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 73
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 140
5.8 Unsupervised Learning Algorithms 146
5.9 Stochastic Gradient Descent 151
5.10 Building a Machine Learning Algorithm 153
5.11 Challenges Motivating Deep Learning 155

II Deep Networks: Modern Practices 166

6 Deep Feedforward Networks 168
6.1 Example: Learning XOR 171
6.2 Gradient-Based Learning 177

6.3 Hidden Units 191
6.4 Architecture Design 197
6.5 Back-Propagation and Other Differentiation Algorithms 204
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 243
7.7 Multi-Task Learning 244
7.8 Early Stopping 246
7.9 Parameter Tying and Parameter Sharing 253
7.10 Sparse Representations 254
7.11 Bagging and Other Ensemble Methods 256
7.12 Dropout 258
7.13 Adversarial Training 268
7.14 Tangent Distance, Tangent Prop, and Manifold Tangent Classifier 270

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 317

9 Convolutional Networks 330
9.1 The Convolution Operation 331
9.2 Motivation 335
9.3 Pooling 339
9.4 Convolution and Pooling as an Infinitely Strong Prior 345
9.5 Variants of the Basic Convolution Function 347
9.6 Structured Outputs 358
9.7 Data Types 360
9.8 Efficient Convolution Algorithms 362
9.9 Random or Unsupervised Features 363

9.10 The Neuroscientific Basis for Convolutional Networks 364
9.11 Convolutional Networks and the History of Deep Learning 371

10 Sequence Modeling: Recurrent and Recursive Nets 373
10.1 Unfolding Computational Graphs 375
10.2 Recurrent Neural Networks 378
10.3 Bidirectional RNNs 394
10.4 Encoder-Decoder Sequence-to-Sequence Architectures 396
10.5 Deep Recurrent Networks 398
10.6 Recursive Neural Networks 400
10.7 The Challenge of Long-Term Dependencies 401
10.8 Echo State Networks 404
10.9 Leaky Units and Other Strategies for Multiple Time Scales 406
10.10 The Long Short-Term Memory and Other Gated RNNs 408
10.11 Optimization for Long-Term Dependencies 413
10.12 Explicit Memory 416

11 Practical Methodology 421
11.1 Performance Metrics 422
11.2 Default Baseline Models 425
11.3 Determining Whether to Gather More Data 426
11.4 Selecting Hyperparameters 427
11.5 Debugging Strategies 436
11.6 Example: Multi-Digit Number Recognition 440

12 Applications 443
12.1 Large-Scale Deep Learning 443
12.2 Computer Vision 452
12.3 Speech Recognition 458
12.4 Natural Language Processing 461
12.5 Other Applications 478

III Deep Learning Research 486

13 Linear Factor Models 489
13.1 Probabilistic PCA and Factor Analysis 490
13.2 Independent Component Analysis (ICA) 491
13.3 Slow Feature Analysis 493
13.4 Sparse Coding 496

13.5 Manifold Interpretation of PCA 499

14 Autoencoders **502**
14.1 Undercomplete Autoencoders 503
14.2 Regularized Autoencoders 504
14.3 Representational Power, Layer Size and Depth 508
14.4 Stochastic Encoders and Decoders 509
14.5 Denoising Autoencoders 510
14.6 Learning Manifolds with Autoencoders 515
14.7 Contractive Autoencoders 521
14.8 Predictive Sparse Decomposition 523
14.9 Applications of Autoencoders 524

15 Representation Learning **526**
15.1 Greedy Layer-Wise Unsupervised Pretraining 528
15.2 Transfer Learning and Domain Adaptation 536
15.3 Semi-Supervised Disentangling of Causal Factors 541
15.4 Distributed Representation 546
15.5 Exponential Gains from Depth 553
15.6 Providing Clues to Discover Underlying Causes 554

16 Structured Probabilistic Models for Deep Learning **558**
16.1 The Challenge of Unstructured Modeling 559
16.2 Using Graphs to Describe Model Structure 563
16.3 Sampling from Graphical Models 580
16.4 Advantages of Structured Modeling 582
16.5 Learning about Dependencies 582
16.6 Inference and Approximate Inference 584
16.7 The Deep Learning Approach to Structured Probabilistic Models 585

17 Monte Carlo Methods **590**
17.1 Sampling and Monte Carlo Methods 590
17.2 Importance Sampling 592
17.3 Markov Chain Monte Carlo Methods 595
17.4 Gibbs Sampling 599
17.5 The Challenge of Mixing between Separated Modes 599

18 Confronting the Partition Function **605**
18.1 The Log-Likelihood Gradient 606
18.2 Stochastic Maximum Likelihood and Contrastive Divergence . . . 607

18.3 Pseudolikelihood 615
18.4 Score Matching and Ratio Matching 617
18.5 Denoising Score Matching 619
18.6 Noise-Contrastive Estimation 620
18.7 Estimating the Partition Function 623

19 Approximate Inference **631**
19.1 Inference as Optimization 633
19.2 Expectation Maximization 634
19.3 MAP Inference and Sparse Coding 635
19.4 Variational Inference and Learning 638
19.5 Learned Approximate Inference 651

20 Deep Generative Models **654**
20.1 Boltzmann Machines 654
20.2 Restricted Boltzmann Machines 656
20.3 Deep Belief Networks 660
20.4 Deep Boltzmann Machines 663
20.5 Boltzmann Machines for Real-Valued Data 676
20.6 Convolutional Boltzmann Machines 683
20.7 Boltzmann Machines for Structured or Sequential Outputs 685
20.8 Other Boltzmann Machines 686
20.9 Back-Propagation through Random Operations 687
20.10 Directed Generative Nets 692
20.11 Drawing Samples from Autoencoders 711
20.12 Generative Stochastic Networks 714
20.13 Other Generation Schemes 716
20.14 Evaluating Generative Models 717
20.15 Conclusion 720

Bibliography **721**

Index **777**

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.