Deep Learning

Ian Goodfellow
Yoshua Bengio
Aaron Courville

Contents

Website vii

Acknowledgments viii

Notation xi

1 Introduction 1

1.1 Who Should Read This Book? 8

1.2 Historical Trends in Deep Learning 11

I Applied Math and Machine Learning Basics 29

2 Linear Algebra 31

2.1 Scalars, Vectors, Matrices and Tensors 31

2.2 Multiplying Matrices and Vectors 34

2.3 Identity and Inverse Matrices 36

2.4 Linear Dependence and Span 37

2.5 Norms 39

2.6 Special Kinds of Matrices and Vectors 40

2.7 Eigendecomposition 42

2.8 Singular Value Decomposition 44

2.9 The Moore-Penrose Pseudoinverse 45

2.10 The Trace Operator 46

2.11 The Determinant 47

2.12 Example: Principal Components Analysis 48

3 Probability and Information Theory 53

3.1 Why Probability? 54

3.2 Random Variables . . . 56
3.3 Probability Distributions . . . 56
3.4 Marginal Probability . . . 58
3.5 Conditional Probability . . . 59
3.6 The Chain Rule of Conditional Probabilities . . . 59
3.7 Independence and Conditional Independence . . . 60
3.8 Expectation, Variance and Covariance . . . 60
3.9 Common Probability Distributions . . . 62
3.10 Useful Properties of Common Functions . . . 67
3.11 Bayes' Rule . . . 70
3.12 Technical Details of Continuous Variables . . . 71
3.13 Information Theory . . . 72
3.14 Structured Probabilistic Models . . . 75

4 Numerical Computation **80**
4.1 Overflow and Underflow . . . 80
4.2 Poor Conditioning . . . 82
4.3 Gradient-Based Optimization . . . 82
4.4 Constrained Optimization . . . 93
4.5 Example: Linear Least Squares . . . 96

5 Machine Learning Basics **98**
5.1 Learning Algorithms . . . 99
5.2 Capacity, Overfitting and Underfitting . . . 110
5.3 Hyperparameters and Validation Sets . . . 120
5.4 Estimators, Bias and Variance . . . 122
5.5 Maximum Likelihood Estimation . . . 131
5.6 Bayesian Statistics . . . 135
5.7 Supervised Learning Algorithms . . . 139
5.8 Unsupervised Learning Algorithms . . . 145
5.9 Stochastic Gradient Descent . . . 150
5.10 Building a Machine Learning Algorithm . . . 152
5.11 Challenges Motivating Deep Learning . . . 154

II Deep Networks: Modern Practices **165**

6 Deep Feedforward Networks **167**
6.1 Example: Learning XOR . . . 170
6.2 Gradient-Based Learning . . . 176

6.3 Hidden Units . . . 190
6.4 Architecture Design . . . 196
6.5 Back-Propagation and Other Differentiation Algorithms . . . 203
6.6 Historical Notes . . . 224

7 Regularization for Deep Learning 228
7.1 Parameter Norm Penalties . . . 230
7.2 Norm Penalties as Constrained Optimization . . . 237
7.3 Regularization and Under-Constrained Problems . . . 239
7.4 Dataset Augmentation . . . 240
7.5 Noise Robustness . . . 242
7.6 Semi-Supervised Learning . . . 244
7.7 Multi-Task Learning . . . 245
7.8 Early Stopping . . . 246
7.9 Parameter Tying and Parameter Sharing . . . 251
7.10 Sparse Representations . . . 253
7.11 Bagging and Other Ensemble Methods . . . 255
7.12 Dropout . . . 257
7.13 Adversarial Training . . . 267
7.14 Tangent Distance, Tangent Prop, and Manifold Tangent Classifier 268

8 Optimization for Training Deep Models 274
8.1 How Learning Differs from Pure Optimization . . . 275
8.2 Challenges in Neural Network Optimization . . . 282
8.3 Basic Algorithms . . . 294
8.4 Parameter Initialization Strategies . . . 301
8.5 Algorithms with Adaptive Learning Rates . . . 306
8.6 Approximate Second-Order Methods . . . 310
8.7 Optimization Strategies and Meta-Algorithms . . . 318

9 Convolutional Networks 331
9.1 The Convolution Operation . . . 332
9.2 Motivation . . . 336
9.3 Pooling . . . 340
9.4 Convolution and Pooling as an Infinitely Strong Prior . . . 346
9.5 Variants of the Basic Convolution Function . . . 348
9.6 Structured Outputs . . . 359
9.7 Data Types . . . 361
9.8 Efficient Convolution Algorithms . . . 363
9.9 Random or Unsupervised Features . . . 364

