

Contents

5	<i>List of illustrations</i>	vi
6	<i>List of tables</i>	x
7	<i>Foreword</i>	1
8	Part I Mathematical Foundations	9
9	1 Introduction and Motivation	11
10	1.1 Finding Words for Intuitions	11
11	1.2 Two Ways to Read this Book	12
12	1.3 Exercises and Feedback	15
13	2 Linear Algebra	17
14	2.1 Systems of Linear Equations	19
15	2.2 Matrices	21
16	2.2.1 Matrix Addition and Multiplication	22
17	2.2.2 Inverse and Transpose	24
18	2.2.3 Multiplication by a Scalar	25
19	2.2.4 Compact Representations of Systems of Linear Equations	26
20	2.3 Solving Systems of Linear Equations	26
21	2.3.1 Particular and General Solution	26
22	2.3.2 Elementary Transformations	28
23	2.3.3 The Minus-1 Trick	32
24	2.3.4 Algorithms for Solving a System of Linear Equations	34
25	2.4 Vector Spaces	35
26	2.4.1 Groups	35
27	2.4.2 Vector Spaces	36
28	2.4.3 Vector Subspaces	38
29	2.5 Linear Independence	39
30	2.6 Basis and Rank	43
31	2.6.1 Generating Set and Basis	43
32	2.6.2 Rank	46
33	2.7 Linear Mappings	47
34	2.7.1 Matrix Representation of Linear Mappings	49
35	2.7.2 Basis Change	51
36	2.7.3 Image and Kernel	56
37	2.8 Affine Spaces	58
38	2.8.1 Affine Subspaces	59

39	2.8.2 Affine Mappings	60
40	Exercises	61
41	3 Analytic Geometry	68
42	3.1 Norms	69
43	3.2 Inner Products	70
44	3.2.1 Dot Product	70
45	3.2.2 General Inner Products	70
46	3.2.3 Symmetric, Positive Definite Matrices	71
47	3.3 Lengths and Distances	73
48	3.4 Angles and Orthogonality	74
49	3.5 Orthonormal Basis	76
50	3.6 Inner Product of Functions	77
51	3.7 Orthogonal Projections	78
52	3.7.1 Projection onto 1-Dimensional Subspaces (Lines)	79
53	3.7.2 Projection onto General Subspaces	82
54	3.7.3 Projection onto Affine Subspaces	85
55	3.8 Rotations	86
56	3.8.1 Rotations in \mathbb{R}^2	87
57	3.8.2 Rotations in \mathbb{R}^3	88
58	3.8.3 Rotations in n Dimensions	89
59	3.8.4 Properties of Rotations	89
60	3.9 Further Reading	90
61	Exercises	90
62	4 Matrix Decompositions	92
63	4.1 Determinant and Trace of a Square Matrix	93
64	4.2 Eigenvalues and Eigenvectors	100
65	4.3 Cholesky Decomposition	108
66	4.4 Eigendecomposition and Diagonalization	110
67	4.5 Singular Value Decomposition	115
68	4.5.1 Geometric Intuitions for the SVD	116
69	4.5.2 Existence and Construction of the SVD	119
70	4.5.3 Eigenvalue Decomposition vs Singular Value Decomposition	123
71	4.6 Matrix Approximation	126
72	4.7 Matrix Phylogeny	132
73	4.8 Further Reading	133
74	Exercises	134
75	5 Vector Calculus	137
76	5.1 Differentiation of Univariate Functions	138
77	5.1.1 Taylor Series	140
78	5.1.2 Differentiation Rules	142
79	5.2 Partial Differentiation and Gradients	143
80	5.2.1 Basic Rules of Partial Differentiation	144
81	5.2.2 Chain Rule	145
82	5.3 Gradients of Vector-Valued Functions	146
83	5.4 Gradients of Matrices	152
84	5.5 Useful Identities for Computing Gradients	155

85	5.6	Backpropagation and Automatic Differentiation	155
86	5.6.1	Gradients in a Deep Network	156
87	5.6.2	Automatic Differentiation	158
88	5.7	Higher-order Derivatives	161
89	5.8	Linearization and Multivariate Taylor Series	162
90	5.9	Further Reading	166
91		Exercises	167
92	6	Probability and Distributions	169
93	6.1	Construction of a Probability Space	169
94	6.1.1	Philosophical Issues	169
95	6.1.2	Probability and Random Variables	171
96	6.1.3	Statistics	172
97	6.2	Discrete and Continuous Probabilities	173
98	6.2.1	Discrete Probabilities	173
99	6.2.2	Continuous Probabilities	175
100	6.2.3	Contrasting Discrete and Continuous Distributions	176
101	6.3	Sum Rule, Product Rule and Bayes' Theorem	178
102	6.4	Summary Statistics and Independence	180
103	6.4.1	Means and Covariances	181
104	6.4.2	Three Expressions for the Variance	183
105	6.4.3	Statistical Independence	184
106	6.4.4	Sums and Transformations of Random Variables	186
107	6.4.5	Inner Products of Random Variables	186
108	6.5	Change of Variables/Inverse transform	188
109	6.5.1	Distribution Function Technique	189
110	6.5.2	Change of Variables	191
111	6.6	Gaussian Distribution	194
112	6.6.1	Marginals and Conditionals of Gaussians are Gaussians	196
113	6.6.2	Product of Gaussians	198
114	6.6.3	Sums and Linear Transformations	199
115	6.6.4	Sampling from Multivariate Gaussian Distributions	201
116	6.7	Conjugacy and the Exponential Family	202
117	6.7.1	Conjugacy	205
118	6.7.2	Sufficient Statistics	206
119	6.7.3	Exponential Family	207
120	6.8	Further Reading	209
121		Exercises	210
122	7	Continuous Optimization	212
123	7.1	Optimization using Gradient Descent	214
124	7.1.1	Stepsize	216
125	7.1.2	Gradient Descent with Momentum	217
126	7.1.3	Stochastic Gradient Descent	217
127	7.2	Constrained Optimization and Lagrange Multipliers	219
128	7.3	Convex Optimization	222
129	7.3.1	Linear Programming	225
130	7.3.2	Quadratic Programming	227
131	7.3.3	Legendre-Fenchel Transform and Convex Conjugate	228

