

Contents

Preface **xv**

1 Introduction **1**

I Fundamentals **3**

2 Probability **5**

2.1	Introduction	5
2.2	Some common univariate distributions	5
2.2.1	Some common discrete distributions	5
2.2.2	Some common continuous distributions	8
2.2.3	Pareto distribution	14
2.3	The multivariate Gaussian (normal) distribution	16
2.3.1	Definition	16
2.3.2	Moment form and canonical form	17
2.3.3	Marginals and conditionals of a MVN	17
2.3.4	Bayes' rule for Gaussians	18
2.3.5	Example: sensor fusion with known measurement noise	19
2.3.6	Handling missing data	19
2.3.7	A calculus for linear Gaussian models	20
2.4	Some other multivariate continuous distributions	23
2.4.1	Multivariate Student distribution	23
2.4.2	Circular normal (von Mises Fisher) distribution	24
2.4.3	Matrix-variate Gaussian (MVG) distribution	24
2.4.4	Wishart distribution	24
2.4.5	Dirichlet distribution	27
2.5	The exponential family	28
2.5.1	Definition	29
2.5.2	Examples	30
2.5.3	Log partition function is cumulant generating function	34
2.5.4	Canonical (natural) vs mean (moment) parameters	36

1			
2	2.5.5	MLE for the exponential family	37
3	2.5.6	Exponential dispersion family	38
4	2.5.7	Maximum entropy derivation of the exponential family	38
5	2.6	Fisher information matrix (FIM)	39
6	2.6.1	Definition	39
7	2.6.2	Equivalence between the FIM and the Hessian of the NLL	39
8	2.6.3	Examples	41
9	2.6.4	Approximating KL divergence using FIM	42
10	2.6.5	Fisher information matrix for exponential family	42
11	2.7	Transformations of random variables	43
12	2.7.1	Invertible transformations (bijections)	44
13	2.7.2	Monte Carlo approximation	44
14	2.7.3	Probability integral transform	44
15	2.8	Markov chains	46
16	2.8.1	Parameterization	46
17	2.8.2	Application: Language modeling	48
18	2.8.3	Parameter estimation	49
19	2.8.4	Stationary distribution of a Markov chain	51
20	2.9	Divergence measures between probability distributions	54
21	2.9.1	f-divergence	55
22	2.9.2	Integral probability metrics	56
23	2.9.3	Maximum mean discrepancy (MMD)	57
24	2.9.4	Total variation distance	60
25	2.9.5	Comparing distributions using binary classifiers	60
26			
27	3	Statistics	63
28	3.1	Introduction	63
29	3.1.1	Frequentist statistics	63
30	3.1.2	Bayesian statistics	63
31	3.1.3	Arguments for the Bayesian approach	64
32	3.1.4	Arguments against the Bayesian approach	65
33	3.1.5	Why not just use MAP estimation?	65
34	3.2	Closed-form analysis using conjugate priors	70
35	3.2.1	The binomial model	70
36	3.2.2	The multinomial model	78
37	3.2.3	The univariate Gaussian model	80
38	3.2.4	The multivariate Gaussian model	85
39	3.2.5	Conjugate-exponential models	90
40	3.3	Beyond conjugate priors	93
41	3.3.1	Robust (heavy-tailed) priors	93
42	3.3.2	Priors for variance parameters	93
43	3.4	Noninformative priors	95
44	3.4.1	Maximum entropy priors	95
45	3.4.2	Jeffreys priors	96
46	3.4.3	Invariant priors	99
47			

1			
2			
3	3.4.4	Reference priors	100
4	3.5	Hierarchical priors	100
5	3.5.1	A hierarchical binomial model	101
6	3.5.2	A hierarchical Gaussian model	103
7	3.6	Empirical Bayes	106
8	3.6.1	A hierarchical binomial model	107
9	3.6.2	A hierarchical Gaussian model	108
10	3.6.3	Hierarchical Bayes for n-gram smoothing	108
11	3.7	Model selection and evaluation	110
12	3.7.1	Bayesian model selection	111
13	3.7.2	Estimating the marginal likelihood	111
14	3.7.3	Connection between cross validation and marginal likelihood	112
15	3.7.4	Pareto-Smoothed Importance Sampling LOO estimate	113
16	3.7.5	Information criteria	115
17	3.7.6	Posterior predictive checks	117
18	3.7.7	Bayesian p-values	118
19	3.8	Bayesian decision theory	120
20	3.8.1	Basics	120
21	3.8.2	Example: COVID-19	121
22	3.8.3	One-shot decision problems	122
23	3.8.4	Multi-stage decision problems	123
24	4	Probabilistic graphical models	125
25	4.1	Introduction	125
26	4.2	Directed graphical models (Bayes nets)	125
27	4.2.1	Representing the joint distribution	125
28	4.2.2	Examples	126
29	4.2.3	Conditional independence properties	131
30	4.2.4	Generation (sampling)	136
31	4.2.5	Inference	136
32	4.2.6	Learning	138
33	4.2.7	Plate notation	143
34	4.3	Undirected graphical models (Markov random fields)	146
35	4.3.1	Representing the joint distribution	147
36	4.3.2	Examples	148
37	4.3.3	Conditional independence properties	155
38	4.3.4	Generation (sampling)	157
39	4.3.5	Inference	157
40	4.3.6	Learning	158
41	4.4	Conditional random fields (CRFs)	162
42	4.4.1	1d CRFs	163
43	4.4.2	2d CRFs	166
44	4.4.3	Parameter estimation	169
45	4.4.4	Other approaches to structured prediction	170
46	4.5	Comparing directed and undirected PGMs	170
47			

