Deep Learning

Ian Goodfellow
Yoshua Bengio
Aaron Courville

Contents

Website viii

Acknowledgments ix

Notation xiii

1 Introduction 1
- 1.1 Who Should Read This Book? . . . 8
- 1.2 Historical Trends in Deep Learning . . . 12

I Applied Math and Machine Learning Basics 27

2 Linear Algebra 29
- 2.1 Scalars, Vectors, Matrices and Tensors . . . 29
- 2.2 Multiplying Matrices and Vectors . . . 32
- 2.3 Identity and Inverse Matrices . . . 34
- 2.4 Linear Dependence and Span . . . 35
- 2.5 Norms . . . 37
- 2.6 Special Kinds of Matrices and Vectors . . . 38
- 2.7 Eigendecomposition . . . 40
- 2.8 Singular Value Decomposition . . . 42
- 2.9 The Moore-Penrose Pseudoinverse . . . 43
- 2.10 The Trace Operator . . . 44
- 2.11 The Determinant . . . 45
- 2.12 Example: Principal Components Analysis . . . 45

3 Probability and Information Theory 51
- 3.1 Why Probability? . . . 52

3.2 Random Variables . . . 54
3.3 Probability Distributions . . . 54
3.4 Marginal Probability . . . 56
3.5 Conditional Probability . . . 57
3.6 The Chain Rule of Conditional Probabilities . . . 57
3.7 Independence and Conditional Independence . . . 58
3.8 Expectation, Variance and Covariance . . . 58
3.9 Common Probability Distributions . . . 60
3.10 Useful Properties of Common Functions . . . 65
3.11 Bayes' Rule . . . 68
3.12 Technical Details of Continuous Variables . . . 69
3.13 Information Theory . . . 71
3.14 Structured Probabilistic Models . . . 73

4 Numerical Computation 78
4.1 Overflow and Underflow . . . 78
4.2 Poor Conditioning . . . 80
4.3 Gradient-Based Optimization . . . 80
4.4 Constrained Optimization . . . 91
4.5 Example: Linear Least Squares . . . 94

5 Machine Learning Basics 96
5.1 Learning Algorithms . . . 97
5.2 Capacity, Overfitting and Underfitting . . . 108
5.3 Hyperparameters and Validation Sets . . . 118
5.4 Estimators, Bias and Variance . . . 120
5.5 Maximum Likelihood Estimation . . . 129
5.6 Bayesian Statistics . . . 133
5.7 Supervised Learning Algorithms . . . 137
5.8 Unsupervised Learning Algorithms . . . 142
5.9 Stochastic Gradient Descent . . . 149
5.10 Building a Machine Learning Algorithm . . . 151
5.11 Challenges Motivating Deep Learning . . . 152

II Deep Networks: Modern Practices 162

6 Deep Feedforward Networks 164
6.1 Example: Learning XOR . . . 167
6.2 Gradient-Based Learning . . . 172

6.3 Hidden Units . . . 187
6.4 Architecture Design . . . 193
6.5 Back-Propagation and Other Differentiation Algorithms . . . 200
6.6 Historical Notes . . . 220

7 Regularization for Deep Learning 224
7.1 Parameter Norm Penalties . . . 226
7.2 Norm Penalties as Constrained Optimization . . . 233
7.3 Regularization and Under-Constrained Problems . . . 235
7.4 Dataset Augmentation . . . 236
7.5 Noise Robustness . . . 238
7.6 Semi-Supervised Learning . . . 240
7.7 Multitask Learning . . . 241
7.8 Early Stopping . . . 241
7.9 Parameter Tying and Parameter Sharing . . . 249
7.10 Sparse Representations . . . 251
7.11 Bagging and Other Ensemble Methods . . . 253
7.12 Dropout . . . 255
7.13 Adversarial Training . . . 265
7.14 Tangent Distance, Tangent Prop and Manifold Tangent Classifier . . . 267

8 Optimization for Training Deep Models 271
8.1 How Learning Differs from Pure Optimization . . . 272
8.2 Challenges in Neural Network Optimization . . . 279
8.3 Basic Algorithms . . . 290
8.4 Parameter Initialization Strategies . . . 296
8.5 Algorithms with Adaptive Learning Rates . . . 302
8.6 Approximate Second-Order Methods . . . 307
8.7 Optimization Strategies and Meta-Algorithms . . . 313

9 Convolutional Networks 326
9.1 The Convolution Operation . . . 327
9.2 Motivation . . . 329
9.3 Pooling . . . 335
9.4 Convolution and Pooling as an Infinitely Strong Prior . . . 339
9.5 Variants of the Basic Convolution Function . . . 342
9.6 Structured Outputs . . . 352
9.7 Data Types . . . 354

9.8 Efficient Convolution Algorithms . . . 356
9.9 Random or Unsupervised Features . . . 356
9.10 The Neuroscientific Basis for Convolutional Networks . . . 358
9.11 Convolutional Networks and the History of Deep Learning . . . 365

