Contents

Preface **xi**

1 Introduction **1**

1.1 What is machine learning? 1
1.2 Supervised learning 1
1.2.1 Classification 2
1.2.2 Regression 7
1.2.3 Overfitting and generalization 10
1.3 Unsupervised learning 11
1.3.1 Clustering 12
1.3.2 Self-supervised learning 12
1.3.3 Evaluating unsupervised learning 13
1.4 Reinforcement learning 13
1.4.1 Challenges in RL 14
1.4.2 Comparing supervised, unsupervised and reinforcement learning 15
1.5 Discussion 15
1.5.1 The relationship between ML and other fields 16
1.5.2 Structure of the book 16
1.5.3 Caveats 16

I Foundations **19**

2 Probabilistic inference **21**

2.1 Introduction 21
2.2 Bayes' rule 21
2.2.1 Example: testing for COVID-19 22
2.2.2 Example: The Monty Hall problem 23
2.2.3 Inverse problems 24
2.3 Bayesian concept learning 25
2.3.1 Learning a discrete concept: the number game 26
2.3.2 Learning a continuous concept: the healthy levels game 32

2.4 Bayesian machine learning 35
2.4.1 Example: scalar input, binary output 37
2.4.2 Example: binary input, scalar output 38
2.4.3 Scaling up 39

3 Probabilistic models 41

3.1 Bernoulli and binomial distributions 41
3.1.1 Definition 41
3.1.2 Sigmoid (logistic) function 42
3.1.3 Binary logistic regression 44
3.2 Categorical and multinomial distributions 45
3.2.1 Definition 45
3.2.2 Softmax function 45
3.2.3 Multiclass logistic regression 46
3.2.4 Log-sum-exp trick 47
3.3 Univariate Gaussian (normal) distribution 48
3.3.1 Cumulative distribution function 48
3.3.2 Probability density function 49
3.3.3 Regression 50
3.3.4 Why is the Gaussian distribution so widely used? 51
3.3.5 Half-normal 52
3.4 Some other common univariate distributions 52
3.4.1 Student t distribution 52
3.4.2 Cauchy distribution 54
3.4.3 Laplace distribution 54
3.4.4 Beta distribution 55
3.4.5 Gamma distribution 55
3.5 The multivariate Gaussian (normal) distribution 57
3.5.1 Definition 57
3.5.2 Mahalanobis distance 58
3.5.3 Marginals and conditionals of an MVN 60
3.5.4 Example: Imputing missing values 60
3.6 Linear Gaussian systems 61
3.6.1 Example: inferring a latent vector from a noisy sensor 62
3.6.2 Example: inferring a latent vector from multiple noisy sensors 63
3.7 Mixture models 64
3.7.1 Gaussian mixture models 65
3.7.2 Mixtures of Bernoullis 67
3.7.3 Gaussian scale mixtures 68
3.8 Probabilistic graphical models 70
3.8.1 Representation 71
3.8.2 Inference 71
3.8.3 Learning 73

4 Parameter estimation 77

4.1 Introduction 77
4.2 Maximum likelihood estimation (MLE) 77
4.2.1 Definition 77
4.2.2 Justification for MLE 78
4.2.3 Example: MLE for the Bernoulli distribution 79
4.2.4 Example: MLE for the categorical distribution 80
4.2.5 Example: MLE for the univariate Gaussian 81
4.2.6 Example: MLE for the multivariate Gaussian 82
4.2.7 Example: MLE for linear regression 84
4.3 Empirical risk minimization (ERM) 85
4.3.1 Example: minimizing the misclassification rate 85
4.3.2 Surrogate loss 85
4.4 Regularization 86
4.4.1 Example: MAP estimation for the Bernoulli distribution 87
4.4.2 Example: MAP estimation for the multivariate Gaussian 88
4.4.3 Example: weight decay 89
4.4.4 Picking the regularizer using a validation set 91
4.4.5 Cross-validation 91
4.4.6 Early stopping 93
4.4.7 Using more data 93
4.5 The method of moments 95
4.5.1 Example: MOM for the univariate Gaussian 95
4.5.2 Example: MOM for the uniform distribution 95
4.6 Online (recursive) estimation 96
4.6.1 Example: recursive MLE for the mean of a Gaussian 96
4.6.2 Exponentially-weighted moving average (EMA) 97
4.6.3 Bayesian inference 98
4.7 Parameter uncertainty 98