1			
2	4.5.1	CI properties	170
3	4.5.2	Converting between a directed and undirected model	171
4	4.5.3	Conditional directed vs undirected PGMs and the label bias problem	173
5	4.5.4	Combining directed and undirected graphs	174
6	4.5.5	Comparing directed and undirected Gaussian PGMs	175
7	4.5.6	Factor graphs	177
8	4.6	Extensions of Bayes nets	180
9	4.6.1	Probabilistic circuits	180
10	4.6.2	Relational probability models	181
11	4.6.3	Open-universe probability models	183
12	4.6.4	Programs as probability models	185
13	4.7	Structural causal models	185
14	4.7.1	Example: causal impact of education on wealth	186
15	4.7.2	Structural equation models	187
16	4.7.3	Do operator and augmented DAGs	187
17	4.7.4	Estimating average treatment effect using path analysis	189
18	4.7.5	Counterfactuals	190
19	5	Information theory	193
20	5.1	KL divergence	193
21	5.1.1	Desiderata	193
22	5.1.2	The KL divergence uniquely satisfies the desiderata	195
23	5.1.3	Thinking about KL	198
24	5.1.4	Properties of KL	200
25	5.1.5	KL divergence and MLE	202
26	5.1.6	KL divergence and Bayesian Inference	203
27	5.1.7	KL divergence and Exponential Families	204
28	5.2	Entropy	205
29	5.2.1	Definition	205
30	5.2.2	Differential entropy for continuous random variables	206
31	5.2.3	Typical sets	207
32	5.2.4	Cross entropy and perplexity	209
33	5.3	Mutual information	210
34	5.3.1	Definition	210
35	5.3.2	Interpretation	210
36	5.3.3	Data processing inequality	211
37	5.3.4	Sufficient Statistics	212
38	5.3.5	Multivariate mutual information	212
39	5.3.6	Variational bounds on mutual information	215
40	5.4	Data compression (source coding)	218
41	5.4.1	Lossless compression	218
42	5.4.2	Lossy compression and the rate-distortion tradeoff	218
43	5.4.3	Bits back coding	221
44	5.5	Error-correcting codes (channel coding)	221
45	5.6	The information bottleneck	223
46			
47			

1			
2	5.6.1	Vanilla IB	223
3	5.6.2	Variational IB	224
4	5.6.3	Conditional entropy bottleneck	225
5			
6	6	Optimization	229
7	6.1	Introduction	229
8	6.2	Automatic differentiation	229
9	6.2.1	Differentiation in functional form	229
10	6.2.2	Differentiating chains, circuits, and programs	233
11	6.3	Stochastic gradient descent	239
12	6.4	Natural gradient descent	239
13	6.4.1	Defining the natural gradient	240
14	6.4.2	Interpretations of NGD	241
15	6.4.3	Benefits of NGD	242
16	6.4.4	Approximating the natural gradient	242
17	6.4.5	Natural gradients for the exponential family	244
18	6.5	Mirror descent	246
19	6.5.1	Bregman divergence	247
20	6.5.2	Proximal point method	248
21	6.5.3	PPM using Bregman divergence	248
22	6.6	Gradients of stochastic functions	248
23	6.6.1	Minibatch approximation to finite-sum objectives	249
24	6.6.2	Optimizing parameters of a distribution	249
25	6.6.3	Score function estimator (likelihood ratio trick)	250
26	6.6.4	Reparameterization trick	251
27	6.6.5	The delta method	253
28	6.6.6	Gumbel softmax trick	253
29	6.6.7	Stochastic computation graphs	254
30	6.6.8	Straight-through estimator	254
31	6.7	Bound optimization (MM) algorithms	255
32	6.7.1	The general algorithm	255
33	6.7.2	Example: logistic regression	256
34	6.7.3	The EM algorithm	258
35	6.7.4	Example: EM for an MVN with missing data	260
36	6.7.5	Example: robust linear regression using Student- <i>t</i> likelihood	262
37	6.7.6	Extensions to EM	263
38	6.8	The Bayesian learning rule	265
39	6.8.1	Deriving inference algorithms from BLR	266
40	6.8.2	Deriving optimization algorithms from BLR	268
41	6.8.3	Variational optimization	271
42	6.9	Bayesian optimization	272
43	6.9.1	Sequential model-based optimization	272
44	6.9.2	Surrogate functions	274
45	6.9.3	Acquisition functions	275
46	6.9.4	Other issues	278
47			

1			
2	6.10	Optimal Transport	279
3	6.10.1	Warm-up: Matching optimally two families of points	279
4	6.10.2	From Optimal Matchings to Kantorovich and Monge formulations	280
5	6.10.3	Solving optimal transport	282
6	6.11	Submodular optimization	287
7	6.11.1	Intuition, Examples, and Background	288
8	6.11.2	Submodular Basic Definitions	290
9	6.11.3	Example Submodular Functions	291
10	6.11.4	Submodular Optimization	294
11	6.11.5	Applications of Submodularity in Machine Learning and AI	298
12	6.11.6	Sketching, CoreSets, Distillation, and Data Subset & Feature Selection	298
13	6.11.7	Combinatorial Information Functions	302
14	6.11.8	Clustering, Data Partitioning, and Parallel Machine Learning	303
15	6.11.9	Active and Semi-Supervised Learning	303
16	6.11.10	Probabilistic Modeling	304
17	6.11.11	Structured Norms and Loss Functions	306
18	6.11.12	Conclusions	306
19	6.12	Derivative free optimization	307
20			
21			
22	II	Inference	309
23			
24	7	Inference algorithms: an overview	311
25	7.1	Introduction	311
26	7.2	Common inference patterns	311
27	7.2.1	Global latents	312
28	7.2.2	Local latents	312
29	7.2.3	Global and local latents	313
30	7.3	Exact inference algorithms	313
31	7.4	Approximate inference algorithms	314
32	7.4.1	MAP estimation	314
33	7.4.2	Grid approximation	314
34	7.4.3	Laplace (quadratic) approximation	315
35	7.4.4	Variational inference	316
36	7.4.5	Markov Chain Monte Carlo (MCMC)	318
37	7.4.6	Sequential Monte Carlo	319
38	7.5	Evaluating approximate inference algorithms	320
39			
40	8	Message passing inference	323
41	8.1	Introduction	323
42	8.2	Belief propagation for discrete chains	324
43	8.2.1	Example: casino HMM	324
44	8.2.2	Forwards filtering	326
45	8.2.3	Backwards smoothing	328
46	8.2.4	The Viterbi algorithm	332
47			

