Contents

Preface **xxix**

1 Introduction **1**

I Fundamentals **3**

2 Probability **5**

2.1 Introduction 5
2.1.1 Basic rules of probability 5
2.1.2 Exchangeability and de Finetti's theorem 8
2.2 Some common probability distributions 9
2.2.1 Discrete distributions 9
2.2.2 Continuous distributions on $\mathbb{R}$ 11
2.2.3 Continuous distributions on $\mathbb{R}^+$ 13
2.2.4 Continuous distributions on $[0, 1]$ 17
2.2.5 Multivariate continuous distributions 17
2.3 Gaussian joint distributions 22
2.3.1 The multivariate normal 22
2.3.2 Linear Gaussian systems 29
2.3.3 A general calculus for linear Gaussian systems 31
2.4 The exponential family 34
2.4.1 Definition 34
2.4.2 Examples 35
2.4.3 Log partition function is cumulant generating function 40
2.4.4 Canonical (natural) vs mean (moment) parameters 41
2.4.5 MLE for the exponential family 42
2.4.6 Exponential dispersion family 43
2.4.7 Maximum entropy derivation of the exponential family 43
2.5 Transformations of random variables 44
2.5.1 Invertible transformations (bijections) 44
2.5.2 Monte Carlo approximation 45
2.5.3 Probability integral transform 45
2.6 Markov chains 46
2.6.1 Parameterization 47
2.6.2 Application: language modeling 49
2.6.3 Parameter estimation 50
2.6.4 Stationary distribution of a Markov chain 52
2.7 Divergence measures between probability distributions 55
2.7.1 f-divergence 56

2.7.2 Integral probability metrics 58
2.7.3 Maximum mean discrepancy (MMD) 59
2.7.4 Total variation distance 61
2.7.5 Density ratio estimation using binary classifiers 61

3 Statistics 65

3.1 Introduction 65
3.2 Bayesian statistics 65
3.2.1 Tossing coins 66
3.2.2 Modeling more complex data 72
3.2.3 Selecting the prior 73
3.2.4 Computational issues 73
3.3 Frequentist statistics 73
3.3.1 Sampling distributions 74
3.3.2 Bootstrap approximation of the sampling distribution 74
3.3.3 Asymptotic normality of the sampling distribution of the MLE 76
3.3.4 Fisher information matrix 76
3.3.5 Counterintuitive properties of frequentist statistics 81
3.3.6 Why isn't everyone a Bayesian? 83
3.4 Conjugate priors 84
3.4.1 The binomial model 85
3.4.2 The multinomial model 85
3.4.3 The univariate Gaussian model 86
3.4.4 The multivariate Gaussian model 91
3.4.5 The exponential family model 97
3.4.6 Beyond conjugate priors 100
3.5 Noninformative priors 103
3.5.1 Maximum entropy priors 103
3.5.2 Jeffreys priors 104
3.5.3 Invariant priors 107
3.5.4 Reference priors 108
3.6 Hierarchical priors 109
3.6.1 A hierarchical binomial model 109
3.6.2 A hierarchical Gaussian model 112
3.6.3 Hierarchical conditional models 115
3.7 Empirical Bayes 115
3.7.1 EB for the hierarchical binomial model 116
3.7.2 EB for the hierarchical Gaussian model 117
3.7.3 EB for Markov models (n-gram smoothing) 117
3.7.4 EB for non-conjugate models 119
3.8 Model selection 119
3.8.1 Bayesian model selection 120
3.8.2 Bayes model averaging 122
3.8.3 Estimating the marginal likelihood 122
3.8.4 Connection between cross validation and marginal likelihood 123
3.8.5 Conditional marginal likelihood 124
3.8.6 Bayesian leave-one-out (LOO) estimate 125
3.8.7 Information criteria 126
3.9 Model checking 128
3.9.1 Posterior predictive checks 129
3.9.2 Bayesian p-values 131
3.10 Hypothesis testing 132
3.10.1 Frequentist approach 132
3.10.2 Bayesian approach 132
3.10.3 Common statistical tests correspond to inference in linear models 137
3.11 Missing data 142

4 Graphical models 145

4.1 Introduction 145
4.2 Directed graphical models (Bayes nets) 145
4.2.1 Representing the joint distribution 145
4.2.2 Examples 146
4.2.3 Gaussian Bayes nets 150
4.2.4 Conditional independence properties 151
4.2.5 Generation (sampling) 156
4.2.6 Inference 157
4.2.7 Learning 157
4.2.8 Plate notation 163
4.3 Undirected graphical models (Markov random fields) 166
4.3.1 Representing the joint distribution 167
4.3.2 Fully visible MRFs (Ising, Potts, Hopfield, etc.) 168
4.3.3 MRFs with latent variables (Boltzmann machines, etc.) 174
4.3.4 Maximum entropy models 176
4.3.5 Gaussian MRFs 179
4.3.6 Conditional independence properties 181
4.3.7 Generation (sampling) 183
4.3.8 Inference 183
4.3.9 Learning 184
4.4 Conditional random fields (CRFs) 187
4.4.1 1d CRFs 188
4.4.2 2d CRFs 191
4.4.3 Parameter estimation 194
4.4.4 Other approaches to structured prediction 195
4.5 Comparing directed and undirected PGMs 195
4.5.1 CI properties 195
4.5.2 Converting between a directed and undirected model 197
4.5.3 Conditional directed vs undirected PGMs and the label bias problem 198
4.5.4 Combining directed and undirected graphs 199
4.5.5 Comparing directed and undirected Gaussian PGMs 201
4.6 PGM extensions 203
4.6.1 Factor graphs 203
4.6.2 Probabilistic circuits 206
4.6.3 Directed relational PGMs 207
4.6.4 Undirected relational PGMs 209
4.6.5 Open-universe probability models 212
4.6.6 Programs as probability models 212
4.7 Structural causal models 213
4.7.1 Example: causal impact of education on wealth 214
4.7.2 Structural equation models 215
4.7.3 Do operator and augmented DAGs 215
4.7.4 Counterfactuals 216