5 Optimization algorithms 99

5.1 Introduction 99
5.1.1 Local vs global optimization 99
5.1.2 Constrained vs unconstrained optimization 101
5.1.3 Convex vs nonconvex optimization 101
5.1.4 Smooth vs nonsmooth optimization 102
5.2 First-order methods 102
5.2.1 Descent direction 103
5.2.2 Step size (learning rate) 103
5.2.3 Convergence rates 105
5.2.4 Momentum methods 106
5.3 Second-order methods 108
5.3.1 Newton's method 108
5.3.2 BFGS and other quasi-Newton methods 109
5.3.3 Trust region methods 110
5.3.4 Natural gradient descent 111

5.4 Stochastic gradient descent 114
5.4.1 Application to finite sum problems 115
5.4.2 Example: SGD for fitting linear regression 115
5.4.3 Choosing the step size 116
5.4.4 Iterate averaging 117
5.4.5 Variance reduction 117
5.4.6 Preconditioned SGD 118
5.5 Constrained optimization 121
5.5.1 Lagrange multipliers 122
5.5.2 The KKT conditions 123
5.5.3 Linear programming 124
5.5.4 Quadratic programming 126
5.5.5 Mixed integer linear programming 127
5.6 Proximal gradient method 127
5.6.1 Projected gradient descent 128
5.6.2 Proximal operator for ℓ_1-norm regularizer 129
5.6.3 Proximal operator for quantization 130
5.7 Bound optimization 131
5.7.1 The general algorithm 131
5.7.2 The EM algorithm 132
5.7.3 Example: EM for a GMM 135
5.7.4 Example: EM for an MVN with missing data 139
5.8 Blackbox and derivative free optimization 142
5.8.1 Grid search and random search 142
5.8.2 Simulated annealing 142
5.8.3 Model-based blackbox optimization 143

6 Information theory 145

6.1 Entropy 145
6.1.1 Entropy for discrete random variables 145
6.1.2 Cross entropy 147
6.1.3 Joint entropy 147
6.1.4 Conditional entropy 148
6.1.5 Perplexity 149
6.1.6 Differential entropy for continuous random variables 149
6.2 Relative entropy (KL divergence) 151
6.2.1 Definition 151
6.2.2 Interpretation 151
6.2.3 Example: KL divergence between two Gaussians 151
6.2.4 Non-negativity of KL 152
6.2.5 KL divergence and MLE 152
6.2.6 Forward vs reverse KL 153
6.3 Mutual information 154
6.3.1 Definition 154
6.3.2 Interpretation 155

6.3.3 Example 155
6.3.4 Conditional mutual information 156
6.3.5 Normalized mutual information 157
6.3.6 MI as a “generalized correlation coefficient” 157
6.3.7 Data processing inequality 159
6.3.8 Sufficient Statistics 160
6.3.9 Fano’s inequality 160