1			
2		8.2.5	Forwards filtering, backwards sampling 335
3	8.3	Belief propagation for Gaussian chains	336
4		8.3.1	Example: tracking SSM 337
5		8.3.2	The Kalman filter 338
6		8.3.3	The Kalman (RTS) smoother 343
7		8.3.4	Extensions to the nonlinear case 344
8	8.4	Belief propagation on trees	345
9		8.4.1	BP for polytrees 345
10		8.4.2	BP for undirected graphs with pairwise potentials 348
11		8.4.3	BP for factor graphs 349
12		8.4.4	Max product belief propagation 349
13		8.4.5	Gaussian and non-Gaussian belief propagation 351
14	8.5	Loopy belief propagation	352
15		8.5.1	Convergence 352
16		8.5.2	Accuracy 355
17		8.5.3	Connection with variational inference 355
18		8.5.4	Generalized belief propagation 355
19		8.5.5	Application: error correcting codes 356
20		8.5.6	Application: Affinity propagation 357
21		8.5.7	Emulating BP with graph neural nets 358
22	8.6	The variable elimination (VE) algorithm	359
23		8.6.1	Derivation of the algorithm 360
24		8.6.2	Computational complexity of VE 362
25		8.6.3	Computational complexity of exact inference 363
26		8.6.4	Drawbacks of VE 364
27	8.7	The junction tree algorithm (JTA)	364
28		8.7.1	Creating a junction tree 365
29		8.7.2	Running belief propagation on a junction tree 369
30		8.7.3	The generalized distributive law 370
31		8.7.4	Other applications of the JTA 371
32	8.8	Inference as backpropagation	372
33			
34	9	Variational inference	375
35	9.1	Introduction	375
36		9.1.1	Variational free energy 375
37		9.1.2	Evidence lower bound (ELBO) 376
38	9.2	Mean field VI	377
39		9.2.1	Coordinate ascent variational inference (CAVI) 377
40		9.2.2	Example: CAVI for the Ising model 378
41		9.2.3	Variational Bayes 380
42		9.2.4	Example: VB for a univariate Gaussian 381
43		9.2.5	Variational Bayes EM 384
44		9.2.6	Example: VBEM for a GMM 385
45		9.2.7	Variational message passing (VMP) 391
46		9.2.8	Autoconj 392
47			

1			
2	9.3	Fixed-form VI	392
3	9.3.1	Black-box variational inference	392
4	9.3.2	Stochastic variational inference	394
5	9.3.3	Reparameterization VI	395
6	9.3.4	Gaussian VI	396
7	9.3.5	Automatic differentiation VI	400
8	9.3.6	Beyond Gaussian posteriors	401
9	9.3.7	Amortized inference	403
10	9.3.8	Exploiting partial conjugacy	404
11	9.3.9	Online variational inference	408
12	9.4	More accurate variational posteriors	411
13	9.4.1	Structured mean field	412
14	9.4.2	Hierarchical (auxiliary variable) posteriors	412
15	9.4.3	Normalizing flow posteriors	412
16	9.4.4	Implicit posteriors	414
17	9.4.5	Combining VI with MCMC inference	414
18	9.5	Lower bounds	415
19	9.5.1	Multi-sample ELBO (IWAE bound)	415
20	9.5.2	The thermodynamic variational objective (TVO)	416
21	9.6	Upper bounds	416
22	9.6.1	Minimizing the χ -divergence upper bound	417
23	9.6.2	Minimizing the evidence upper bound	418
24	9.7	Expectation propagation (EP)	419
25	9.7.1	Minimizing forwards vs reverse KL	419
26	9.7.2	EP as generalized ADF	421
27	9.7.3	Algorithm	421
28	9.7.4	Example	422
29	9.7.5	Optimization issues	422
30	9.7.6	Power EP and α -divergence	423
31	9.7.7	Stochastic EP	423
32	9.7.8	Applications	424
33			
34	10	Monte Carlo inference	425
35	10.1	Introduction	425
36	10.2	Monte Carlo integration	425
37	10.2.1	Example: estimating π by Monte Carlo integration	426
38	10.2.2	Accuracy of Monte Carlo integration	426
39	10.3	Generating random samples from simple distributions	428
40	10.3.1	Sampling using the inverse cdf	428
41	10.3.2	Sampling from a Gaussian (Box-Muller method)	429
42	10.4	Rejection sampling	429
43	10.4.1	Basic idea	430
44	10.4.2	Example	431
45	10.4.3	Adaptive rejection sampling	431
46	10.4.4	Rejection sampling in high dimensions	432
47			

1			
2	10.5	Importance sampling	432
3	10.5.1	Direct importance sampling	433
4	10.5.2	Self-normalized importance sampling	433
5	10.5.3	Choosing the proposal	434
6	10.5.4	Annealed importance sampling (AIS)	434
7	10.6	Controlling Monte Carlo variance	436
8	10.6.1	Rao-Blackwellisation	436
9	10.6.2	Control variates	437
10	10.6.3	Antithetic sampling	438
11	10.6.4	Quasi Monte Carlo (QMC)	439
12	11	Markov Chain Monte Carlo inference	441
13	11.1	Introduction	441
14	11.2	Metropolis Hastings algorithm	441
15	11.2.1	Basic idea	442
16	11.2.2	Why MH works	443
17	11.2.3	Proposal distributions	444
18	11.2.4	Initialization	446
19	11.2.5	Simulated annealing	447
20	11.3	Gibbs sampling	449
21	11.3.1	Basic idea	449
22	11.3.2	Gibbs sampling is a special case of MH	450
23	11.3.3	Example: Gibbs sampling for Ising models	450
24	11.3.4	Example: Gibbs sampling for Potts models	452
25	11.3.5	Example: Gibbs sampling for GMMs	452
26	11.3.6	Sampling from the full conditionals	454
27	11.3.7	Blocked Gibbs sampling	455
28	11.3.8	Collapsed Gibbs sampling	456
29	11.4	Auxiliary variable MCMC	458
30	11.4.1	Slice sampling	459
31	11.4.2	Swendsen Wang	460
32	11.5	Hamiltonian Monte Carlo (HMC)	462
33	11.5.1	Hamiltonian mechanics	462
34	11.5.2	Integrating Hamilton's equations	463
35	11.5.3	The HMC algorithm	464
36	11.5.4	Tuning HMC	465
37	11.5.5	Riemann Manifold HMC	466
38	11.5.6	Langevin Monte Carlo (MALA)	467
39	11.5.7	Connection between SGD and Langevin sampling	468
40	11.5.8	Applying HMC to constrained parameters	470
41	11.5.9	Speeding up HMC	470
42	11.6	MCMC convergence	471
43	11.6.1	Mixing rates of Markov chains	472
44	11.6.2	Practical convergence diagnostics	472
45	11.6.3	Improving speed of convergence	479
46			
47			