5 Information theory 219

5.1 KL divergence 219
5.1.1 Desiderata 220
5.1.2 The KL divergence uniquely satisfies the desiderata 221
5.1.3 Thinking about KL 224
5.1.4 Minimizing KL 225
5.1.5 Properties of KL 228
5.1.6 KL divergence and MLE 230
5.1.7 KL divergence and Bayesian inference 231
5.1.8 KL divergence and exponential families 232
5.1.9 Approximating KL divergence using the Fisher information matrix 233

5.1.10 Bregman divergence 233
5.2 Entropy 234
5.2.1 Definition 235
5.2.2 Differential entropy for continuous random variables 235
5.2.3 Typical sets 236
5.2.4 Cross entropy and perplexity 237
5.3 Mutual information 238
5.3.1 Definition 238
5.3.2 Interpretation 239
5.3.3 Data processing inequality 239
5.3.4 Sufficient statistics 240
5.3.5 Multivariate mutual information 241
5.3.6 Variational bounds on mutual information 244
5.3.7 Relevance networks 246
5.4 Data compression (source coding) 247
5.4.1 Lossless compression 247
5.4.2 Lossy compression and the rate-distortion tradeoff 248
5.4.3 Bits back coding 250
5.5 Error-correcting codes (channel coding) 251
5.6 The information bottleneck 252
5.6.1 Vanilla IB 252
5.6.2 Variational IB 253
5.6.3 Conditional entropy bottleneck 254

6 Optimization 257

6.1 Introduction 257
6.2 Automatic differentiation 257
6.2.1 Differentiation in functional form 257
6.2.2 Differentiating chains, circuits, and programs 262
6.3 Stochastic optimization 267
6.3.1 Stochastic gradient descent 267
6.3.2 SGD for optimizing a finite-sum objective 269
6.3.3 SGD for optimizing the parameters of a distribution 269
6.3.4 Score function estimator (likelihood ratio trick) 270
6.3.5 Reparameterization trick 271
6.3.6 Gumbel softmax trick 273
6.3.7 Stochastic computation graphs 274
6.3.8 Straight-through estimator 275
6.4 Natural gradient descent 275
6.4.1 Defining the natural gradient 276
6.4.2 Interpretations of NGD 277
6.4.3 Benefits of NGD 278
6.4.4 Approximating the natural gradient 278
6.4.5 Natural gradients for the exponential family 280
6.5 Bound optimization (MM) algorithms 283
6.5.1 The general algorithm 283
6.5.2 Example: logistic regression 284
6.5.3 The EM algorithm 285
6.5.4 Example: EM for an MVN with missing data 287
6.5.5 Example: robust linear regression using Student likelihood 289
6.5.6 Extensions to EM 291
6.6 Bayesian optimization 293
6.6.1 Sequential model-based optimization 294
6.6.2 Surrogate functions 294
6.6.3 Acquisition functions 296
6.6.4 Other issues 299

6.7 Derivative-free optimization 300
6.7.1 Local search 300
6.7.2 Simulated annealing 303
6.7.3 Evolutionary algorithms 303
6.7.4 Estimation of distribution (EDA) algorithms 306
6.7.5 Cross-entropy method 307
6.7.6 Evolutionary strategies 308
6.8 Optimal transport 309
6.8.1 Warm-up: matching optimally two families of points 309
6.8.2 From optimal matchings to Kantorovich and Monge formulations 310
6.8.3 Solving optimal transport 313
6.9 Submodular optimization 317
6.9.1 Intuition, examples, and background 318
6.9.2 Submodular basic definitions 320
6.9.3 Example submodular functions 322
6.9.4 Submodular optimization 324
6.9.5 Applications of submodularity in machine learning and AI 328
6.9.6 Sketching, coresets, distillation, and data subset and feature Selection 329
6.9.7 Combinatorial information functions 332
6.9.8 Clustering, data partitioning, and parallel machine learning 334
6.9.9 Active and semi-supervised learning 334
6.9.10 Probabilistic modeling 335
6.9.11 Structured norms and loss functions 336
6.9.12 Conclusions 337

II Inference 339

7 Inference algorithms: an overview 341
7.1 Introduction 341
7.2 Common inference patterns 341
7.2.1 Global latents 342
7.2.2 Local latents 342
7.2.3 Global and local latents 343
7.3 Exact inference algorithms 343
7.4 Approximate inference algorithms 344
7.4.1 The MAP approximation and its problems 344
7.4.2 Grid approximation 346
7.4.3 Laplace (quadratic) approximation 347
7.4.4 Variational inference 348
7.4.5 Markov chain Monte Carlo (MCMC) 349
7.4.6 Sequential Monte Carlo 351
7.4.7 Challenging posteriors 351
7.5 Evaluating approximate inference algorithms 352

8 Gaussian filtering and smoothing 353
8.1 Introduction 353
8.1.1 Inferential goals 353
8.1.2 Bayesian filtering equations 355
8.1.3 Bayesian smoothing equations 356
8.1.4 The Gaussian ansatz 357
8.2 Inference for linear-Gaussian SSMs 357
8.2.1 Examples 358
8.2.2 The Kalman filter 359
8.2.3 The Kalman (RTS) smoother 363
8.2.4 Information form filtering and smoothing 366
8.3 Inference based on local linearization 369