7 Bayesian statistics 163

7.1 Introduction 163
7.1.1 Computing the posterior 163
7.1.2 Summarizing the posterior 163
7.2 Conjugate priors 167
7.2.1 The beta-binomial model 168
7.2.2 The Dirichlet-multinomial model 174
7.2.3 The Gaussian-Gaussian model 178
7.2.4 The multivariate Gaussian-Gaussian model 183
7.2.5 Beyond conjugate priors 189
7.3 Noninformative priors 189
7.3.1 Jeffreys priors 190
7.3.2 Invariant priors 192
7.3.3 Reference priors 193
7.4 Hierarchical priors 193
7.4.1 A hierarchical binomial model 194
7.4.2 A hierarchical Gaussian model 195
7.5 Empirical priors 199
7.5.1 A hierarchical binomial model 199
7.5.2 A hierarchical Gaussian model 200
7.6 Bayesian model comparison 201
7.6.1 Bayesian model selection 201
7.6.2 Bayes model averaging 202
7.6.3 Occam’s razor 203
7.6.4 Connection between cross validation and marginal likelihood 204
7.6.5 Information criteria 206
7.6.6 Bayesian hypothesis testing 208
7.6.7 Group comparisons 210
7.6.8 Posterior predictive checks 212
7.7 Approximate inference algorithms 214
7.7.1 Grid approximation 215
7.7.2 Laplace approximation 215
7.7.3 Variational approximation 216
7.7.4 Markov Chain Monte Carlo (MCMC) approximation 218
7.7.5 Online inference using assumed density filtering 219

8 Bayesian decision theory 221

8.1 Bayesian decision theory 221
8.1.1 Basics 221
8.1.2 Classification problems 222
8.1.3 ROC curves 224
8.1.4 Precision-recall curves 226
8.1.5 Regression problems 228
8.1.6 Probabilistic prediction problems 229
8.2 A/B testing 231
8.2.1 A Bayesian approach 232
8.2.2 Example 234
8.3 Bandit problems 236
8.3.1 Contextual bandits 236
8.3.2 Markov decision processes 237
8.3.3 Exploration-exploitation tradeoff 238
8.3.4 Optimal solution 238
8.3.5 Regret 240
8.3.6 Upper confidence bounds (UCB) 241
8.3.7 Thompson sampling 242
8.3.8 Simple heuristics 243
8.4 Discussion 244
8.4.1 The separation principle and its limits 244
8.4.2 Optimality of the Bayesian approach and its limits 245

II Linear models 247

9 Linear discriminant analysis 249

9.1 Introduction 249
9.2 Gaussian discriminant analysis 249
9.2.1 Quadratic decision boundaries 250
9.2.2 Linear decision boundaries 251
9.2.3 The connection between LDA and logistic regression 251
9.2.4 Model fitting 252
9.2.5 Nearest centroid classifier 254
9.2.6 Fisher's linear discriminant analysis 254
9.3 Naive Bayes classifiers 258
9.3.1 Example models 259
9.3.2 Model fitting 260
9.3.3 Bayesian naive Bayes 261
9.3.4 The connection between naive Bayes and logistic regression 261
9.4 Generative vs discriminative classifiers 262
9.4.1 Advantages of discriminative classifiers 262
9.4.2 Advantages of generative classifiers 263
9.4.3 Handling missing features 263

10 Logistic regression 265

10.1 Introduction 265
10.2 Binary logistic regression 265
10.2.1 Linear classifiers 265
10.2.2 Nonlinear classifiers 266
10.2.3 Maximum likelihood estimation 268
10.2.4 Stochastic gradient descent 271
10.2.5 Perceptron algorithm 272
10.2.6 Iteratively reweighted least squares 272
10.2.7 MAP estimation 274
10.2.8 Standardization 275
10.3 Multinomial logistic regression 276
10.3.1 Linear and nonlinear classifiers 276
10.3.2 Maximum likelihood estimation 277
10.3.3 Gradient-based optimization 279
10.3.4 Bound optimization 279
10.3.5 MAP estimation 281
10.3.6 Maximum entropy classifiers 281
10.3.7 Hierarchical classification 282
10.3.8 Handling large numbers of classes 283
10.4 Preprocessing discrete input data 285
10.4.1 One-hot encoding 285
10.4.2 Feature crosses 285
10.4.3 Dealing with text 286
10.5 Robust logistic regression 288
10.5.1 Mixture model for the likelihood 288
10.5.2 Bi-tempered loss 288
10.6 Bayesian logistic regression 292
10.6.1 Approximating the posterior predictive 292
10.6.2 Laplace approximation 293
10.6.3 MCMC approximation 295
10.6.4 Variational inference 297
10.6.5 Online inference using assumed density filtering 301