1			
2		11.6.4 Non-centered parameterizations and Neal’s funnel	480
3	11.7	Stochastic gradient MCMC	481
4		11.7.1 Stochastic Gradient Langevin Dynamics (SGLD)	482
5		11.7.2 Preconditioning	482
6		11.7.3 Reducing the variance of the gradient estimate	483
7		11.7.4 SG-HMC	484
8		11.7.5 Underdamped Langevin Dynamics	485
9	11.8	Reversible jump (trans-dimensional) MCMC	486
10		11.8.1 Basic idea	486
11		11.8.2 Example	488
12		11.8.3 Discussion	489
13	11.9	Annealing methods	489
14		11.9.1 Parallel tempering	489
15	12	Sequential Monte Carlo inference	491
16	12.1	Introduction	491
17		12.1.1 Problem statement	491
18		12.1.2 Particle filtering for state-space models	491
19		12.1.3 SMC samplers for static parameter estimation	493
20	12.2	Basics of SMC	493
21		12.2.1 Importance sampling	493
22		12.2.2 Sequential importance sampling	494
23		12.2.3 Sequential importance sampling with resampling	495
24		12.2.4 Resampling methods	498
25		12.2.5 Adaptive resampling	500
26	12.3	Some applications of particle filtering	501
27		12.3.1 1d pendulum model with outliers	501
28		12.3.2 Visual object tracking	502
29		12.3.3 Robot localization	503
30		12.3.4 Online parameter estimation	504
31	12.4	Proposal distributions	505
32		12.4.1 Locally optimal proposal	506
33		12.4.2 Proposals based on the Laplace approximation	506
34		12.4.3 Proposals based on the extended and unscented Kalman filter	508
35		12.4.4 Proposals based on SMC	508
36		12.4.5 Neural adaptive SMC	509
37		12.4.6 Amortized adaptive SMC	509
38		12.4.7 Variational SMC	510
39	12.5	Rao-Blackwellised particle filtering (RBPF)	511
40		12.5.1 Mixture of Kalman filters	511
41		12.5.2 FastSLAM	515
42	12.6	SMC samplers	516
43		12.6.1 Ingredients of an SMC sampler	517
44		12.6.2 Likelihood tempering (geometric path)	518
45		12.6.3 Data tempering	520
46			
47			

1			
2	12.6.4	Sampling rare events and extrema	522
3	12.6.5	SMC-ABC and likelihood-free inference	523
4	12.6.6	SMC ²	523
5	12.7	Particle MCMC methods	523
6	12.7.1	Particle Marginal Metropolis Hastings	524
7	12.7.2	Particle Independent Metropolis Hastings	525
8	12.7.3	Particle Gibbs	526
9			
10			
11	III	Prediction	527
12	13	Predictive models: an overview	529
13	13.1	Introduction	529
14	13.1.1	Types of model	529
15	13.1.2	Model fitting using ERM, MLE and MAP	530
16	13.1.3	Model fitting using Bayes, VI and generalized Bayes	531
17	13.2	Evaluating predictive models	532
18	13.2.1	Proper scoring rules	532
19	13.2.2	Calibration	532
20	13.2.3	Beyond evaluating marginal probabilities	536
21	13.3	Conformal prediction	539
22	13.3.1	Conformalizing classification	540
23	13.3.2	Conformalizing regression	541
24	13.3.3	Conformalizing Bayes	542
25	13.3.4	What do we do if we don't have a calibration set?	543
26			
27	14	Generalized linear models	545
28	14.1	Introduction	545
29	14.1.1	Examples	545
30	14.1.2	GLMs with non-canonical link functions	548
31	14.1.3	Maximum likelihood estimation	548
32	14.1.4	Bayesian inference	549
33	14.2	Linear regression	550
34	14.2.1	Conjugate priors	550
35	14.2.2	Uninformative priors	552
36	14.2.3	Informative priors	554
37	14.2.4	Online Bayesian inference using the Kalman filter (recursive least squares)	556
38	14.2.5	Spike and slab prior	558
39	14.2.6	Laplace prior (Bayesian lasso)	559
40	14.2.7	Horseshoe prior	560
41	14.2.8	Automatic relevancy determination	561
42	14.3	Logistic regression	564
43	14.3.1	Binary logistic regression	564
44	14.3.2	Multinomial logistic regression	564
45	14.3.3	Priors	565
46			
47			

1			
2	14.3.4	Posteriors	566
3	14.3.5	Laplace approximation	566
4	14.3.6	MCMC inference	569
5	14.3.7	Variational inference	570
6	14.3.8	Assumed density filtering	570
7	14.4	Probit regression	574
8	14.4.1	Latent variable interpretation	575
9	14.4.2	Maximum likelihood estimation	575
10	14.4.3	Bayesian inference	577
11	14.4.4	Ordinal probit regression	577
12	14.4.5	Multinomial probit models	578
13	14.5	Multi-level GLMs	578
14	14.5.1	Generalized linear mixed models (GLMMs)	578
15	14.5.2	Model fitting	579
16	14.5.3	Example: radon regression	579
17	15	Deep neural networks	583
18	15.1	Introduction	583
19	15.2	Building blocks of differentiable circuits	583
20	15.2.1	Linear layers	584
21	15.2.2	Non-linearities	584
22	15.2.3	Convolutional layers	585
23	15.2.4	Residual (skip) connections	586
24	15.2.5	Normalization layers	587
25	15.2.6	Dropout layers	587
26	15.2.7	Attention layers	588
27	15.2.8	Recurrent layers	591
28	15.2.9	Multiplicative layers	591
29	15.2.10	Implicit layers	592
30	15.3	Canonical examples of neural networks	592
31	15.3.1	Multi-layer perceptrons (MLP)	593
32	15.3.2	Convolutional neural networks (CNN)	593
33	15.3.3	Recurrent neural networks (RNN)	594
34	15.3.4	Transformers	595
35	15.3.5	Graph neural networks (GNNs)	597
36	16	Bayesian neural networks	603
37	16.1	Introduction	603
38	16.2	Priors for BNNs	603
39	16.2.1	Gaussian priors	604
40	16.2.2	Sparsity-promoting priors	605
41	16.2.3	Learning the prior	605
42	16.2.4	Priors in function space	606
43	16.2.5	Architectural priors	606
44	16.3	Likelihoods for BNNs	607
45			
46			
47			