8.3.1 Taylor series expansion 369
8.3.2 The extended Kalman filter (EKF) 370
8.3.3 The extended Kalman smoother (EKS) 373
8.4 Inference based on the unscented transform 373
8.4.1 The unscented transform 373
8.4.2 The unscented Kalman filter (UKF) 376
8.4.3 The unscented Kalman smoother (UKS) 376
8.5 Other variants of the Kalman filter 376
8.5.1 General Gaussian filtering 376
8.5.2 Conditional moment Gaussian filtering 379
8.5.3 Iterated filters and smoothers 380
8.5.4 Ensemble Kalman filter 382
8.5.5 Robust Kalman filters 383
8.5.6 Dual EKF 383
8.6 Assumed density filtering 383
8.6.1 Connection with Gaussian filtering 385
8.6.2 ADF for SLDS (Gaussian sum filter) 386
8.6.3 ADF for online logistic regression 387
8.6.4 ADF for online DNNs 390
8.7 Other inference methods for SSMs 390
8.7.1 Grid-based approximations 390
8.7.2 Expectation propagation 391
8.7.3 Variational inference 392
8.7.4 MCMC 392
8.7.5 Particle filtering 392

9 Message passing algorithms 395

9.1 Introduction 395
9.2 Belief propagation on chains 395
9.2.1 Hidden Markov Models 396
9.2.2 The forwards algorithm 397
9.2.3 The forwards-backwards algorithm 398
9.2.4 Forwards filtering backwards smoothing 401
9.2.5 Time and space complexity 402
9.2.6 The Viterbi algorithm 403
9.2.7 Forwards filtering backwards sampling 406
9.3 Belief propagation on trees 406
9.3.1 Directed vs undirected trees 406
9.3.2 Sum-product algorithm 408
9.3.3 Max-product algorithm 409
9.4 Loopy belief propagation 411
9.4.1 Loopy BP for pairwise undirected graphs 412
9.4.2 Loopy BP for factor graphs 412
9.4.3 Gaussian belief propagation 413
9.4.4 Convergence 415
9.4.5 Accuracy 417
9.4.6 Generalized belief propagation 418
9.4.7 Convex BP 418
9.4.8 Application: error correcting codes 418
9.4.9 Application: affinity propagation 420
9.4.10 Emulating BP with graph neural nets 421
9.5 The variable elimination (VE) algorithm 422
9.5.1 Derivation of the algorithm 422
9.5.2 Computational complexity of VE 424
9.5.3 Picking a good elimination order 426
9.5.4 Computational complexity of exact inference 426

9.5.5 Drawbacks of VE 427
9.6 The junction tree algorithm (JTA) 428
9.7 Inference as optimization 429
9.7.1 Inference as backpropagation 429
9.7.2 Perturb and MAP 430

10 Variational inference 433

10.1 Introduction 433
10.1.1 The variational objective 433
10.1.2 Form of the variational posterior 435
10.1.3 Parameter estimation using variational EM 436
10.1.4 Stochastic VI 438
10.1.5 Amortized VI 438
10.1.6 Semi-amortized inference 439
10.2 Gradient-based VI 439
10.2.1 Reparameterized VI 440
10.2.2 Automatic differentiation VI 446
10.2.3 Blackbox variational inference 448
10.3 Coordinate ascent VI 449
10.3.1 Derivation of CAVI algorithm 450
10.3.2 Example: CAVI for the Ising model 452
10.3.3 Variational Bayes 453
10.3.4 Example: VB for a univariate Gaussian 454
10.3.5 Variational Bayes EM 457
10.3.6 Example: VBEM for a GMM 458
10.3.7 Variational message passing (VMP) 464
10.3.8 Autoconj 465
10.4 More accurate variational posteriors 465
10.4.1 Structured mean field 465
10.4.2 Hierarchical (auxiliary variable) posteriors 465
10.4.3 Normalizing flow posteriors 466
10.4.4 Implicit posteriors 466
10.4.5 Combining VI with MCMC inference 466
10.5 Tighter bounds 467
10.5.1 Multi-sample ELBO (IWAE bound) 467
10.5.2 The thermodynamic variational objective (TVO) 468
10.5.3 Minimizing the evidence upper bound 468
10.6 Wake-sleep algorithm 469
10.6.1 Wake phase 469
10.6.2 Sleep phase 470
10.6.3 Daydream phase 471
10.6.4 Summary of algorithm 471
10.7 Expectation propagation (EP) 472
10.7.1 Algorithm 472
10.7.2 Example 474
10.7.3 EP as generalized ADF 474
10.7.4 Optimization issues 475
10.7.5 Power EP and α-divergence 475
10.7.6 Stochastic EP 475

11 Monte Carlo methods 477

11.1 Introduction 477
11.2 Monte Carlo integration 477
11.2.1 Example: estimating π by Monte Carlo integration 478
11.2.2 Accuracy of Monte Carlo integration 478
11.3 Generating random samples from simple distributions 480
11.3.1 Sampling using the inverse cdf 480

11.3.2 Sampling from a Gaussian (Box-Muller method) 481
11.4 Rejection sampling 481
11.4.1 Basic idea 482
11.4.2 Example 483
11.4.3 Adaptive rejection sampling 483
11.4.4 Rejection sampling in high dimensions 484
11.5 Importance sampling 484
11.5.1 Direct importance sampling 485
11.5.2 Self-normalized importance sampling 485
11.5.3 Choosing the proposal 486
11.5.4 Annealed importance sampling (AIS) 486
11.6 Controlling Monte Carlo variance 488
11.6.1 Common random numbers 488
11.6.2 Rao-Blackwellization 488
11.6.3 Control variates 489
11.6.4 Antithetic sampling 490
11.6.5 Quasi-Monte Carlo (QMC) 491