11 Linear regression 305

11.1 Introduction 305
11.2 Standard linear regression 305
11.2.1 Terminology 305
11.2.2 Least squares estimation 306
11.2.3 Other approaches to computing the MLE 310
11.2.4 Measuring goodness of fit 313
11.3 Ridge regression 315
11.3.1 Computing the MAP estimate 316
11.3.2 Connection between ridge regression and PCA 317
11.3.3 Choosing the strength of the regularizer 319

11.4 Robust linear regression 319
11.4.1 Robust regression using the Student t distribution 319
11.4.2 Robust regression using the Laplace distribution 321
11.4.3 Robust regression using Huber loss 322
11.4.4 Robust regression by randomly or iteratively removing outliers 323
11.5 Lasso regression 323
11.5.1 MAP estimation with a Laplace prior (ℓ_1 regularization) 323
11.5.2 Why does ℓ_1 regularization yield sparse solutions? 324
11.5.3 Hard vs soft thresholding 325
11.5.4 Regularization path 327
11.5.5 Comparison of least squares, lasso, ridge and subset selection 328
11.5.6 Variable selection consistency 329
11.5.7 Group lasso 330
11.5.8 Elastic net (ridge and lasso combined) 333
11.5.9 Optimization algorithms 333
11.6 Bayesian linear regression 335
11.6.1 Computing $p(\mathbf{w}|\mathcal{D}, \sigma^2)$ with Gaussian prior 335
11.6.2 Computing $p(\mathbf{w}, \sigma^2|\mathcal{D})$ with Gaussian-Gamma prior 339
11.6.3 Uninformative priors 341
11.6.4 Sparsity-promoting priors 343
11.6.5 Hierarchical priors 346
11.6.6 Empirical Bayes (Automatic relevancy determination) 348
11.6.7 Online inference (recursive least squares) 351

12 Generalized linear models 353

12.1 Introduction 353
12.2 The exponential family 353
12.2.1 Definition 353
12.2.2 Examples 354
12.2.3 Log partition function is cumulant generating function 358
12.2.4 MLE for the exponential family 360
12.2.5 Exponential dispersion family 360
12.3 Generalized linear models (GLMs) 361
12.3.1 Examples 361
12.3.2 Maximum likelihood estimation 363
12.3.3 GLMs with non-canonical link functions 364
12.4 Probit regression 365
12.4.1 Latent variable interpretation 365
12.4.2 Maximum likelihood estimation 366
12.4.3 Bayesian inference 367
12.4.4 Ordinal probit regression 367
12.4.5 Multinomial probit models 368

III Deep neural networks 369

13 Neural networks for unstructured data 371

13.1 Introduction 371
13.2 Multilayer perceptrons (MLPs) 372
13.2.1 The XOR problem 372
13.2.2 Differentiable MLPs 373
13.2.3 Activation functions 374
13.2.4 Example models 376
13.2.5 The importance of depth 380
13.2.6 Connections with biology 382
13.3 Backpropagation 384
13.3.1 Forwards pass 384
13.3.2 Backwards pass 384
13.3.3 Automatic differentiation 386
13.3.4 Computation graphs 388
13.4 Training neural networks 390
13.4.1 Tuning the learning rate 390
13.4.2 Vanishing and exploding gradients 390
13.4.3 Residual connections 392
13.4.4 Batch normalization 393
13.4.5 Parameter initialization 395
13.5 Regularization 397
13.5.1 Early stopping 397
13.5.2 Weight decay 397
13.5.3 Sparse DNNs 397
13.5.4 Dropout 399
13.5.5 Bayesian neural networks 399
13.6 Other kinds of feedforward networks 400
13.6.1 Radial basis function networks 400
13.6.2 Mixtures of experts 401

14 Neural networks for images 407

14.1 Introduction 407
14.2 Basics 407
14.2.1 Convolution in 1d 407
14.2.2 Convolution in 2d 409
14.2.3 Convolution as matrix-vector multiplication 410
14.2.4 Boundary conditions and strides 410
14.2.5 Pooling layers 413
14.2.6 Normalization layers 414
14.2.7 Putting it altogether 415
14.3 Image classification using CNNs 415
14.3.1 Common datasets 415
14.3.2 Common models 419

