

Contents

Website	xiii
Acknowledgments	xv
Notation	xix
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	36
2.6 Special Kinds of Matrices and Vectors	38
2.7 Eigendecomposition	39
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	68
3.13	Information Theory	70
3.14	Structured Probabilistic Models	74
4	Numerical Computation	77
4.1	Overflow and Underflow	77
4.2	Poor Conditioning	79
4.3	Gradient-Based Optimization	79
4.4	Constrained Optimization	89
4.5	Example: Linear Least Squares	92
5	Machine Learning Basics	95
5.1	Learning Algorithms	96
5.2	Capacity, Overfitting and Underfitting	107
5.3	Hyperparameters and Validation Sets	117
5.4	Estimators, Bias and Variance	119
5.5	Maximum Likelihood Estimation	128
5.6	Bayesian Statistics	132
5.7	Supervised Learning Algorithms	136
5.8	Unsupervised Learning Algorithms	142

5.9	Stochastic Gradient Descent	147
5.10	Building a Machine Learning Algorithm	149
5.11	Challenges Motivating Deep Learning	151
II	Deep Networks: Modern Practices	161
6	Deep Feedforward Networks	163
6.1	Example: Learning XOR	166
6.2	Gradient-Based Learning	171
6.3	Hidden Units	185
6.4	Architecture Design	191
6.5	Back-Propagation and Other Differentiation Algorithms	197
6.6	Historical Notes	217
7	Regularization for Deep Learning	221
7.1	Parameter Norm Penalties	223
7.2	Norm Penalties as Constrained Optimization	230
7.3	Regularization and Under-Constrained Problems	232
7.4	Dataset Augmentation	233
7.5	Noise Robustness	235
7.6	Semi-Supervised Learning	236
7.7	Multitask Learning	237
7.8	Early Stopping	239
7.9	Parameter Tying and Parameter Sharing	246
7.10	Sparse Representations	247
7.11	Bagging and Other Ensemble Methods	249
7.12	Dropout	251
7.13	Adversarial Training	261
7.14	Tangent Distance, Tangent Prop and Manifold Tangent Classifier .	263
8	Optimization for Training Deep Models	267
8.1	How Learning Differs from Pure Optimization	268

8.2	Challenges in Neural Network Optimization	275
8.3	Basic Algorithms	286
8.4	Parameter Initialization Strategies	292
8.5	Algorithms with Adaptive Learning Rates	298
8.6	Approximate Second-Order Methods	302
8.7	Optimization Strategies and Meta-Algorithms	309
9	Convolutional Networks	321
9.1	The Convolution Operation	322
9.2	Motivation	324
9.3	Pooling	330
9.4	Convolution and Pooling as an Infinitely Strong Prior	334
9.5	Variants of the Basic Convolution Function	337
9.6	Structured Outputs	347
9.7	Data Types	348
9.8	Efficient Convolution Algorithms	350
9.9	Random or Unsupervised Features	351
9.10	The Neuroscientific Basis for Convolutional Networks	353
9.11	Convolutional Networks and the History of Deep Learning	359
10	Sequence Modeling: Recurrent and Recursive Nets	363
10.1	Unfolding Computational Graphs	365
10.2	Recurrent Neural Networks	368
10.3	Bidirectional RNNs	383
10.4	Encoder-Decoder Sequence-to-Sequence Architectures	385
10.5	Deep Recurrent Networks	387
10.6	Recursive Neural Networks	388
10.7	The Challenge of Long-Term Dependencies	390
10.8	Echo State Networks	392
10.9	Leaky Units and Other Strategies for Multiple Time Scales	395
10.10	The Long Short-Term Memory and Other Gated RNNs	397

10.11 Optimization for Long-Term Dependencies	401
10.12 Explicit Memory	405
11 Practical Methodology	409
11.1 Performance Metrics	410
11.2 Default Baseline Models	413
11.3 Determining Whether to Gather More Data	414
11.4 Selecting Hyperparameters	415
11.5 Debugging Strategies	424
11.6 Example: Multi-Digit Number Recognition	428
12 Applications	431
12.1 Large-Scale Deep Learning	431
12.2 Computer Vision	440
12.3 Speech Recognition	446
12.4 Natural Language Processing	448
12.5 Other Applications	465
III Deep Learning Research	475
13 Linear Factor Models	479
13.1 Probabilistic PCA and Factor Analysis	480
13.2 Independent Component Analysis (ICA)	481
13.3 Slow Feature Analysis	484
13.4 Sparse Coding	486
13.5 Manifold Interpretation of PCA	489
14 Autoencoders	493
14.1 Undercomplete Autoencoders	494
14.2 Regularized Autoencoders	495
14.3 Representational Power, Layer Size and Depth	499
14.4 Stochastic Encoders and Decoders	500

14.5	Denoising Autoencoders	501
14.6	Learning Manifolds with Autoencoders	506
14.7	Contractive Autoencoders	510
14.8	Predictive Sparse Decomposition	514
14.9	Applications of Autoencoders	515
15	Representation Learning	517
15.1	Greedy Layer-Wise Unsupervised Pretraining	519
15.2	Transfer Learning and Domain Adaptation	526
15.3	Semi-Supervised Disentangling of Causal Factors	532
15.4	Distributed Representation	536
15.5	Exponential Gains from Depth	543
15.6	Providing Clues to Discover Underlying Causes	544
16	Structured Probabilistic Models for Deep Learning	549
16.1	The Challenge of Unstructured Modeling	550
16.2	Using Graphs to Describe Model Structure	554
16.3	Sampling from Graphical Models	570
16.4	Advantages of Structured Modeling	572
16.5	Learning about Dependencies	572
16.6	Inference and Approximate Inference	573
16.7	The Deep Learning Approach to Structured Probabilistic Models .	575
17	Monte Carlo Methods	581
17.1	Sampling and Monte Carlo Methods	581
17.2	Importance Sampling	583
17.3	Markov Chain Monte Carlo Methods	586
17.4	Gibbs Sampling	590
17.5	The Challenge of Mixing between Separated Modes	591
18	Confronting the Partition Function	597
18.1	The Log-Likelihood Gradient	598
18.2	Stochastic Maximum Likelihood and Contrastive Divergence . . .	599

18.3	Pseudolikelihood	607
18.4	Score Matching and Ratio Matching	609
18.5	Denoising Score Matching	611
18.6	Noise-Contrastive Estimation	612
18.7	Estimating the Partition Function	614
19	Approximate Inference	623
19.1	Inference as Optimization	624
19.2	Expectation Maximization	626
19.3	MAP Inference and Sparse Coding	627
19.4	Variational Inference and Learning	629
19.5	Learned Approximate Inference	642
20	Deep Generative Models	645
20.1	Boltzmann Machines	645
20.2	Restricted Boltzmann Machines	647
20.3	Deep Belief Networks	651
20.4	Deep Boltzmann Machines	654
20.5	Boltzmann Machines for Real-Valued Data	667
20.6	Convolutional Boltzmann Machines	673
20.7	Boltzmann Machines for Structured or Sequential Outputs	675
20.8	Other Boltzmann Machines	677
20.9	Back-Propagation through Random Operations	678
20.10	Directed Generative Nets	682
20.11	Drawing Samples from Autoencoders	701
20.12	Generative Stochastic Networks	704
20.13	Other Generation Schemes	706
20.14	Evaluating Generative Models	707
20.15	Conclusion	710
	Bibliography	711
	Index	767