12 Markov chain Monte Carlo 493

12.1 Introduction 493
12.2 Metropolis-Hastings algorithm 494
12.2.1 Basic idea 494
12.2.2 Why MH works 495
12.2.3 Proposal distributions 496
12.2.4 Initialization 498
12.3 Gibbs sampling 499
12.3.1 Basic idea 499
12.3.2 Gibbs sampling is a special case of MH 499
12.3.3 Example: Gibbs sampling for Ising models 500
12.3.4 Example: Gibbs sampling for Potts models 502
12.3.5 Example: Gibbs sampling for GMMs 502
12.3.6 Metropolis within Gibbs 504
12.3.7 Blocked Gibbs sampling 504
12.3.8 Collapsed Gibbs sampling 505
12.4 Auxiliary variable MCMC 507
12.4.1 Slice sampling 507
12.4.2 Swendsen-Wang 509
12.5 Hamiltonian Monte Carlo (HMC) 510
12.5.1 Hamiltonian mechanics 511
12.5.2 Integrating Hamilton's equations 511
12.5.3 The HMC algorithm 513
12.5.4 Tuning HMC 514
12.5.5 Riemann manifold HMC 515
12.5.6 Langevin Monte Carlo (MALA) 515
12.5.7 Connection between SGD and Langevin sampling 516
12.5.8 Applying HMC to constrained parameters 517
12.5.9 Speeding up HMC 518
12.6 MCMC convergence 518
12.6.1 Mixing rates of Markov chains 519
12.6.2 Practical convergence diagnostics 520
12.6.3 Effective sample size 523
12.6.4 Improving speed of convergence 525
12.6.5 Non-centered parameterizations and Neal's funnel 525
12.7 Stochastic gradient MCMC 526
12.7.1 Stochastic gradient Langevin dynamics (SGLD) 527
12.7.2 Preconditionining 527

12.7.3 Reducing the variance of the gradient estimate 528
12.7.4 SG-HMC 529
12.7.5 Underdamped Langevin dynamics 529
12.8 Reversible jump (transdimensional) MCMC 530
12.8.1 Basic idea 531
12.8.2 Example 531
12.8.3 Discussion 533
12.9 Annealing methods 533
12.9.1 Simulated annealing 533
12.9.2 Parallel tempering 536

13 Sequential Monte Carlo 537

13.1 Introduction 537
13.1.1 Problem statement 537
13.1.2 Particle filtering for state-space models 537
13.1.3 SMC samplers for static parameter estimation 539
13.2 Particle filtering 539
13.2.1 Importance sampling 539
13.2.2 Sequential importance sampling 541
13.2.3 Sequential importance sampling with resampling 542
13.2.4 Resampling methods 545
13.2.5 Adaptive resampling 547
13.3 Proposal distributions 547
13.3.1 Locally optimal proposal 548
13.3.2 Proposals based on the extended and unscented Kalman filter 549
13.3.3 Proposals based on the Laplace approximation 549
13.3.4 Proposals based on SMC (nested SMC) 551
13.4 Rao-Blackwellized particle filtering (RBPF) 551
13.4.1 Mixture of Kalman filters 551
13.4.2 Example: tracking a maneuvering object 553
13.4.3 Example: FastSLAM 554
13.5 Extensions of the particle filter 557
13.6 SMC samplers 557
13.6.1 Ingredients of an SMC sampler 558
13.6.2 Likelihood tempering (geometric path) 559
13.6.3 Data tempering 561
13.6.4 Sampling rare events and extrema 562
13.6.5 SMC-ABC and likelihood-free inference 563
13.6.6 SMC^2 563
13.6.7 Variational filtering SMC 563
13.6.8 Variational smoothing SMC 564

III Prediction 567

14 Predictive models: an overview 569

14.1 Introduction 569
14.1.1 Types of model 569
14.1.2 Model fitting using ERM, MLE, and MAP 570
14.1.3 Model fitting using Bayes, VI, and generalized Bayes 571
14.2 Evaluating predictive models 572
14.2.1 Proper scoring rules 572
14.2.2 Calibration 572
14.2.3 Beyond evaluating marginal probabilities 576
14.3 Conformal prediction 579
14.3.1 Conformalizing classification 581

14.3.2 Conformalizing regression 581

15 Generalized linear models 583

15.1 Introduction 583
15.1.1 Some popular GLMs 583
15.1.2 GLMs with noncanonical link functions 586
15.1.3 Maximum likelihood estimation 587
15.1.4 Bayesian inference 587
15.2 Linear regression 588
15.2.1 Ordinary least squares 588
15.2.2 Conjugate priors 589
15.2.3 Uninformative priors 591
15.2.4 Informative priors 593
15.2.5 Spike and slab prior 595
15.2.6 Laplace prior (Bayesian lasso) 596
15.2.7 Horseshoe prior 597
15.2.8 Automatic relevancy determination 598
15.2.9 Multivariate linear regression 600
15.3 Logistic regression 602
15.3.1 Binary logistic regression 602
15.3.2 Multinomial logistic regression 603
15.3.3 Dealing with class imbalance and the long tail 604
15.3.4 Parameter priors 604
15.3.5 Laplace approximation to the posterior 605
15.3.6 Approximating the posterior predictive distribution 607
15.3.7 MCMC inference 609
15.3.8 Other approximate inference methods 610
15.3.9 Case study: is Berkeley admissions biased against women? 611
15.4 Probit regression 613
15.4.1 Latent variable interpretation 613
15.4.2 Maximum likelihood estimation 614
15.4.3 Bayesian inference 616
15.4.4 Ordinal probit regression 616
15.4.5 Multinomial probit models 617
15.5 Multilevel (hierarchical) GLMs 617
15.5.1 Generalized linear mixed models (GLMMs) 618
15.5.2 Example: radon regression 618

16 Deep neural networks 623

16.1 Introduction 623
16.2 Building blocks of differentiable circuits 623
16.2.1 Linear layers 624
16.2.2 Nonlinearities 624
16.2.3 Convolutional layers 625
16.2.4 Residual (skip) connections 626
16.2.5 Normalization layers 627
16.2.6 Dropout layers 627
16.2.7 Attention layers 628
16.2.8 Recurrent layers 630
16.2.9 Multiplicative layers 631
16.2.10 Implicit layers 632
16.3 Canonical examples of neural networks 632
16.3.1 Multilayer perceptrons (MLPs) 632
16.3.2 Convolutional neural networks (CNNs) 633
16.3.3 Autoencoders 634
16.3.4 Recurrent neural networks (RNNs) 636
16.3.5 Transformers 636