9.10 The Neuroscientific Basis for Convolutional Networks 365
9.11 Convolutional Networks and the History of Deep Learning 372

10 Sequence Modeling: Recurrent and Recursive Nets 374
10.1 Unfolding Computational Graphs 376
10.2 Recurrent Neural Networks 379
10.3 Bidirectional RNNs 396
10.4 Encoder-Decoder Sequence-to-Sequence Architectures 397
10.5 Deep Recurrent Networks 399
10.6 Recursive Neural Networks 401
10.7 The Challenge of Long-Term Dependencies 403
10.8 Echo State Networks 406
10.9 Leaky Units and Other Strategies for Multiple Time Scales 409
10.10 The Long Short-Term Memory and Other Gated RNNs 411
10.11 Optimization for Long-Term Dependencies 415
10.12 Explicit Memory 419

11 Practical methodology 424
11.1 Performance Metrics 425
11.2 Default Baseline Models 428
11.3 Determining Whether to Gather More Data 429
11.4 Selecting Hyperparameters 430
11.5 Debugging Strategies 439
11.6 Example: Multi-Digit Number Recognition 443

12 Applications 446
12.1 Large Scale Deep Learning 446
12.2 Computer Vision 455
12.3 Speech Recognition 461
12.4 Natural Language Processing 464
12.5 Other Applications 480

III Deep Learning Research 489

13 Linear Factor Models 492
13.1 Probabilistic PCA and Factor Analysis 493
13.2 Independent Component Analysis (ICA) 494
13.3 Slow Feature Analysis 496
13.4 Sparse Coding 499

13.5 Manifold Interpretation of PCA 502

14 Autoencoders 505
14.1 Undercomplete Autoencoders 506
14.2 Regularized Autoencoders 507
14.3 Representational Power, Layer Size and Depth 511
14.4 Stochastic Encoders and Decoders 512
14.5 Denoising Autoencoders 513
14.6 Learning Manifolds with Autoencoders 518
14.7 Contractive Autoencoders 524
14.8 Predictive Sparse Decomposition 526
14.9 Applications of Autoencoders 527

15 Representation Learning 529
15.1 Greedy Layer-Wise Unsupervised Pretraining 531
15.2 Transfer Learning and Domain Adaptation 539
15.3 Semi-Supervised Disentangling of Causal Factors 544
15.4 Distributed Representation 549
15.5 Exponential Gains from Depth 556
15.6 Providing Clues to Discover Underlying Causes 557

16 Structured Probabilistic Models for Deep Learning 561
16.1 The Challenge of Unstructured Modeling 562
16.2 Using Graphs to Describe Model Structure 566
16.3 Sampling from Graphical Models 583
16.4 Advantages of Structured Modeling 584
16.5 Learning about Dependencies 585
16.6 Inference and Approximate Inference 586
16.7 The Deep Learning Approach to Structured Probabilistic Models 587

17 Monte Carlo Methods 593
17.1 Sampling and Monte Carlo Methods 593
17.2 Importance Sampling 595
17.3 Markov Chain Monte Carlo Methods 598
17.4 Gibbs Sampling 602
17.5 The Challenge of Mixing between Separated Modes 602

18 Confronting the Partition Function 608
18.1 The Log-Likelihood Gradient 609
18.2 Stochastic Maximum Likelihood and Contrastive Divergence . . . 610

18.3 Pseudolikelihood 618
18.4 Score Matching and Ratio Matching 620
18.5 Denoising Score Matching 622
18.6 Noise-Contrastive Estimation 623
18.7 Estimating the Partition Function 626

19 Approximate inference 634
19.1 Inference as Optimization 636
19.2 Expectation Maximization 637
19.3 MAP Inference and Sparse Coding 638
19.4 Variational Inference and Learning 641
19.5 Learned Approximate Inference 653

20 Deep Generative Models 656
20.1 Boltzmann Machines 656
20.2 Restricted Boltzmann Machines 658
20.3 Deep Belief Networks 662
20.4 Deep Boltzmann Machines 665
20.5 Boltzmann Machines for Real-Valued Data 678
20.6 Convolutional Boltzmann Machines 685
20.7 Boltzmann Machines for Structured or Sequential Outputs 687
20.8 Other Boltzmann Machines 688
20.9 Back-Propagation through Random Operations 689
20.10 Directed Generative Nets 694
20.11 Drawing Samples from Autoencoders 712
20.12 Generative Stochastic Networks 716
20.13 Other Generation Schemes 717
20.14 Evaluating Generative Models 719
20.15 Conclusion 721

Bibliography 723

Index 780

Website

www.deeplearningbook.org

This book is accompanied by the above website. The website provides a variety of supplementary material, including exercises, lecture slides, corrections of mistakes, and other resources that should be useful to both readers and instructors.