

Contents

Preface	v
Contents	xi
1 The Deep Learning Revolution	1
1.1 The Impact of Deep Learning	2
1.1.1 Medical diagnosis	2
1.1.2 Protein structure	3
1.1.3 Image synthesis	4
1.1.4 Large language models	5
1.2 A Tutorial Example	6
1.2.1 Synthetic data	6
1.2.2 Linear models	8
1.2.3 Error function	8
1.2.4 Model complexity	9
1.2.5 Regularization	12
1.2.6 Model selection	14
1.3 A Brief History of Machine Learning	16
1.3.1 Single-layer networks	17
1.3.2 Backpropagation	18
1.3.3 Deep networks	20
2 Probabilities	23
2.1 The Rules of Probability	25
2.1.1 A medical screening example	25
2.1.2 The sum and product rules	26
2.1.3 Bayes' theorem	28
2.1.4 Medical screening revisited	30
2.1.5 Prior and posterior probabilities	31

2.1.6	Independent variables	31
2.2	Probability Densities	32
2.2.1	Example distributions	33
2.2.2	Expectations and covariances	34
2.3	The Gaussian Distribution	36
2.3.1	Mean and variance	37
2.3.2	Likelihood function	37
2.3.3	Bias of maximum likelihood	39
2.3.4	Linear regression	40
2.4	Transformation of Densities	42
2.4.1	Multivariate distributions	44
2.5	Information Theory	46
2.5.1	Entropy	46
2.5.2	Physics perspective	47
2.5.3	Differential entropy	49
2.5.4	Maximum entropy	50
2.5.5	Kullback–Leibler divergence	51
2.5.6	Conditional entropy	53
2.5.7	Mutual information	54
2.6	Bayesian Probabilities	54
2.6.1	Model parameters	55
2.6.2	Regularization	56
2.6.3	Bayesian machine learning	57
	Exercises	58
3	Standard Distributions	65
3.1	Discrete Variables	66
3.1.1	Bernoulli distribution	66
3.1.2	Binomial distribution	67
3.1.3	Multinomial distribution	68
3.2	The Multivariate Gaussian	70
3.2.1	Geometry of the Gaussian	71
3.2.2	Moments	74
3.2.3	Limitations	75
3.2.4	Conditional distribution	76
3.2.5	Marginal distribution	79
3.2.6	Bayes' theorem	81
3.2.7	Maximum likelihood	84
3.2.8	Sequential estimation	85
3.2.9	Mixtures of Gaussians	86
3.3	Periodic Variables	89
3.3.1	Von Mises distribution	89
3.4	The Exponential Family	94
3.4.1	Sufficient statistics	97
3.5	Nonparametric Methods	98

3.5.1	Histograms	98
3.5.2	Kernel densities	100
3.5.3	Nearest-neighbours	103
	Exercises	105
4	Single-layer Networks: Regression	111
4.1	Linear Regression	112
4.1.1	Basis functions	112
4.1.2	Likelihood function	114
4.1.3	Maximum likelihood	115
4.1.4	Geometry of least squares	117
4.1.5	Sequential learning	117
4.1.6	Regularized least squares	118
4.1.7	Multiple outputs	119
4.2	Decision theory	120
4.3	The Bias–Variance Trade-off	123
	Exercises	128
5	Single-layer Networks: Classification	131
5.1	Discriminant Functions	132
5.1.1	Two classes	132
5.1.2	Multiple classes	134
5.1.3	1-of- K coding	135
5.1.4	Least squares for classification	136
5.2	Decision Theory	138
5.2.1	Misclassification rate	139
5.2.2	Expected loss	140
5.2.3	The reject option	142
5.2.4	Inference and decision	143
5.2.5	Classifier accuracy	147
5.2.6	ROC curve	148
5.3	Generative Classifiers	150
5.3.1	Continuous inputs	152
5.3.2	Maximum likelihood solution	153
5.3.3	Discrete features	156
5.3.4	Exponential family	156
5.4	Discriminative Classifiers	157
5.4.1	Activation functions	158
5.4.2	Fixed basis functions	158
5.4.3	Logistic regression	159
5.4.4	Multi-class logistic regression	161
5.4.5	Probit regression	163
5.4.6	Canonical link functions	164
	Exercises	166