16.3.6 Graph neural networks (GNNs) 637

17 Bayesian neural networks 639

17.1 Introduction 639
17.2 Priors for BNNs 639
17.2.1 Gaussian priors 640
17.2.2 Sparsity-promoting priors 642
17.2.3 Learning the prior 642
17.2.4 Priors in function space 642
17.2.5 Architectural priors 643
17.3 Posteriors for BNNs 643
17.3.1 Monte Carlo dropout 643
17.3.2 Laplace approximation 644
17.3.3 Variational inference 645
17.3.4 Expectation propagation 646
17.3.5 Last layer methods 646
17.3.6 SNGP 647
17.3.7 MCMC methods 647
17.3.8 Methods based on the SGD trajectory 648
17.3.9 Deep ensembles 649
17.3.10 Approximating the posterior predictive distribution 653
17.3.11 Tempered and cold posteriors 656
17.4 Generalization in Bayesian deep learning 657
17.4.1 Sharp vs flat minima 657
17.4.2 Mode connectivity and the loss landscape 658
17.4.3 Effective dimensionality of a model 658
17.4.4 The hypothesis space of DNNs 660
17.4.5 PAC-Bayes 660
17.4.6 Out-of-distribution generalization for BNNs 661
17.4.7 Model selection for BNNs 663
17.5 Online inference 663
17.5.1 Sequential Laplace for DNNs 664
17.5.2 Extended Kalman filtering for DNNs 665
17.5.3 Assumed density filtering for DNNs 667
17.5.4 Online variational inference for DNNs 668
17.6 Hierarchical Bayesian neural networks 669
17.6.1 Example: multimoons classification 670

18 Gaussian processes 673

18.1 Introduction 673
18.1.1 GPs: what and why? 673
18.2 Mercer kernels 675
18.2.1 Stationary kernels 676
18.2.2 Nonstationary kernels 681
18.2.3 Kernels for nonvectorial (structured) inputs 682
18.2.4 Making new kernels from old 682
18.2.5 Mercer's theorem 683
18.2.6 Approximating kernels with random features 684
18.3 GPs with Gaussian likelihoods 685
18.3.1 Predictions using noise-free observations 685
18.3.2 Predictions using noisy observations 686
18.3.3 Weight space vs function space 687
18.3.4 Semiparametric GPs 688
18.3.5 Marginal likelihood 689
18.3.6 Computational and numerical issues 689
18.3.7 Kernel ridge regression 690
18.4 GPs with non-Gaussian likelihoods 693

18.4.1 Binary classification 694
18.4.2 Multiclass classification 695
18.4.3 GPs for Poisson regression (Cox process) 696
18.4.4 Other likelihoods 696
18.5 Scaling GP inference to large datasets 697
18.5.1 Subset of data 697
18.5.2 Nyström approximation 698
18.5.3 Inducing point methods 699
18.5.4 Sparse variational methods 702
18.5.5 Exploiting parallelization and structure via kernel matrix multiplies 706
18.5.6 Converting a GP to an SSM 708
18.6 Learning the kernel 709
18.6.1 Empirical Bayes for the kernel parameters 709
18.6.2 Bayesian inference for the kernel parameters 712
18.6.3 Multiple kernel learning for additive kernels 713
18.6.4 Automatic search for compositional kernels 714
18.6.5 Spectral mixture kernel learning 717
18.6.6 Deep kernel learning 718
18.7 GPs and DNNs 720
18.7.1 Kernels derived from infinitely wide DNNs (NN-GP) 721
18.7.2 Neural tangent kernel (NTK) 723
18.7.3 Deep GPs 723
18.8 Gaussian processes for time series forecasting 724
18.8.1 Example: Mauna Loa 724

19 Beyond the iid assumption 727

19.1 Introduction 727
19.2 Distribution shift 727
19.2.1 Motivating examples 727
19.2.2 A causal view of distribution shift 729
19.2.3 The four main types of distribution shift 730
19.2.4 Selection bias 732
19.3 Detecting distribution shifts 732
19.3.1 Detecting shifts using two-sample testing 733
19.3.2 Detecting single out-of-distribution (OOD) inputs 733
19.3.3 Selective prediction 736
19.3.4 Open set and open world recognition 737
19.4 Robustness to distribution shifts 737
19.4.1 Data augmentation 738
19.4.2 Distributionally robust optimization 738
19.5 Adapting to distribution shifts 738
19.5.1 Supervised adaptation using transfer learning 738
19.5.2 Weighted ERM for covariate shift 740
19.5.3 Unsupervised domain adaptation for covariate shift 741
19.5.4 Unsupervised techniques for label shift 742
19.5.5 Test-time adaptation 742
19.6 Learning from multiple distributions 743
19.6.1 Multitask learning 743
19.6.2 Domain generalization 744
19.6.3 Invariant risk minimization 746
19.6.4 Meta learning 747
19.7 Continual learning 750
19.7.1 Domain drift 750
19.7.2 Concept drift 751
19.7.3 Task incremental learning 752
19.7.4 Catastrophic forgetting 753

19.7.5 Online learning 755
19.8 Adversarial examples 756
19.8.1 Whitebox (gradient-based) attacks 758
19.8.2 Blackbox (gradient-free) attacks 759
19.8.3 Real world adversarial attacks 760
19.8.4 Defenses based on robust optimization 760
19.8.5 Why models have adversarial examples 761