14.4 Solving other discriminative vision tasks with CNNs 423
14.4.1 Image tagging 423
14.4.2 Object detection 424
14.4.3 Human pose estimation 425
14.4.4 Image segmentation 425
14.5 Generating images by inverting CNNs 428
14.5.1 Converting a trained classifier into a generative model 428
14.5.2 Image priors 428
14.5.3 Visualizing the features learned by a CNN 430
14.5.4 Deep Dream 431
14.5.5 Neural style transfer 432
14.6 Adversarial Examples 435
14.6.1 Whitebox (gradient-based) attacks 436
14.6.2 Blackbox (gradient-free) attacks 437
14.6.3 Real world adversarial attacks 438
14.6.4 Defenses based on robust optimization 438
14.6.5 Why models have adversarial examples 439

15 Neural networks for sequences 443

15.1 Introduction 443
15.2 Recurrent neural networks (RNNs) 443
15.2.1 Vec2Seq (sequence generation) 443
15.2.2 Seq2Vec (sequence classification) 446
15.2.3 Seq2Seq (sequence translation) 447
15.2.4 Beam search 449
15.2.5 Backpropagation through time 449
15.2.6 Gating and long term memory 450
15.3 1d CNNs 452
15.3.1 1d CNNs for sequence classification 452
15.3.2 Causal 1d CNNs for sequence generation 453
15.4 Attention 454
15.4.1 Seq2seq with attention 455
15.4.2 Seq2vec with attention 456
15.4.3 Attention as a soft dictionary lookup 456
15.4.4 Soft vs hard attention 458
15.5 Transformers 458
15.5.1 Self-attention 459
15.5.2 Multi-headed attention 460
15.5.3 Positional encoding 461
15.5.4 Putting it altogether 461
15.5.5 Comparing transformers, CNNs and RNNs 462
15.6 Efficient transformers 463
15.6.1 Fixed non-learnable localized attention patterns 463
15.6.2 Learnable sparse attention patterns 464
15.6.3 Memory and recurrence methods 465

15.6.4 Low-rank and kernel methods 465

IV Nonparametric models 469

16 Exemplar-based methods 471

16.1 K nearest neighbor (KNN) classification 471
16.1.1 Example 472
16.1.2 The curse of dimensionality 472
16.1.3 Reducing the speed and memory requirements 474
16.1.4 Open set recognition 474
16.2 Learning distance metrics 475
16.2.1 Linear and convex methods 476
16.2.2 Deep metric learning 477
16.2.3 Classification losses 478
16.2.4 Ranking losses 478
16.2.5 Speeding up ranking loss optimization 480
16.2.6 Other training tricks for DML 483
16.3 Kernel density estimation (KDE) 483
16.3.1 Density kernels 484
16.3.2 Parzen window density estimator 484
16.3.3 How to choose the bandwidth parameter 486
16.3.4 From KDE to KNN classification 486
16.3.5 Kernel regression 487

17 Kernel methods 491

17.1 Inferring functions from data 491
17.1.1 Smoothness prior 492
17.1.2 Inference from noise-free observations 492
17.1.3 Inference from noisy observations 494
17.2 Mercer kernels 494
17.2.1 Mercer's theorem 494
17.2.2 Some popular Mercer kernels 495
17.3 Gaussian processes 500
17.3.1 Noise-free observations 500
17.3.2 Noisy observations 501
17.3.3 Comparison to kernel regression 502
17.3.4 Weight space vs function space 503
17.3.5 Numerical issues 503
17.3.6 Estimating the kernel 504
17.3.7 GPs for classification 507
17.3.8 Connections with deep learning 508
17.4 Scaling GPs to large datasets 508
17.4.1 (Sparse) variational inference 509
17.4.2 Exploiting structure in the kernel matrix 513

