Deep Learning

Ian Goodfellow Yoshua Bengio Aaron Courville

Contents

Website			viii	
A	cknow	ledgments	ix	
N	otatio	n	xiii	
1	Intro 1.1 1.2	Oduction Who Should Read This Book?		
Ι	Appl	ied Math and Machine Learning Basics	27	
2	Linear Algebra			
	2.1	Scalars, Vectors, Matrices and Tensors	29	
	2.2	Multiplying Matrices and Vectors		
	2.3	Identity and Inverse Matrices		
	2.4	Linear Dependence and Span		
	2.5	Norms		
	2.6	Special Kinds of Matrices and Vectors		
	2.7	Eigendecomposition		
	2.8	Singular Value Decomposition		
	2.9	The Moore-Penrose Pseudoinverse		
	2.10	The Trace Operator		
	2.11	The Determinant		
	2.12	Example: Principal Components Analysis		
3	Probability and Information Theory			
	3.1	Why Probability?	52	

		Gradient-Based Learning	179
6	Deep 6.1	Feedforward Networks Example: Learning XOR	1 64 . 167
II		p Networks: Modern Practices	162
	0.11	Chancing to thou vacing Deep Dearling	. 102
	5.11	Challenges Motivating Deep Learning	
	5.10	Building a Machine Learning Algorithm	
	5.9	Stochastic Gradient Descent	
	5.8	Unsupervised Learning Algorithms	
	5.0 5.7	Supervised Learning Algorithms	
	5.6	Bayesian Statistics	
	5.4 5.5	Estimators, Bias and Variance	
	5.3 5.4	Hyperparameters and Validation Sets	
	5.2 5.3	Capacity, Overfitting and Underfitting	
	5.1	Learning Algorithms	
5		hine Learning Basics	96
_	7N /F 1		
	4.5	Example: Linear Least Squares	
	4.4	Constrained Optimization	
	4.3	Gradient-Based Optimization	
	4.2	Poor Conditioning	
	4.1	Overflow and Underflow	
4	Num	nerical Computation	78
	3.14	Structured Probabilistic Models	. 73
	3.13	Information Theory	
	3.12	Technical Details of Continuous Variables	
	3.11	Bayes' Rule	
	3.10	Useful Properties of Common Functions	
	3.9	Common Probability Distributions	. 60
	3.8	Expectation, Variance and Covariance	. 58
	3.7	Independence and Conditional Independence	. 58
	3.6	The Chain Rule of Conditional Probabilities	. 57
	3.5	Conditional Probability	
	3.4	Marginal Probability	
	3.3	Probability Distributions	
	3.2	Random Variables	. 54

	6.3	Hidden Units			
	6.4	Architecture Design			
	6.5	Back-Propagation and Other Differentiation			
		Algorithms			
	6.6	Historical Notes			
7	Regi	ularization for Deep Learning 224			
	7.1	Parameter Norm Penalties			
	7.2	Norm Penalties as Constrained Optimization			
	7.3	Regularization and Under-Constrained Problems			
	7.4	Dataset Augmentation			
	7.5	Noise Robustness			
	7.6	Semi-Supervised Learning			
	7.7	Multitask Learning			
	7.8	Early Stopping			
	7.9	Parameter Tying and Parameter Sharing			
	7.10	Sparse Representations			
	7.11	Bagging and Other Ensemble Methods			
	7.12	<u>Dropout</u>			
	7.13	Adversarial Training			
	7.14	Tangent Distance, Tangent Prop and Manifold			
		Tangent Classifier			
8	Optimization for Training Deep Models 271				
	8.1	How Learning Differs from Pure Optimization			
	8.2	Challenges in Neural Network Optimization			
	8.3	Basic Algorithms			
	8.4	Parameter Initialization Strategies			
	8.5	Algorithms with Adaptive Learning Rates			
	8.6	Approximate Second-Order Methods			
	8.7	Optimization Strategies and Meta-Algorithms			
9	Convolutional Networks 326				
	9.1	The Convolution Operation			
	9.2	Motivation			
	9.3	Pooling			
	9.4	Convolution and Pooling as an Infinitely Strong Prior 339			
	9.5	Variants of the Basic Convolution Function			
	9.6	Structured Outputs			
	0.7	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	Sequ	ence Modeling: Recurrent and Recursive Nets	367
	10.1	Unfolding Computational Graphs	369
	10.2	Recurrent Neural Networks	
	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
		Optimization for Long-Term Dependencies	
		Explicit Memory	
11	Practical Methodology		416
	11.1	Performance Metrics	417
	11.2	Default Baseline Models	
	11.3	Determining Whether to Gather More Data	421
	11.4	Selecting Hyperparameters	
	11.5	Debugging Strategies	
	11.6	Example: Multi-Digit Number Recognition	
f 12	Applications 438		
	12.1	Large-Scale Deep Learning	438
	12.2	Computer Vision	
	12.3	Speech Recognition	
	12.4	Natural Language Processing	
	12.5	Other Applications	

Ш	De	ep Learning Research	482
13	Line	ar Factor Models	485
	13.1	Probabilistic PCA and Factor Analysis	486
	13.2	Independent Component Analysis (ICA)	487
	13.3	Slow Feature Analysis	
	13.4	Sparse Coding	492
	13.5	Manifold Interpretation of PCA	
14	Auto	pencoders	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	
	14.7	Contractive Autoencoders	518
	14.8	Predictive Sparse Decomposition	521
	14.9	Applications of Autoencoders	522
15	Rep	resentation 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	Stru	ctured 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	Mon	te Carlo Methods	587
	17.1	Sampling and Monte Carlo Methods	587

	17.2 17.3 17.4 17.5	Importance Sampling	592 596
18	Confi	ronting 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	Appr	roximate 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	
	20.5	Boltzmann Machines for Real-Valued Data	673
	20.6	Convolutional Boltzmann Machines	679
	20.7	Boltzmann Machines for Structured or Sequential Outputs	
	20.8	Other Boltzmann Machines	
	20.9	Back-Propagation through Random Operations	
		Directed Generative Nets	
		Drawing Samples from Autoencoders	
		Generative Stochastic Networks	
		Other Generation Schemes	
		Evaluating Generative Models	
	20.15	Conclusion	716
Bil	oliogra	aphy	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.