IV Generation 763

20 Generative models: an overview 765

20.1 Introduction 765
20.2 Types of generative model 765
20.3 Goals of generative modeling 767
20.3.1 Generating data 767
20.3.2 Density estimation 769
20.3.3 Imputation 770
20.3.4 Structure discovery 771
20.3.5 Latent space interpolation 771
20.3.6 Latent space arithmetic 773
20.3.7 Generative design 774
20.3.8 Model-based reinforcement learning 774
20.3.9 Representation learning 774
20.3.10 Data compression 774
20.4 Evaluating generative models 774
20.4.1 Likelihood-based evaluation 775
20.4.2 Distances and divergences in feature space 776
20.4.3 Precision and recall metrics 777
20.4.4 Statistical tests 778
20.4.5 Challenges with using pretrained classifiers 779
20.4.6 Using model samples to train classifiers 779
20.4.7 Assessing overfitting 779
20.4.8 Human evaluation 780

21 Variational autoencoders 781

21.1 Introduction 781
21.2 VAE basics 781
21.2.1 Modeling assumptions 782
21.2.2 Model fitting 783
21.2.3 Comparison of VAEs and autoencoders 783
21.2.4 VAEs optimize in an augmented space 784
21.3 VAE generalizations 786
21.3.1 β-VAE 787
21.3.2 InfoVAE 789
21.3.3 Multimodal VAEs 790
21.3.4 Semisupervised VAEs 793
21.3.5 VAEs with sequential encoders/decoders 794
21.4 Avoiding posterior collapse 796
21.4.1 KL annealing 797
21.4.2 Lower bounding the rate 798
21.4.3 Free bits 798
21.4.4 Adding skip connections 798
21.4.5 Improved variational inference 798
21.4.6 Alternative objectives 799
21.5 VAEs with hierarchical structure 799

21.5.1 Bottom-up vs top-down inference 800
21.5.2 Example: very deep VAE 801
21.5.3 Connection with autoregressive models 802
21.5.4 Variational pruning 804
21.5.5 Other optimization difficulties 804
21.6 Vector quantization VAE 805
21.6.1 Autoencoder with binary code 805
21.6.2 VQ-VAE model 805
21.6.3 Learning the prior 807
21.6.4 Hierarchical extension (VQ-VAE-2) 807
21.6.5 Discrete VAE 808
21.6.6 VQ-GAN 809

22 Autoregressive models 811
22.1 Introduction 811
22.2 Neural autoregressive density estimators (NADE) 812
22.3 Causal CNNs 812
22.3.1 1d causal CNN (convolutional Markov models) 813
22.3.2 2d causal CNN (PixelCNN) 813
22.4 Transformers 814
22.4.1 Text generation (GPT, etc.) 815
22.4.2 Image generation (DALL-E, etc.) 816
22.4.3 Other applications 817

23 Normalizing flows 819
23.1 Introduction 819
23.1.1 Preliminaries 819
23.1.2 How to train a flow model 821
23.2 Constructing flows 822
23.2.1 Affine flows 822
23.2.2 Elementwise flows 822
23.2.3 Coupling flows 825
23.2.4 Autoregressive flows 826
23.2.5 Residual flows 832
23.2.6 Continuous-time flows 834
23.3 Applications 836
23.3.1 Density estimation 836
23.3.2 Generative modeling 836
23.3.3 Inference 837

24 Energy-based models 839
24.1 Introduction 839
24.1.1 Example: products of experts (PoE) 840
24.1.2 Computational difficulties 840
24.2 Maximum likelihood training 841
24.2.1 Gradient-based MCMC methods 842
24.2.2 Contrastive divergence 842
24.3 Score matching (SM) 846
24.3.1 Basic score matching 846
24.3.2 Denoising score matching (DSM) 847
24.3.3 Sliced score matching (SSM) 848
24.3.4 Connection to contrastive divergence 849
24.3.5 Score-based generative models 850
24.4 Noise contrastive estimation 850
24.4.1 Connection to score matching 852
24.5 Other methods 852
24.5.1 Minimizing Differences/Derivatives of KL Divergences 853

24.5.2 Minimizing the Stein discrepancy 853
24.5.3 Adversarial training 854

25 Diffusion models 857

25.1 Introduction 857
25.2 Denoising diffusion probabilistic models (DDPMs) 857
25.2.1 Encoder (forwards diffusion) 858
25.2.2 Decoder (reverse diffusion) 859
25.2.3 Model fitting 860
25.2.4 Learning the noise schedule 861
25.2.5 Example: image generation 863
25.3 Score-based generative models (SGMs) 864
25.3.1 Example 864
25.3.2 Adding noise at multiple scales 864
25.3.3 Equivalence to DDPM 866
25.4 Continuous time models using differential equations 867
25.4.1 Forwards diffusion SDE 867
25.4.2 Forwards diffusion ODE 868
25.4.3 Reverse diffusion SDE 869
25.4.4 Reverse diffusion ODE 870
25.4.5 Comparison of the SDE and ODE approach 871
25.4.6 Example 871
25.5 Speeding up diffusion models 871
25.5.1 DDIM sampler 872
25.5.2 Non-Gaussian decoder networks 872
25.5.3 Distillation 873
25.5.4 Latent space diffusion 874
25.6 Conditional generation 875
25.6.1 Conditional diffusion model 875
25.6.2 Classifier guidance 875
25.6.3 Classifier-free guidance 876
25.6.4 Generating high resolution images 876
25.7 Diffusion for discrete state spaces 877
25.7.1 Discrete Denoising Diffusion Probabilistic Models 877
25.7.2 Choice of Markov transition matrices for the forward processes 878
25.7.3 Parameterization of the reverse process 879
25.7.4 Noise schedules 880
25.7.5 Connections to other probabilistic models for discrete sequences 880

