4	Learning Functional Causal Models	781
20.8	Summary	783

21 Utilities and Decisions 787

21.1	Foundations: Maximizing Expected Utility	787
------	--	-----

21.1.1	Decision Making Under Uncertainty	787
21.1.2	Theoretical Justification	789
21.2	Utility Curves	791
21.2.1	Utility of money	791
21.2.2	Attitudes towards risk	793
21.2.3	Rationality	794
21.3	Utilities of Complex Outcomes	795
21.3.1	Preference and Utility Independence	795
21.3.2	Additive Independence Properties	797
21.4	Utility Elicitation	803
21.4.1	Utility Elicitation Procedures	803
21.4.2	Utility of Human Life	803
21.5	Summary	805
22	<i>Structured Decision Problems</i>	809
22.1	Decision Trees	809
22.1.1	Representation	809
22.1.2	Backward Induction Algorithm	811
22.2	Influence Diagrams	811
22.2.1	Basic Representation	812
22.2.2	Decision Rules	813
22.2.3	Time and Recall	814
22.2.4	Semantics and Optimality Criterion	815
22.3	Backward Induction in Influence Diagrams	816
22.3.1	Decision Trees for Influence Diagrams	817
22.3.2	Sum-Max-Sum Rule	818
22.4	Computing Expected Utilities	820
22.4.1	Simple Variable Elimination	821
22.4.2	Multiple Utility Variables: Simple Approaches	822
22.4.3	Generalized Variable Elimination	822
22.5	Optimization in Influence Diagrams	826
22.5.1	Optimizing a Single Decision Rule	827
22.5.2	Iterated Optimization Algorithm	827
22.5.3	Strategic Relevance and Global Optimality	829
22.6	Ignoring Irrelevant Information	834
22.7	Value of Information	836
22.7.1	Single Observations	836
22.7.2	Multiple Observations	839
22.8	Summary	840
<i>Index</i>	877	