1			
2	16.4	Posteriors for BNNs	608
3	16.4.1	Laplace approximation	608
4	16.4.2	Variational inference	609
5	16.4.3	Expectation propagation	610
6	16.4.4	Last layer methods	610
7	16.4.5	Dropout	610
8	16.4.6	MCMC methods	611
9	16.4.7	Methods based on the SGD trajectory	611
10	16.4.8	Deep ensembles	612
11	16.4.9	Approximating the posterior predictive distribution	615
12	16.5	Generalization in Bayesian deep learning	616
13	16.5.1	Sharp vs flat minima	616
14	16.5.2	Effective dimensionality of a model	618
15	16.5.3	The hypothesis space of DNNs	619
16	16.5.4	Double descent	620
17	16.5.5	A Bayesian Resolution to Double Descent	623
18	16.5.6	PAC-Bayes	624
19	16.5.7	Out-of-Distribution Generalization for BNNs	625
20	16.6	Online inference	627
21	16.6.1	Extended Kalman Filtering for DNNs	627
22	16.6.2	Assumed Density Filtering for DNNs	630
23	16.6.3	Sequential Laplace for DNNs	631
24	16.6.4	Variational methods	632
25	16.7	Hierarchical Bayesian neural networks	632
26	16.7.1	Solving multiple related classification problems	632
27	17	Gaussian processes	637
28	17.1	Introduction	637
29	17.1.1	GPs: What and why?	637
30	17.2	Mercer kernels	639
31	17.2.1	Some popular Mercer kernels	640
32	17.2.2	Mercer's theorem	646
33	17.2.3	Kernels from Spectral Densities	647
34	17.3	GPs with Gaussian likelihoods	648
35	17.3.1	Predictions using noise-free observations	648
36	17.3.2	Predictions using noisy observations	649
37	17.3.3	Weight space vs function space	650
38	17.3.4	Semi-parametric GPs	651
39	17.3.5	Marginal likelihood	652
40	17.3.6	Computational and numerical issues	652
41	17.3.7	Kernel ridge regression	653
42	17.4	GPs with non-Gaussian likelihoods	656
43	17.4.1	Binary classification	656
44	17.4.2	Multi-class classification	658
45	17.4.3	GPs for Poisson regression (Cox process)	658
46			
47			

1				
2	17.5	Scaling GP inference to large datasets	659	
3	17.5.1	Subset of data	660	
4	17.5.2	Nystrom approximation	661	
5	17.5.3	Inducing point methods	662	
6	17.5.4	Sparse variational methods	665	
7	17.5.5	Exploiting parallelization and structure via kernel matrix multiplies	668	
8	17.5.6	Converting a GP to a SSM	670	
9	17.6	Learning the kernel	671	
10	17.6.1	Empirical Bayes for the kernel parameters	672	
11	17.6.2	Bayesian inference for the kernel parameters	673	
12	17.6.3	Multiple kernel learning for additive kernels	675	
13	17.6.4	Automatic search for compositional kernels	676	
14	17.6.5	Spectral mixture kernel learning	678	
15	17.6.6	Deep kernel learning	680	
16	17.6.7	Functional kernel learning	682	
17	17.7	GPs and DNNs	682	
18	17.7.1	Kernels derived from random DNNs (NN-GP)	683	
19	17.7.2	Kernels derived from trained DNNs (neural tangent kernel)	686	
20	17.7.3	Deep GPs	688	
21	18	Beyond the iid assumption	695	
22	18.1	Introduction	695	
23	18.2	Distribution shift	695	
24	18.2.1	Motivating examples	695	
25	18.2.2	A causal view of distribution shift	697	
26	18.2.3	Covariate shift	698	
27	18.2.4	Domain shift	698	
28	18.2.5	Label / prior shift	699	
29	18.2.6	Concept shift	699	
30	18.2.7	Manifestation shift	699	
31	18.2.8	Selection bias	700	
32	18.3	Training-time techniques for distribution shift	700	
33	18.3.1	Importance weighting for covariate shift	701	
34	18.3.2	Domain adaptation	702	
35	18.3.3	Domain randomization	702	
36	18.3.4	Data augmentation	703	
37	18.3.5	Unsupervised label shift estimation	703	
38	18.3.6	Distributionally robust optimization	704	
39	18.4	Test-time techniques for distribution shift	704	
40	18.4.1	Detecting shifts using two-sample testing	704	
41	18.4.2	Detecting single out-of-distribution (OOD) inputs	704	
42	18.4.3	Selective prediction	707	
43	18.4.4	Open world recognition	709	
44	18.4.5	Online adaptation	709	
45	18.5	Learning from multiple distributions	710	
46				
47				

1			
2	18.5.1	Transfer learning	710
3	18.5.2	Few-shot learning	711
4	18.5.3	Prompt tuning	711
5	18.5.4	Zero-shot learning	712
6	18.5.5	Multi-task learning	713
7	18.5.6	Domain generalization	713
8	18.5.7	Invariant risk minimization	714
9	18.6	Meta-learning	715
10	18.6.1	Meta-learning as probabilistic inference for prediction	715
11	18.6.2	Gradient-based meta-learning	717
12	18.6.3	Metric-based few-shot learning	717
13	18.6.4	VERSA	718
14	18.6.5	Neural processes	718
15	18.7	Continual learning	718
16	18.7.1	Domain drift	718
17	18.7.2	Concept drift	719
18	18.7.3	Task incremental learning	720
19	18.7.4	Catastrophic forgetting	722
20	18.7.5	Online learning	723
21	18.8	Adversarial examples	725
22	18.8.1	Whitebox (gradient-based) attacks	726
23	18.8.2	Blackbox (gradient-free) attacks	727
24	18.8.3	Real world adversarial attacks	728
25	18.8.4	Defenses based on robust optimization	729
26	18.8.5	Why models have adversarial examples	729
27			
28			
29	IV	Generation	731
30			
31	19	Generative models: an overview	733
32	19.1	Introduction	733
33	19.2	Types of generative model	733
34	19.3	Goals of generative modeling	735
35	19.3.1	Generating data	735
36	19.3.2	Density estimation	736
37	19.3.3	Imputation	737
38	19.3.4	Structure discovery	737
39	19.3.5	Latent space interpolation	738
40	19.3.6	Representation learning	739
41	19.4	Evaluating generative models	739
42	19.4.1	Likelihood	740
43	19.4.2	Distances and divergences in feature space	742
44	19.4.3	Precision and recall metrics	743
45	19.4.4	Statistical tests	744
46	19.4.5	Challenges with using pretrained classifiers	744
47			