26 Generative adversarial networks 883

26.1 Introduction 883
26.2 Learning by comparison 884
26.2.1 Guiding principles 885
26.2.2 Density ratio estimation using binary classifiers 886
26.2.3 Bounds on f-divergences 888
26.2.4 Integral probability metrics 890
26.2.5 Moment matching 892
26.2.6 On density ratios and differences 892
26.3 Generative adversarial networks 894
26.3.1 From learning principles to loss functions 894
26.3.2 Gradient descent 895
26.3.3 Challenges with GAN training 897
26.3.4 Improving GAN optimization 898
26.3.5 Convergence of GAN training 898
26.4 Conditional GANs 902
26.5 Inference with GANs 903
26.6 Neural architectures in GANs 904

26.6.1 The importance of discriminator architectures 904
26.6.2 Architectural inductive biases 905
26.6.3 Attention in GANs 905
26.6.4 Progressive generation 906
26.6.5 Regularization 907
26.6.6 Scaling up GAN models 908
26.7 Applications 908
26.7.1 GANs for image generation 908
26.7.2 Video generation 911
26.7.3 Audio generation 912
26.7.4 Text generation 912
26.7.5 Imitation learning 913
26.7.6 Domain adaptation 914
26.7.7 Design, art and creativity 914

V Discovery 915

27 Discovery methods: an overview 917
27.1 Introduction 917
27.2 Overview of Part V 918

28 Latent factor models 919
28.1 Introduction 919
28.2 Mixture models 919
28.2.1 Gaussian mixture models (GMMs) 920
28.2.2 Bernoulli mixture models 922
28.2.3 Gaussian scale mixtures (GSMs) 922
28.2.4 Using GMMs as a prior for inverse imaging problems 924
28.2.5 Using mixture models for classification problems 927
28.3 Factor analysis 929
28.3.1 Factor analysis: the basics 929
28.3.2 Probabilistic PCA 934
28.3.3 Mixture of factor analyzers 936
28.3.4 Factor analysis models for paired data 943
28.3.5 Factor analysis with exponential family likelihoods 945
28.3.6 Factor analysis with DNN likelihoods (VAEs) 948
28.3.7 Factor analysis with GP likelihoods (GP-LVM) 948
28.4 LFMs with non-Gaussian priors 949
28.4.1 Non-negative matrix factorization (NMF) 949
28.4.2 Multinomial PCA 950
28.5 Topic models 953
28.5.1 Latent Dirichlet allocation (LDA) 953
28.5.2 Correlated topic model 957
28.5.3 Dynamic topic model 957
28.5.4 LDA-HMM 958
28.6 Independent components analysis (ICA) 962
28.6.1 Noiseless ICA model 962
28.6.2 The need for non-Gaussian priors 963
28.6.3 Maximum likelihood estimation 964
28.6.4 Alternatives to MLE 965
28.6.5 Sparse coding 966
28.6.6 Nonlinear ICA 967

29 State-space models 969
29.1 Introduction 969
29.2 Hidden Markov models (HMMs) 970

29.2.1 Conditional independence properties 970
29.2.2 State transition model 970
29.2.3 Discrete likelihoods 971
29.2.4 Gaussian likelihoods 972
29.2.5 Autoregressive likelihoods 972
29.2.6 Neural network likelihoods 973
29.3 HMMs: applications 974
29.3.1 Time series segmentation 974
29.3.2 Protein sequence alignment 976
29.3.3 Spelling correction 978
29.4 HMMs: parameter learning 980
29.4.1 The Baum-Welch (EM) algorithm 980
29.4.2 Parameter estimation using SGD 983
29.4.3 Parameter estimation using spectral methods 984
29.4.4 Bayesian HMMs 985
29.5 HMMs: generalizations 987
29.5.1 Hidden semi-Markov model (HSMM) 987
29.5.2 Hierarchical HMMs 989
29.5.3 Factorial HMMs 991
29.5.4 Coupled HMMs 992
29.5.5 Dynamic Bayes nets (DBN) 992
29.5.6 Changepoint detection 993
29.6 Linear dynamical systems (LDSs) 996
29.6.1 Conditional independence properties 996
29.6.2 Parameterization 996
29.7 LDS: applications 997
29.7.1 Object tracking and state estimation 997
29.7.2 Online Bayesian linear regression (recursive least squares) 998
29.7.3 Adaptive filtering 1000
29.7.4 Time series forecasting 1000
29.8 LDS: parameter learning 1001
29.8.1 EM for LDS 1001
29.8.2 Subspace identification methods 1003
29.8.3 Ensuring stability of the dynamical system 1003
29.8.4 Bayesian LDS 1004
29.9 Switching linear dynamical systems (SLDSs) 1005
29.9.1 Parameterization 1005
29.9.2 Posterior inference 1006
29.9.3 Application: Multitarget tracking 1006
29.10 Nonlinear SSMs 1009
29.10.1 Example: object tracking and state estimation 1010
29.10.2 Posterior inference 1010
29.11 Non-Gaussian SSMs 1010
29.11.1 Example: spike train modeling 1011
29.11.2 Example: stochastic volatility models 1012
29.11.3 Posterior inference 1012
29.12 Structural time series models 1012
29.12.1 Introduction 1013
29.12.2 Structural building blocks 1013
29.12.3 Model fitting 1016
29.12.4 Forecasting 1016
29.12.5 Examples 1016
29.12.6 Causal impact of a time series intervention 1020
29.12.7 Prophet 1024
29.12.8 Neural forecasting methods 1024
29.13 Deep SSMs 1025

29.13.1 Deep Markov models 1026
29.13.2 Recurrent SSM 1027
29.13.3 Improving multistep predictions 1027
29.13.4 Variational RNNs 1028

30 Graph learning 1031

30.1 Introduction 1031
30.2 Latent variable models for graphs 1031
30.3 Graphical model structure learning 1031

31 Nonparametric Bayesian models 1035

31.1 Introduction 1035

32 Representation learning 1037

32.1 Introduction 1037
32.2 Evaluating and comparing learned representations 1037
32.2.1 Downstream performance 1038
32.2.2 Representational similarity 1040
32.3 Approaches for learning representations 1044
32.3.1 Supervised representation learning and transfer 1045
32.3.2 Generative representation learning 1047
32.3.3 Self-supervised representation learning 1049
32.3.4 Multiview representation learning 1052
32.4 Theory of representation learning 1057
32.4.1 Identifiability 1057
32.4.2 Information maximization 1058