17.4.3 Random feature approximation 514
17.5 Support vector machines (SVMs) 516
17.5.1 Large margin classifiers 516
17.5.2 The dual problem 518
17.5.3 Soft margin classifiers 520
17.5.4 The kernel trick 521
17.5.5 Converting SVM outputs into probabilitues 522
17.5.6 Connection with logistic regression 522
17.5.7 Multi-class classification with SVMs 523
17.5.8 How to choose the regularizer C 524
17.5.9 Kernel ridge regression 525
17.5.10 SVMs for regression 526
17.6 Sparse vector machines 528
17.6.1 Relevance vector machines (RVMs) 529
17.6.2 Comparison of sparse and dense kernel methods 529

18 Trees, forests, bagging and boosting 533

18.1 Classification and regression trees (CART) 533
18.1.1 Model definition 533
18.1.2 Model fitting 534
18.1.3 Regularization 535
18.1.4 Pros and cons 535
18.2 Ensemble learning 537
18.2.1 Stacking 537
18.2.2 Ensembling is not Bayes model averaging 537
18.3 Bagging 538
18.4 Random forests 539
18.5 Boosting 540
18.5.1 Forward stagewise additive modeling 541
18.5.2 Quadratic loss and least squares boosting 541
18.5.3 Exponential loss and AdaBoost 541
18.5.4 LogitBoost 544
18.5.5 Gradient boosting 545
18.6 Interpreting tree ensembles 549
18.6.1 Feature importance 549
18.6.2 Partial dependency plots 550

V Beyond supervised learning 553

19 Learning with fewer labeled examples 555

19.1 Data augmentation 555
19.1.1 Examples 555
19.1.2 Theoretical justification 556
19.2 Transfer learning 557

19.2.1 Fine-tuning 557
19.2.2 Supervised pre-training 558
19.2.3 Unsupervised pre-training (self-supervised learning) 559
19.2.4 Domain adaptation 562
19.3 Meta-learning 562
19.3.1 Model-agnostic meta-learning (MAML) 563
19.4 Few-shot learning 564
19.4.1 Matching networks 565
19.5 Word embeddings 566
19.5.1 Methods based on SVD 566
19.5.2 Word2vec 568
19.5.3 RAND-WALK model of word embeddings 570
19.5.4 Word analogies 571
19.5.5 Contextual word embeddings 572
19.6 Semi-supervised learning 576
19.6.1 Self-training and pseudo-labeling 577
19.6.2 Entropy minimization 578
19.6.3 Co-training 580
19.6.4 Label propagation on graphs 581
19.6.5 Consistency regularization 582
19.6.6 Deep generative models 583
19.6.7 Combining self-supervised and semi-supervised learning 586
19.7 Active learning 587
19.7.1 Decision-theoretic approach 587
19.7.2 Information-theoretic approach 588
19.7.3 Batch active learning 588

20 Dimensionality reduction 591

20.1 Principal components analysis (PCA) 591
20.1.1 Examples 591
20.1.2 Derivation of the algorithm 593
20.1.3 Computational issues 596
20.1.4 Choosing the number of latent dimensions 598
20.2 Factor analysis 600
20.2.1 Generative model 600
20.2.2 Probabilistic PCA 602
20.2.3 EM algorithm for FA/PPCA 603
20.2.4 Unidentifiability of the parameters 605
20.2.5 Nonlinear factor analysis 607
20.2.6 Mixtures of factor analysers 608
20.2.7 Exponential family factor analysis 609
20.2.8 Factor analysis models for paired data 610
20.3 Autoencoders 613
20.3.1 Bottleneck autoencoders 614
20.3.2 Denoising autoencoders 615

20.3.3 Contractive autoencoders 615
20.3.4 Sparse autoencoders 617
20.3.5 Variational autoencoders 618
20.4 Manifold learning 623
20.4.1 What are manifolds? 623
20.4.2 The manifold hypothesis 623
20.4.3 Approaches to manifold learning 624
20.4.4 Multi-dimensional scaling (MDS) 625
20.4.5 Isomap 628
20.4.6 Kernel PCA 628
20.4.7 Maximum variance unfolding (MVU) 630
20.4.8 Local linear embedding (LLE) 631
20.4.9 Laplacian eigenmaps 632
20.4.10 t-SNE 635