1			
2	19.4.6	Using model samples to train classifiers	744
3	19.4.7	Assessing overfitting	745
4	19.4.8	Human evaluation	745
5	20	Variational autoencoders	747
6	20.1	Introduction	747
7	20.2	VAE basics	747
8	20.2.1	Modeling assumptions	747
9	20.2.2	Evidence lower bound	749
10	20.2.3	Optimization	750
11	20.2.4	The reparameterization trick	750
12	20.2.5	Computing the reparameterized ELBO	752
13	20.2.6	Comparison of VAEs and autoencoders	754
14	20.2.7	VAEs optimize in an augmented space	755
15	20.3	VAE generalizations	757
16	20.3.1	σ -VAE	757
17	20.3.2	β -VAE	758
18	20.3.3	InfoVAE	760
19	20.3.4	Multi-modal VAEs	763
20	20.3.5	VAEs with missing data	765
21	20.3.6	Semi-supervised VAEs	766
22	20.3.7	VAEs with sequential encoders/decoders	768
23	20.4	Avoiding posterior collapse	771
24	20.4.1	KL annealing	772
25	20.4.2	Lower bounding the rate	772
26	20.4.3	Free bits	772
27	20.4.4	Adding skip connections	772
28	20.4.5	Improved variational inference	772
29	20.4.6	Alternative objectives	773
30	20.4.7	Enforcing identifiability	773
31	20.5	VAEs with hierarchical structure	774
32	20.5.1	Bottom-up vs top-down inference	775
33	20.5.2	Example: Very deep VAE	776
34	20.5.3	Connection with autoregressive models	778
35	20.5.4	Variational pruning	779
36	20.5.5	Other optimization difficulties	779
37	20.6	Vector quantization VAE	780
38	20.6.1	Autoencoder with binary code	780
39	20.6.2	VQ-VAE model	781
40	20.6.3	Learning the prior	783
41	20.6.4	Hierarchical extension (VQ-VAE-2)	783
42	20.6.5	Discrete VAE	784
43	20.6.6	VQ-GAN	785
44	20.7	Wake-sleep algorithm	785
45	20.7.1	Wake phase	786
46			
47			

1			
2	20.7.2	Sleep phase	787
3	20.7.3	Daydream phase	788
4	20.7.4	Summary of algorithm	788
5			
6	21	Auto-regressive models	791
7	21.1	Introduction	791
8	21.2	Neural autoregressive density estimators (NADE)	792
9	21.3	Causal CNNs	792
10	21.3.1	1d causal CNN (Convolutional Markov models)	793
11	21.3.2	2d causal CNN (PixelCNN)	793
12	21.4	Transformer decoders	794
13	21.4.1	Text generation (GPT)	795
14	21.4.2	Music generation	795
15	21.4.3	Text-to-image generation (DALL-E)	796
16			
17	22	Normalizing Flows	799
18	22.1	Introduction	799
19	22.1.1	Preliminaries	799
20	22.1.2	Example	801
21	22.1.3	How to train a flow model	802
22	22.2	Constructing Flows	803
23	22.2.1	Affine flows	803
24	22.2.2	Elementwise flows	804
25	22.2.3	Coupling flows	806
26	22.2.4	Autoregressive flows	808
27	22.2.5	Residual flows	813
28	22.2.6	Continuous-time flows	815
29	22.3	Applications	817
30	22.3.1	Density estimation	817
31	22.3.2	Generative Modeling	818
32	22.3.3	Inference	818
33			
34	23	Energy-based models	819
35	23.1	Introduction	819
36	23.1.1	Example: Products of experts (PoE)	819
37	23.1.2	Computational difficulties	820
38	23.2	Maximum Likelihood Training	820
39	23.2.1	Gradient-based MCMC methods	822
40	23.2.2	Contrastive divergence	822
41	23.3	Score Matching (SM)	825
42	23.3.1	Basic score matching	826
43	23.3.2	Denoising Score Matching (DSM)	826
44	23.3.3	Sliced Score Matching (SSM)	828
45	23.3.4	Connection to Contrastive Divergence	829
46	23.3.5	Score-Based Generative Models	830
47			

1			
2	23.4	Noise Contrastive Estimation	832
3	23.4.1	Connection to Score Matching	834
4	23.5	Other Methods	835
5	23.5.1	Minimizing Differences/Derivatives of KL Divergences	835
6	23.5.2	Minimizing the Stein Discrepancy	835
7	23.5.3	Adversarial Training	836
8	24	Denoising diffusion models	839
9	24.1	Model definition	839
10	24.2	Examples	841
11	24.3	Model training	842
12	24.4	Connections with other generative models	844
13	24.4.1	Connection with score matching	844
14	24.4.2	Connection with VAEs	845
15	24.4.3	Connection with flow models	845
16	25	Generative adversarial networks	847
17	25.1	Introduction	847
18	25.2	Learning by Comparison	848
19	25.2.1	Guiding principles	849
20	25.2.2	Class probability estimation	850
21	25.2.3	Bounds on f -divergences	853
22	25.2.4	Integral probability metrics	854
23	25.2.5	Moment matching	856
24	25.2.6	On density ratios and differences	857
25	25.3	Generative Adversarial Networks	858
26	25.3.1	From learning principles to loss functions	859
27	25.3.2	Gradient Descent	860
28	25.3.3	Challenges with GAN training	861
29	25.3.4	Improving GAN optimization	863
30	25.3.5	Convergence of GAN training	863
31	25.4	Conditional GANs	866
32	25.5	Inference with GANs	868
33	25.6	Neural architectures in GANs	868
34	25.6.1	The importance of discriminator architectures	869
35	25.6.2	Architectural inductive biases	869
36	25.6.3	Attention in GANs	869
37	25.6.4	Progressive generation	870
38	25.6.5	Regularization	871
39	25.6.6	Scaling up GAN models	872
40	25.7	Applications	872
41	25.7.1	GANs for image generation	873
42	25.7.2	Video generation	875
43	25.7.3	Audio generation	876
44	25.7.4	Text generation	877
45			
46			
47			