33 Interpretability 1061

33.1 Introduction 1061
33.1.1 The role of interpretability: unknowns and under-specifications 1062
33.1.2 Terminology and framework 1063
33.2 Methods for interpretable machine learning 1066
33.2.1 Inherently interpretable models: the model is its explanation 1067
33.2.2 Semi-inherently interpretable models: example-based methods 1069
33.2.3 Post-hoc or joint training: the explanation gives a partial view of the model 1069
33.2.4 Transparency and visualization 1073
33.3 Properties: the abstraction between context and method 1074
33.3.1 Properties of explanations from interpretable machine learning 1074
33.3.2 Properties of explanations from cognitive science 1076
33.4 Evaluation of interpretable machine learning models 1077
33.4.1 Computational evaluation: does the method have desired properties? 1078
33.4.2 User study-based evaluation: does the method help a user perform a target task? 1082
33.5 Discussion: how to think about interpretable machine learning 1086

VI Action 1091

34 Decision making under uncertainty 1093

34.1 Statistical decision theory 1093
34.1.1 Basics 1093
34.1.2 Frequentist decision theory 1093
34.1.3 Bayesian decision theory 1094
34.1.4 Frequentist optimality of the Bayesian approach 1095
34.1.5 Examples of one-shot decision making problems 1095
34.2 Decision (influence) diagrams 1099
34.2.1 Example: oil wildcatter 1100
34.2.2 Information arcs 1101

34.2.3 Value of information 1101
34.2.4 Computing the optimal policy 1102
34.3 A/B testing 1103
34.3.1 A Bayesian approach 1103
34.3.2 Example 1106
34.4 Contextual bandits 1107
34.4.1 Types of bandit 1108
34.4.2 Applications 1109
34.4.3 Exploration-exploitation tradeoff 1109
34.4.4 The optimal solution 1110
34.4.5 Upper confidence bounds (UCBs) 1111
34.4.6 Thompson sampling 1113
34.4.7 Regret 1114
34.5 Markov decision problems 1116
34.5.1 Basics 1116
34.5.2 Partially observed MDPs 1117
34.5.3 Episodes and returns 1117
34.5.4 Value functions 1119
34.5.5 Optimal value functions and policies 1119
34.6 Planning in an MDP 1120
34.6.1 Value iteration 1121
34.6.2 Policy iteration 1122
34.6.3 Linear programming 1123
34.7 Active learning 1124
34.7.1 Active learning scenarios 1124
34.7.2 Relationship to other forms of sequential decision making 1125
34.7.3 Acquisition strategies 1126
34.7.4 Batch active learning 1128

35 Reinforcement learning 1133

35.1 Introduction 1133
35.1.1 Overview of methods 1133
35.1.2 Value-based methods 1135
35.1.3 Policy search methods 1135
35.1.4 Model-based RL 1135
35.1.5 Exploration-exploitation tradeoff 1136
35.2 Value-based RL 1138
35.2.1 Monte Carlo RL 1138
35.2.2 Temporal difference (TD) learning 1138
35.2.3 TD learning with eligibility traces 1139
35.2.4 SARSA: on-policy TD control 1140
35.2.5 Q-learning: off-policy TD control 1141
35.2.6 Deep Q-network (DQN) 1142
35.3 Policy-based RL 1144
35.3.1 The policy gradient theorem 1145
35.3.2 REINFORCE 1146
35.3.3 Actor-critic methods 1146
35.3.4 Bound optimization methods 1148
35.3.5 Deterministic policy gradient methods 1150
35.3.6 Gradient-free methods 1151
35.4 Model-based RL 1151
35.4.1 Model predictive control (MPC) 1151
35.4.2 Combining model-based and model-free 1153
35.4.3 MBRL using Gaussian processes 1154
35.4.4 MBRL using DNNs 1155
35.4.5 MBRL using latent-variable models 1156

35.4.6 Robustness to model errors 1158
35.5 Off-policy learning 1158
35.5.1 Basic techniques 1159
35.5.2 The curse of horizon 1162
35.5.3 The deadly triad 1163
35.6 Control as inference 1165
35.6.1 Maximum entropy reinforcement learning 1165
35.6.2 Other approaches 1167
35.6.3 Imitation learning 1168

36 Causality 1171
36.1 Introduction 1171
36.2 Causal formalism 1173
36.2.1 Structural causal models 1173
36.2.2 Causal DAGs 1174
36.2.3 Identification 1176
36.2.4 Counterfactuals and the causal hierarchy 1178
36.3 Randomized control trials 1180
36.4 Confounder adjustment 1181
36.4.1 Causal estimand, statistical estimand, and identification 1181
36.4.2 ATE estimation with observed confounders 1184
36.4.3 Uncertainty quantification 1189
36.4.4 Matching 1189
36.4.5 Practical considerations and procedures 1190
36.4.6 Summary and practical advice 1193
36.5 Instrumental variable strategies 1195
36.5.1 Additive unobserved confounding 1196
36.5.2 Instrument monotonicity and local average treatment effect 1198
36.5.3 Two stage least squares 1201
36.6 Difference in differences 1202
36.6.1 Estimation 1205
36.7 Credibility checks 1206
36.7.1 Placebo checks 1206
36.7.2 Sensitivity analysis to unobserved confounding 1207
36.8 The do-calculus 1215
36.8.1 The three rules 1215
36.8.2 Revisiting backdoor adjustment 1216
36.8.3 Frontdoor adjustment 1217
36.9 Further reading 1218

Index 1221

Bibliography 1239