21 Clustering 639
21.1 Introduction 639
21.1.1 Evaluating the output of clustering methods 639
21.2 Hierarchical agglomerative clustering 641
21.2.1 The algorithm 642
21.2.2 Example 644
21.3 K means clustering 645
21.3.1 The algorithm 646
21.3.2 Examples 646
21.3.3 Vector quantization 647
21.3.4 The K-means++ algorithm 649
21.3.5 The K-medoids algorithm 649
21.3.6 Speedup tricks 650
21.3.7 Choosing the number of clusters K 650
21.4 Clustering using mixture models 653
21.4.1 Mixtures of Gaussians 654
21.5 Spectral clustering 654
21.5.1 Normalized cuts 654
21.5.2 Eigenvectors of the graph Laplacian encode the clustering 656
21.5.3 Example 656
21.5.4 Connection with other methods 657
21.6 Biclustering 658
21.6.1 Basic biclustering 658
21.6.2 Nested partition models (Crosscat) 659

22 Recommender systems 663
22.1 Explicit feedback 663
22.1.1 Datasets 663
22.1.2 Collaborative filtering 664
22.1.3 Matrix factorization 665

22.1.4 Autoencoders 667
22.2 Implicit feedback 669
22.2.1 Bayesian personalized ranking 669
22.2.2 Factorization machines 670
22.2.3 Neural matrix factorization 671
22.3 Leveraging side information 671
22.4 Exploration-exploitation tradeoff 672

23 Graph embeddings 675
23.1 Introduction 675
23.2 Graph Embedding as an Encoder/Decoder Problem 676
23.3 Shallow graph embeddings 678
23.3.1 Unsupervised embeddings 678
23.3.2 Distance-based: Euclidean methods 679
23.3.3 Distance-based: non-Euclidean methods 680
23.3.4 Outer product-based: Matrix factorization methods 680
23.3.5 Outer product-based: Skip-gram methods 681
23.3.6 Supervised embeddings 682
23.4 Graph Neural Networks 683
23.4.1 Message passing GNNs 683
23.4.2 Spectral Graph Convolutions 685
23.4.3 Spatial Graph Convolutions 685
23.4.4 Non-Euclidean Graph Convolutions 687
23.5 Deep graph embeddings 687
23.5.1 Unsupervised embeddings 687
23.5.2 Semi-supervised embeddings 690
23.6 Applications 691
23.6.1 Unsupervised applications 691
23.6.2 Supervised applications 693

Appendices 695

VI Appendix: Mathematical backgound 697

A Some useful mathematics 699
A.1 Introduction 699
A.2 Sets, functions and relations 699
A.2.1 Functions 699
A.2.2 Relations 703
A.3 Matrix calculus 704
A.3.1 Derivatives 704
A.3.2 Gradients 705
A.3.3 Jacobian 706
A.3.4 Hessian 706
A.3.5 Gradients of commonly used functions 706

A.4 Convexity 708
A.4.1 Convex sets 708
A.4.2 Convex functions 710
A.4.3 Jensen's inequality 712
A.4.4 Subgradients 712
A.4.5 Taylor series approximation 713
A.4.6 Bregman divergence 714
A.4.7 Conjugate duality 715