132	7.4	Further Reading	232
133		Exercises	233
134		Part II Central Machine Learning Problems	235
135	8	When Models meet Data	237
136	8.1	Empirical Risk Minimization	243
137	8.1.1	Hypothesis Class of Functions	244
138	8.1.2	Loss Function for Training	245
139	8.1.3	Regularization to Reduce Overfitting	247
140	8.1.4	Cross Validation to Assess the Generalization Performance	248
141	8.2	Parameter Estimation	250
142	8.2.1	Maximum Likelihood Estimation	250
143	8.2.2	Maximum A Posteriori Estimation	252
144	8.3	Probabilistic Modeling	255
145	8.3.1	MLE, MAP, and Bayesian Inference	255
146	8.3.2	Latent Variables	256
147	8.4	Directed Graphical Models	258
148	8.4.1	Graph Semantics	259
149	8.4.2	Conditional Independence and D-Separation	262
150	8.5	Model Selection	264
151	8.5.1	Nested Cross Validation	264
152	8.5.2	Bayesian Model Selection	265
153	8.5.3	Bayes Factors for Model Comparison	267
154	9	Linear Regression	269
155	9.1	Problem Formulation	271
156	9.2	Parameter Estimation	272
157	9.2.1	Maximum Likelihood Estimation	272
158	9.2.2	Overfitting in Linear Regression	277
159	9.2.3	Regularization and Maximum A Posteriori Estimation	279
160	9.3	Bayesian Linear Regression	282
161	9.3.1	Model	283
162	9.3.2	Prior Predictions	283
163	9.3.3	Posterior Distribution	284
164	9.3.4	Posterior Predictions	288
165	9.3.5	Computing the Marginal Likelihood	290
166	9.4	Maximum Likelihood as Orthogonal Projection	292
167	9.5	Further Reading	294
168	10	Dimensionality Reduction with Principal Component Analysis	297
169	10.1	Problem Setting	298
170	10.2	Maximum Variance Perspective	300
171	10.2.1	Direction with Maximal Variance	301
172	10.2.2	M -dimensional Subspace with Maximal Variance	303
173	10.3	Projection Perspective	305
174	10.3.1	Setting and Objective	305
175	10.3.2	Optimization	307

176	10.4	Eigenvector Computation and Low-Rank Approximations	313
177	10.4.1	PCA using Low-rank Matrix Approximations	314
178	10.4.2	Practical Aspects	314
179	10.5	PCA in High Dimensions	315
180	10.6	Key Steps of PCA in Practice	316
181	10.7	Latent Variable Perspective	320
182	10.7.1	Generative Process and Probabilistic Model	320
183	10.7.2	Likelihood and Joint Distribution	322
184	10.7.3	Posterior Distribution	323
185	10.8	Further Reading	324
186	11	Density Estimation with Gaussian Mixture Models	329
187	11.1	Gaussian Mixture Model	330
188	11.2	Parameter Learning via Maximum Likelihood	331
189	11.3	EM Algorithm	341
190	11.4	Latent Variable Perspective	343
191	11.4.1	Prior	344
192	11.4.2	Marginal	345
193	11.4.3	Posterior	345
194	11.4.4	Extension to a Full Dataset	346
195	11.4.5	EM Algorithm Revisited	346
196	11.5	Further Reading	347
197	12	Classification with Support Vector Machines	349
198	12.1	Separating Hyperplanes	351
199	12.2	Primal Support Vector Machine	353
200	12.2.1	Concept Of The Margin	353
201	12.2.2	Traditional Derivation Of The Margin	355
202	12.2.3	Why We Can Set The Margin To 1	357
203	12.2.4	Soft Margin SVM: Geometric View	358
204	12.2.5	Soft Margin SVM: Loss Function View	359
205	12.3	Dual Support Vector Machine	361
206	12.3.1	Convex Duality Via Lagrange Multipliers	362
207	12.3.2	Soft Margin SVM: Convex Hull View	364
208	12.3.3	Kernels	367
209	12.3.4	Numerical Solution	369
210	12.4	Further Reading	371
211		References	373