6	Deep Neural Networks	171
6.1	Limitations of Fixed Basis Functions	172
6.1.1	The curse of dimensionality	172
6.1.2	High-dimensional spaces	175
6.1.3	Data manifolds	176
6.1.4	Data-dependent basis functions	178
6.2	Multilayer Networks	180
6.2.1	Parameter matrices	181
6.2.2	Universal approximation	181
6.2.3	Hidden unit activation functions	182
6.2.4	Weight-space symmetries	185
6.3	Deep Networks	186
6.3.1	Hierarchical representations	187
6.3.2	Distributed representations	187
6.3.3	Representation learning	188
6.3.4	Transfer learning	189
6.3.5	Contrastive learning	191
6.3.6	General network architectures	193
6.3.7	Tensors	194
6.4	Error Functions	194
6.4.1	Regression	194
6.4.2	Binary classification	196
6.4.3	multiclass classification	197
6.5	Mixture Density Networks	198
6.5.1	Robot kinematics example	198
6.5.2	Conditional mixture distributions	199
6.5.3	Gradient optimization	201
6.5.4	Predictive distribution	202
	Exercises	204
7	Gradient Descent	209
7.1	Error Surfaces	210
7.1.1	Local quadratic approximation	211
7.2	Gradient Descent Optimization	213
7.2.1	Use of gradient information	214
7.2.2	Batch gradient descent	214
7.2.3	Stochastic gradient descent	214
7.2.4	Mini-batches	216
7.2.5	Parameter initialization	216
7.3	Convergence	218
7.3.1	Momentum	220
7.3.2	Learning rate schedule	222
7.3.3	RMSProp and Adam	223
7.4	Normalization	224
7.4.1	Data normalization	226

7.4.2	Batch normalization	227
7.4.3	Layer normalization	229
	Exercises	230
8	Backpropagation	233
8.1	Evaluation of Gradients	234
8.1.1	Single-layer networks	234
8.1.2	General feed-forward networks	235
8.1.3	A simple example	238
8.1.4	Numerical differentiation	239
8.1.5	The Jacobian matrix	240
8.1.6	The Hessian matrix	242
8.2	Automatic Differentiation	244
8.2.1	Forward-mode automatic differentiation	246
8.2.2	Reverse-mode automatic differentiation	249
	Exercises	250
9	Regularization	253
9.1	Inductive Bias	254
9.1.1	Inverse problems	254
9.1.2	No free lunch theorem	255
9.1.3	Symmetry and invariance	256
9.1.4	Equivariance	259
9.2	Weight Decay	260
9.2.1	Consistent regularizers	262
9.2.2	Generalized weight decay	264
9.3	Learning Curves	266
9.3.1	Early stopping	266
9.3.2	Double descent	268
9.4	Parameter Sharing	270
9.4.1	Soft weight sharing	271
9.5	Residual Connections	274
9.6	Model Averaging	277
9.6.1	Dropout	279
	Exercises	281
10	Convolutional Networks	287
10.1	Computer Vision	288
10.1.1	Image data	289
10.2	Convolutional Filters	290
10.2.1	Feature detectors	290
10.2.2	Translation equivariance	291
10.2.3	Padding	294
10.2.4	Strided convolutions	294
10.2.5	Multi-dimensional convolutions	295
10.2.6	Pooling	296