10 Sequence Modeling: Recurrent and Recursive Nets 367
10.1 Unfolding Computational Graphs . . . 369
10.2 Recurrent Neural Networks . . . 372
10.3 Bidirectional RNNs . . . 388
10.4 Encoder-Decoder Sequence-to-Sequence Architectures . . . 390
10.5 Deep Recurrent Networks . . . 392
10.6 Recursive Neural Networks . . . 394
10.7 The Challenge of Long-Term Dependencies . . . 396
10.8 Echo State Networks . . . 399
10.9 Leaky Units and Other Strategies for Multiple Time Scales . . . 402
10.10 The Long Short-Term Memory and Other Gated RNNs . . . 404
10.11 Optimization for Long-Term Dependencies . . . 408
10.12 Explicit Memory . . . 412

11 Practical Methodology 416
11.1 Performance Metrics . . . 417
11.2 Default Baseline Models . . . 420
11.3 Determining Whether to Gather More Data . . . 421
11.4 Selecting Hyperparameters . . . 422
11.5 Debugging Strategies . . . 431
11.6 Example: Multi-Digit Number Recognition . . . 435

12 Applications 438
12.1 Large-Scale Deep Learning . . . 438
12.2 Computer Vision . . . 447
12.3 Speech Recognition . . . 453
12.4 Natural Language Processing . . . 456
12.5 Other Applications . . . 473

III Deep Learning Research **482**

13 Linear Factor Models **485**
13.1 Probabilistic PCA and Factor Analysis . . . 486
13.2 Independent Component Analysis (ICA) . . . 487
13.3 Slow Feature Analysis . . . 489
13.4 Sparse Coding . . . 492
13.5 Manifold Interpretation of PCA . . . 496

14 Autoencoders **499**
14.1 Undercomplete Autoencoders . . . 500
14.2 Regularized Autoencoders . . . 501
14.3 Representational Power, Layer Size and Depth . . . 505
14.4 Stochastic Encoders and Decoders . . . 506
14.5 Denoising Autoencoders . . . 507
14.6 Learning Manifolds with Autoencoders . . . 513
14.7 Contractive Autoencoders . . . 518
14.8 Predictive Sparse Decomposition . . . 521
14.9 Applications of Autoencoders . . . 522

15 Representation Learning **524**
15.1 Greedy Layer-Wise Unsupervised Pretraining . . . 526
15.2 Transfer Learning and Domain Adaptation . . . 534
15.3 Semi-Supervised Disentangling of Causal Factors . . . 539
15.4 Distributed Representation . . . 544
15.5 Exponential Gains from Depth . . . 550
15.6 Providing Clues to Discover Underlying Causes . . . 552

16 Structured Probabilistic Models for Deep Learning **555**
16.1 The Challenge of Unstructured Modeling . . . 556
16.2 Using Graphs to Describe Model Structure . . . 560
16.3 Sampling from Graphical Models . . . 577
16.4 Advantages of Structured Modeling . . . 579
16.5 Learning about Dependencies . . . 579
16.6 Inference and Approximate Inference . . . 580
16.7 The Deep Learning Approach to Structured Probabilistic Models . . . 581

17 Monte Carlo Methods **587**
17.1 Sampling and Monte Carlo Methods . . . 587

17.2 Importance Sampling . . . 589
17.3 Markov Chain Monte Carlo Methods . . . 592
17.4 Gibbs Sampling . . . 596
17.5 The Challenge of Mixing between Separated Modes . . . 597

18 Confronting the Partition Function **603**
18.1 The Log-Likelihood Gradient . . . 604
18.2 Stochastic Maximum Likelihood and Contrastive Divergence . . . 605
18.3 Pseudolikelihood . . . 613
18.4 Score Matching and Ratio Matching . . . 615
18.5 Denoising Score Matching . . . 617
18.6 Noise-Contrastive Estimation . . . 618
18.7 Estimating the Partition Function . . . 621

19 Approximate Inference **629**
19.1 Inference as Optimization . . . 631
19.2 Expectation Maximization . . . 632
19.3 MAP Inference and Sparse Coding . . . 633
19.4 Variational Inference and Learning . . . 636
19.5 Learned Approximate Inference . . . 648

20 Deep Generative Models **651**
20.1 Boltzmann Machines . . . 651
20.2 Restricted Boltzmann Machines . . . 653
20.3 Deep Belief Networks . . . 657
20.4 Deep Boltzmann Machines . . . 660
20.5 Boltzmann Machines for Real-Valued Data . . . 673
20.6 Convolutional Boltzmann Machines . . . 679
20.7 Boltzmann Machines for Structured or Sequential Outputs . . . 681
20.8 Other Boltzmann Machines . . . 683
20.9 Back-Propagation through Random Operations . . . 684
20.10 Directed Generative Nets . . . 688
20.11 Drawing Samples from Autoencoders . . . 707
20.12 Generative Stochastic Networks . . . 710
20.13 Other Generation Schemes . . . 712
20.14 Evaluating Generative Models . . . 713
20.15 Conclusion . . . 716

Bibliography **717**

Index **773**

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.