B Linear algebra 719
B.1 Introduction 719
B.1.1 Notation 719
B.1.2 Vector spaces 722
B.1.3 Norms of a vector and matrix 724
B.1.4 Properties of a matrix 726
B.1.5 Special types of matrices 728
B.2 Matrix multiplication 732
B.2.1 Vector-Vector Products 732
B.2.2 Matrix-Vector Products 732
B.2.3 Matrix-Matrix Products 733
B.2.4 Application: manipulating data matrices 735
B.2.5 Kronecker products 737
B.2.6 Einstein summation 738
B.3 Matrix inversion 739
B.3.1 The inverse of a square matrix 739
B.3.2 Schur complements 739
B.3.3 The matrix inversion lemma 741
B.3.4 Matrix determinant lemma 741
B.4 Eigenvalue decomposition (EVD) 742
B.4.1 Basics 742
B.4.2 Diagonalization 743
B.4.3 Eigenvalues and eigenvectors of symmetric matrices 743
B.4.4 Geometry of quadratic forms 744
B.4.5 Standardizing and whitening data 744
B.4.6 Power method 746
B.4.7 Deflation 747
B.4.8 Eigenvectors optimize quadratic forms 747
B.5 Singular value decomposition (SVD) 747
B.5.1 Basics 747
B.5.2 Connection between SVD and EVD 748
B.5.3 Pseudo inverse 749
B.5.4 SVD and the range and null space of a matrix 750
B.5.5 Truncated SVD 751
B.6 Other matrix decompositions 752
B.6.1 LU factorization 752

B.6.2 QR decomposition 752
B.6.3 Cholesky decomposition 753
B.7 Solving systems of linear equations 754
B.7.1 Solving square systems 755
B.7.2 Solving underconstrained systems (least norm estimation) 755
B.7.3 Solving overconstrained systems (least squares estimation) 756

C Probability 759

C.1 Introduction 759
C.1.1 What is probability? 759
C.1.2 Types of uncertainty 759
C.1.3 Fundamental rules of probability 760
C.2 Random variables 761
C.2.1 Discrete random variables 761
C.2.2 Continuous random variables 762
C.3 Sets of related random variables 764
C.3.1 Joint, marginal and conditional distributions 764
C.3.2 Bayes' rule 765
C.3.3 Independence and conditional independence 765
C.4 Properties of a distribution 766
C.4.1 Moments of a distribution 766
C.4.2 Covariance 769
C.4.3 Correlation 769
C.4.4 Uncorrelated does not imply indpendent 770
C.4.5 Correlation does not imply causation 771
C.4.6 Simpsons' paradox 771
C.5 Transformations of random variables 772
C.5.1 Discrete case 773
C.5.2 Continuous case 773
C.5.3 Invertible transformations (bijectors) 773
C.5.4 Moments of a linear transformation 775
C.5.5 The convolution theorem 776
C.5.6 Central limit theorem 778

D Frequentist statistics 779

D.1 Introduction 779
D.2 Fisher information matrix (FIM) 779
D.2.1 Definition 779
D.2.2 Connection between the FIM and the Hessian of the NLL 780
D.2.3 Examples 781
D.2.4 Connection between FIM and KL divergence 782
D.3 Sampling distributions 783
D.3.1 Exact sampling distribution of the MLE 783
D.3.2 Large sample approximation 785
D.3.3 Bootstrap approximation 786

D.3.4 Confidence intervals 788

D.4 Bias and variance 789

D.4.1 Bias of an estimator 789

D.4.2 Variance of an estimator 789

D.4.3 The bias-variance tradeoff 790

D.4.4 Jacknife 792

D.5 Frequentist decision theory 794

D.5.1 Computing the risk of an estimator 794

D.5.2 Consistent estimators 797

D.5.3 Admissible estimators 797

D.5.4 Stein's paradox 798

D.6 Empirical risk minimization 800

D.6.1 Empirical risk 800

D.6.2 Structural risk 802

D.6.3 Cross-validation 802

D.6.4 Statistical learning theory 803

D.7 Hypothesis testing 804

D.7.1 Likelihood ratio test 805

D.7.2 Null hypothesis significance testing (NHST) 806

D.7.3 t-test 806

D.7.4 χ^2 test 807

D.7.5 p-values 808

D.8 Pathologies of frequentist statistics 808

D.8.1 Confidence intervals are not credible 809

D.8.2 p-values confuse deduction with induction 810

D.8.3 p-values overstate evidence against the null hypothesis 810

D.8.4 p-values depend on the stopping rule 811

D.8.5 Why isn't everyone a Bayesian? 812

E Exercises 815

Bibliography 851