10.2.7	Multilayer convolutions	298
10.2.8	Example network architectures	299
10.3	Visualizing Trained CNNs	302
10.3.1	Visual cortex	302
10.3.2	Visualizing trained filters	303
10.3.3	Saliency maps	305
10.3.4	Adversarial attacks	306
10.3.5	Synthetic images	308
10.4	Object Detection	308
10.4.1	Bounding boxes	309
10.4.2	Intersection-over-union	310
10.4.3	Sliding windows	311
10.4.4	Detection across scales	313
10.4.5	Non-max suppression	314
10.4.6	Fast region CNNs	314
10.5	Image Segmentation	315
10.5.1	Convolutional segmentation	315
10.5.2	Up-sampling	316
10.5.3	Fully convolutional networks	318
10.5.4	The U-net architecture	319
10.6	Style Transfer	320
	Exercises	322
11	Structured Distributions	325
11.1	Graphical Models	326
11.1.1	Directed graphs	326
11.1.2	Factorization	327
11.1.3	Discrete variables	329
11.1.4	Gaussian variables	332
11.1.5	Binary classifier	334
11.1.6	Parameters and observations	334
11.1.7	Bayes' theorem	336
11.2	Conditional Independence	337
11.2.1	Three example graphs	338
11.2.2	Explaining away	341
11.2.3	D-separation	343
11.2.4	Naive Bayes	344
11.2.5	Generative models	346
11.2.6	Markov blanket	347
11.2.7	Graphs as filters	348
11.3	Sequence Models	349
11.3.1	Hidden variables	352
	Exercises	353

12	Transformers	357
12.1	Attention	358
12.1.1	Transformer processing	360
12.1.2	Attention coefficients	361
12.1.3	Self-attention	362
12.1.4	Network parameters	363
12.1.5	Scaled self-attention	366
12.1.6	Multi-head attention	366
12.1.7	Transformer layers	368
12.1.8	Computational complexity	370
12.1.9	Positional encoding	371
12.2	Natural Language	374
12.2.1	Word embedding	375
12.2.2	Tokenization	377
12.2.3	Bag of words	378
12.2.4	Autoregressive models	379
12.2.5	Recurrent neural networks	380
12.2.6	Backpropagation through time	381
12.3	Transformer Language Models	382
12.3.1	Decoder transformers	383
12.3.2	Sampling strategies	386
12.3.3	Encoder transformers	388
12.3.4	Sequence-to-sequence transformers	390
12.3.5	Large language models	390
12.4	Multimodal Transformers	394
12.4.1	Vision transformers	395
12.4.2	Generative image transformers	396
12.4.3	Audio data	399
12.4.4	Text-to-speech	400
12.4.5	Vision and language transformers	402
	Exercises	403
13	Graph Neural Networks	407
13.1	Machine Learning on Graphs	409
13.1.1	Graph properties	410
13.1.2	Adjacency matrix	410
13.1.3	Permutation equivariance	411
13.2	Neural Message-Passing	412
13.2.1	Convolutional filters	413
13.2.2	Graph convolutional networks	414
13.2.3	Aggregation operators	416
13.2.4	Update operators	418
13.2.5	Node classification	419
13.2.6	Edge classification	420
13.2.7	Graph classification	420