25.7.5	Imitation Learning	878
25.7.6	Domain Adaptation	878
25.7.7	Design, Art and Creativity	878
V Discovery 881		
26	Discovery methods: an overview	883
26.1	Introduction	883
26.2	Overview of Part V	884
27	Latent variable models	885
27.1	Introduction	885
27.2	Mixture models	885
27.2.1	Gaussian mixture models (GMMs)	886
27.2.2	Bernoulli mixture models	888
27.2.3	Gaussian scale mixtures	888
27.2.4	Using GMMs as a prior for inverse imaging problems	890
27.3	Factor analysis	893
27.3.1	Vanilla factor analysis	893
27.3.2	Probabilistic PCA	897
27.3.3	Factor analysis models for paired data	900
27.3.4	Factor analysis with exponential family likelihoods	902
27.3.5	Factor analysis with DNN likelihoods	905
27.3.6	Factor analysis with GP likelihoods (GP-LVM)	905
27.4	Mixture of factor analysers	906
27.4.1	Model definition	906
27.4.2	Model fitting	907
27.4.3	MixFA for image generation	910
27.5	LVMs with non-Gaussian priors	912
27.5.1	Non-negative matrix factorization (NMF)	913
27.5.2	Multinomial PCA	913
27.5.3	Latent Dirichlet Allocation (LDA)	916
27.6	Independent components analysis (ICA)	916
27.6.1	Noiseless ICA model	916
27.6.2	The need for non-Gaussian priors	918
27.6.3	Maximum likelihood estimation	919
27.6.4	Alternatives to MLE	920
27.6.5	Sparse coding	921
27.6.6	Nonlinear ICA	922
28	Hidden Markov models	925
28.1	Introduction	925
28.2	HMMs: parameterization	925
28.2.1	Transition model	925
28.2.2	Observation model	926

<u>1</u>			
<u>2</u>	28.3	HMMs: Applications	929
<u>3</u>		28.3.1 Segmentation of time series data	929
<u>4</u>		28.3.2 Spelling correction	931
<u>5</u>		28.3.3 Protein sequence alignment	933
<u>6</u>	28.4	HMMs: parameter learning	935
<u>7</u>		28.4.1 The Baum-Welch (EM) algorithm	935
<u>8</u>		28.4.2 Parameter estimation using SGD	940
<u>9</u>		28.4.3 Parameter estimation using spectral methods	942
<u>10</u>		28.4.4 Bayesian parameter inference	942
<u>11</u>	28.5	HMMs: Generalizations	943
<u>12</u>		28.5.1 Hidden semi-Markov model (HSMM)	943
<u>13</u>		28.5.2 HSMMs for changepoint detection	945
<u>14</u>		28.5.3 Hierarchical HMMs	948
<u>15</u>		28.5.4 Factorial HMMs	950
<u>16</u>		28.5.5 Coupled HMMs	952
<u>17</u>		28.5.6 Dynamic Bayes nets (DBN)	954
<u>18</u>	29	State-space models	955
<u>19</u>		29.1 Introduction	955
<u>20</u>		29.2 Linear dynamical systems	955
<u>21</u>		29.2.1 Example: Noiseless 1d spring-mass system	956
<u>22</u>		29.2.2 Example: Noisy 2d tracking problem	957
<u>23</u>		29.2.3 Parameter estimation	960
<u>24</u>	29.3	Non-linear dynamical systems	962
<u>25</u>		29.3.1 Example: nonlinear 2d tracking problem	962
<u>26</u>		29.3.2 Example: Simultaneous localization and mapping (SLAM)	963
<u>27</u>		29.3.3 Example: stochastic volatility models	965
<u>28</u>		29.3.4 Example: Multi-target tracking	966
<u>29</u>	29.4	Other kinds of SSM	968
<u>30</u>		29.4.1 Exponential family SSM	968
<u>31</u>		29.4.2 Bayesian SSM	972
<u>32</u>		29.4.3 GP-SSM	972
<u>33</u>	29.5	Deep SSMs	972
<u>34</u>		29.5.1 Deep Markov models	973
<u>35</u>		29.5.2 Recurrent SSM	974
<u>36</u>		29.5.3 Improving multi-step predictions	975
<u>37</u>		29.5.4 Variational RNNs	976
<u>38</u>	29.6	Time series forecasting	977
<u>39</u>		29.6.1 Structural time series models	977
<u>40</u>		29.6.2 Causal impact of a time series intervention	983
<u>41</u>		29.6.3 Prophet	988
<u>42</u>		29.6.4 Gaussian processes for timeseries forecasting	988
<u>43</u>		29.6.5 Neural forecasting methods	990
<u>44</u>	30	Graph learning	993
<u>45</u>			
<u>46</u>			
<u>47</u>			

1			
2	30.1	Introduction	993
3	30.2	Latent variable models for graphs	993
4	30.2.1	Stochastic block model	993
5	30.2.2	Mixed membership stochastic block model	995
6	30.2.3	Infinite relational model	997
7	30.3	Graphical model structure learning	999
8	30.3.1	Applications	999
9	30.3.2	Relevance networks	1001
10	30.3.3	Learning sparse PGMs	1002
11	31	Non-parametric Bayesian models	1003
12	31.1	Introduction	1003
13	31.2	Dirichlet process	1004
14	31.2.1	Definition	1004
15	31.2.2	Stick breaking construction of the DP	1006
16	31.2.3	The Chinese restaurant process (CRP)	1007
17	31.2.4	Dirichlet process mixture models	1009
18	31.3	Generalizations of the Dirichlet process	1014
19	31.3.1	Pitman-Yor process	1015
20	31.3.2	Dependent random probability measures	1016
21	31.4	The Indian buffet process and the Beta process	1018
22	31.5	Small-variance asymptotics	1021
23	31.6	Completely random measures	1024
24	31.7	Lévy processes	1025
25	31.8	Point processes with repulsion and reinforcement	1027
26	31.8.1	Poisson process	1027
27	31.8.2	Renewal process	1028
28	31.8.3	Hawkes process	1029
29	31.8.4	Gibbs point process	1031
30	31.8.5	Determinantal point process	1032
31	32	Representation learning (Unfinished)	1035
32	32.1	CLIP	1035
33	33	Interpretability	1037
34	33.1	Introduction	1037
35	33.1.1	The Role of Interpretability	1038
36	33.1.2	Terminology and Framework	1039
37	33.2	Methods for Interpretable Machine Learning	1043
38	33.2.1	Inherently Interpretable Models: The Model is its Explanation	1043
39	33.2.2	Semi-Inherently Interpretable Models: Example-Based Methods	1045
40	33.2.3	Post-hoc or Joint training: The Explanation gives a Partial View of the Model	1046
41	33.2.4	Transparency and Visualization	1050
42	33.3	Properties: The Abstraction Between Context and Method	1051
43			
44			
45			
46			
47			