13.3	General Graph Networks	420
13.3.1	Graph attention networks	421
13.3.2	Edge embeddings	421
13.3.3	Graph embeddings	422
13.3.4	Over-smoothing	422
13.3.5	Regularization	423
13.3.6	Geometric deep learning	424
	Exercises	425
14	Sampling	429
14.1	Basic Sampling Algorithms	430
14.1.1	Expectations	430
14.1.2	Standard distributions	431
14.1.3	Rejection sampling	433
14.1.4	Adaptive rejection sampling	435
14.1.5	Importance sampling	437
14.1.6	Sampling-importance-resampling	439
14.2	Markov Chain Monte Carlo	440
14.2.1	The Metropolis algorithm	441
14.2.2	Markov chains	442
14.2.3	The Metropolis–Hastings algorithm	445
14.2.4	Gibbs sampling	446
14.2.5	Ancestral sampling	450
14.3	Langevin Sampling	451
14.3.1	Energy-based models	452
14.3.2	Maximizing the likelihood	453
14.3.3	Langevin dynamics	454
	Exercises	456
15	Discrete Latent Variables	459
15.1	K -means Clustering	460
15.1.1	Image segmentation	464
15.2	Mixtures of Gaussians	466
15.2.1	Likelihood function	468
15.2.2	Maximum likelihood	470
15.3	Expectation–Maximization Algorithm	474
15.3.1	Gaussian mixtures	478
15.3.2	Relation to K -means	480
15.3.3	Mixtures of Bernoulli distributions	481
15.4	Evidence Lower Bound	485
15.4.1	EM revisited	486
15.4.2	Independent and identically distributed data	488
15.4.3	Parameter priors	489
15.4.4	Generalized EM	489
15.4.5	Sequential EM	490
	Exercises	490

16	Continuous Latent Variables	495
16.1	Principal Component Analysis	497
16.1.1	Maximum variance formulation	497
16.1.2	Minimum-error formulation	499
16.1.3	Data compression	501
16.1.4	Data whitening	502
16.1.5	High-dimensional data	504
16.2	Probabilistic Latent Variables	506
16.2.1	Generative model	506
16.2.2	Likelihood function	507
16.2.3	Maximum likelihood	509
16.2.4	Factor analysis	513
16.2.5	Independent component analysis	514
16.2.6	Kalman filters	515
16.3	Evidence Lower Bound	516
16.3.1	Expectation maximization	518
16.3.2	EM for PCA	519
16.3.3	EM for factor analysis	520
16.4	Nonlinear Latent Variable Models	522
16.4.1	Nonlinear manifolds	522
16.4.2	Likelihood function	524
16.4.3	Discrete data	526
16.4.4	Four approaches to generative modelling	527
	Exercises	527
17	Generative Adversarial Networks	533
17.1	Adversarial Training	534
17.1.1	Loss function	535
17.1.2	GAN training in practice	536
17.2	Image GANs	539
17.2.1	CycleGAN	539
	Exercises	544
18	Normalizing Flows	547
18.1	Coupling Flows	549
18.2	Autoregressive Flows	552
18.3	Continuous Flows	554
18.3.1	Neural differential equations	554
18.3.2	Neural ODE backpropagation	555
18.3.3	Neural ODE flows	557
	Exercises	559

19 Autoencoders	563
19.1 Deterministic Autoencoders	564
19.1.1 Linear autoencoders	564
19.1.2 Deep autoencoders	565
19.1.3 Sparse autoencoders	566
19.1.4 Denoising autoencoders	567
19.1.5 Masked autoencoders	567
19.2 Variational Autoencoders	569
19.2.1 Amortized inference	572
19.2.2 The reparameterization trick	574
Exercises	578
20 Diffusion Models	581
20.1 Forward Encoder	582
20.1.1 Diffusion kernel	583
20.1.2 Conditional distribution	584
20.2 Reverse Decoder	585
20.2.1 Training the decoder	587
20.2.2 Evidence lower bound	588
20.2.3 Rewriting the ELBO	589
20.2.4 Predicting the noise	591
20.2.5 Generating new samples	592
20.3 Score Matching	594
20.3.1 Score loss function	595
20.3.2 Modified score loss	596
20.3.3 Noise variance	597
20.3.4 Stochastic differential equations	598
20.4 Guided Diffusion	599
20.4.1 Classifier guidance	600
20.4.2 Classifier-free guidance	600
Exercises	603
Appendix A Linear Algebra	609
A.1 Matrix Identities	609
A.2 Traces and Determinants	610
A.3 Matrix Derivatives	611
A.4 Eigenvectors	612
Appendix B Calculus of Variations	617
Appendix C Lagrange Multipliers	621
Bibliography	625
Index	641