1			
2	33.3.1	Properties of Explanations from Interpretable Machine Learning	1051
3	33.3.2	Properties of Explanations from Cognitive Science	1054
4	33.4	Evaluation of Interpretable Machine Learning Models	1055
5	33.4.1	Computational Evaluation: Does the Method have Desired Properties?	1056
6	33.4.2	User Study-based Evaluation: Does the Method Help a User Perform a Task?	1060
7			
8	33.5	Discussion: How to Think about Interpretable Machine Learning	1064
9			
10			
11	VI	Decision making	1071
12	34	Multi-step decision problems	1073
13	34.1	Introduction	1073
14	34.2	Decision (influence) diagrams	1073
15	34.2.1	Example: oil wildcatter	1073
16	34.2.2	Information arcs	1074
17	34.2.3	Value of information	1075
18	34.2.4	Computing the optimal policy	1076
19	34.3	A/B testing	1076
20	34.3.1	A Bayesian approach	1077
21	34.3.2	Example	1080
22	34.4	Contextual bandits	1081
23	34.4.1	Types of bandit	1081
24	34.4.2	Applications	1083
25	34.4.3	Exploration-exploitation tradeoff	1083
26	34.4.4	The optimal solution	1083
27	34.4.5	Upper confidence bounds (UCB)	1085
28	34.4.6	Thompson sampling	1087
29	34.4.7	Regret	1088
30	34.5	Markov decision problems	1089
31	34.5.1	Basics	1090
32	34.5.2	Partially observed MDPs	1091
33	34.5.3	Episodes and returns	1092
34	34.5.4	Value functions	1092
35	34.5.5	Optimal value functions and policies	1093
36	34.6	Planning in an MDP	1094
37	34.6.1	Value iteration	1095
38	34.6.2	Policy iteration	1096
39	34.6.3	Linear programming	1097
40			
41	35	Reinforcement learning	1099
42	35.1	Introduction	1099
43	35.1.1	Overview of methods	1099
44	35.1.2	Value based methods	1100
45	35.1.3	Policy search methods	1100
46			
47			

1			
2		35.1.4	Model-based RL 1101
3		35.1.5	Exploration-exploitation tradeoff 1101
4	35.2	Value-based RL	1103
5		35.2.1	Monte Carlo RL 1103
6		35.2.2	Temporal difference (TD) learning 1104
7		35.2.3	TD learning with eligibility traces 1105
8		35.2.4	SARSA: on-policy TD control 1105
9		35.2.5	Q-learning: off-policy TD control 1106
10		35.2.6	Deep Q-network (DQN) 1109
11	35.3	Policy-based RL	1110
12		35.3.1	The policy gradient theorem 1110
13		35.3.2	REINFORCE 1111
14		35.3.3	Actor-critic methods 1111
15		35.3.4	Bound optimization methods 1113
16		35.3.5	Deterministic policy gradient methods 1115
17		35.3.6	Gradient-free methods 1116
18	35.4	Model-based RL	1116
19		35.4.1	Model predictive control (MPC) 1117
20		35.4.2	Combining model-based and model-free 1118
21		35.4.3	MBRL using Gaussian processes 1119
22		35.4.4	MBRL using DNNs 1120
23		35.4.5	MBRL using latent-variable models 1121
24		35.4.6	Robustness to model errors 1123
25	35.5	Off-policy learning	1123
26		35.5.1	Basic techniques 1124
27		35.5.2	The curse of horizon 1127
28		35.5.3	The deadly triad 1128
29	35.6	Control as inference	1130
30		35.6.1	Maximum entropy reinforcement learning 1130
31		35.6.2	Other approaches 1132
32		35.6.3	Imitation learning 1133
33	36	Causality	1137
34		36.1	Introduction 1137
35		36.1.1	Why is causality different than other forms of ML? 1137
36		36.2	Causal Formalism 1139
37		36.2.1	Structural Causal Models 1139
38		36.2.2	Causal DAGs 1141
39		36.2.3	Identification 1143
40		36.2.4	Counterfactuals and the Causal Hierarchy 1144
41	36.3	Randomized Control Trials	1146
42	36.4	Confounder Adjustment	1147
43		36.4.1	Causal Estimand, Statistical Estimand, and Identification 1147
44		36.4.2	ATE Estimation with Observed Confounders 1150
45		36.4.3	Uncertainty Quantification 1155
46			
47			

<u>1</u>				
<u>2</u>	36.4.4	Matching	1156	
<u>3</u>	36.4.5	Practical Considerations and Procedures	1157	
<u>4</u>	36.4.6	Summary and Practical Advice	1160	
<u>5</u>	36.5	Instrumental Variable Strategies	1161	
<u>6</u>	36.5.1	Additive Unobserved Confounding	1163	
<u>7</u>	36.5.2	Instrument Monotonicity and Local Average Treatment Effect	1164	
<u>8</u>	36.5.3	Two Stage Least Squares	1168	
<u>9</u>	36.6	Difference in Differences	1168	
<u>10</u>	36.6.1	Estimation	1172	
<u>11</u>	36.7	Credibility Checks	1172	
<u>12</u>	36.7.1	Placebo Checks	1173	
<u>13</u>	36.7.2	Sensitivity Analysis to Unobserved Confounding	1173	
<u>14</u>	36.8	The Do Calculus	1181	
<u>15</u>	36.8.1	The three rules	1181	
<u>16</u>	36.8.2	Revisiting Backdoor Adjustment	1182	
<u>17</u>	36.8.3	Frontdoor Adjustment	1183	
<u>18</u>	36.9	Further Reading	1185	
<u>19</u>	Bibliography	1199		

20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47