
Contents

<i>List of illustrations</i>	vi
<i>List of tables</i>	x
<i>Foreword</i>	1
Part I Mathematical Foundations	11
1 Introduction and Motivation	13
1.1 Finding Words for Intuitions	13
1.2 Two Ways to Read this Book	15
1.3 Exercises and Feedback	18
2 Linear Algebra	19
2.1 Systems of Linear Equations	21
2.2 Matrices	24
2.2.1 Matrix Addition and Multiplication	24
2.2.2 Inverse and Transpose	26
2.2.3 Multiplication by a Scalar	27
2.2.4 Compact Representations of Systems of Linear Equations	28
2.3 Solving Systems of Linear Equations	29
2.3.1 Particular and General Solution	29
2.3.2 Elementary Transformations	30
2.3.3 The Minus-1 Trick	34
2.3.4 Algorithms for Solving a System of Linear Equations	36
2.4 Vector Spaces	37
2.4.1 Groups	38
2.4.2 Vector Spaces	39
2.4.3 Vector Subspaces	41
2.5 Linear Independence	42
2.6 Basis and Rank	46
2.6.1 Generating Set and Basis	46
2.6.2 Rank	49
2.7 Linear Mappings	50
2.7.1 Matrix Representation of Linear Mappings	52
2.7.2 Basis Change	55
2.7.3 Image and Kernel	60
2.8 Affine Spaces	63
2.8.1 Affine Subspaces	63

2.8.2	Affine Mappings	65
2.9	Further Reading	65
	Exercises	65
3	Analytic Geometry	72
3.1	Norms	73
3.2	Inner Products	74
	3.2.1 Dot Product	74
	3.2.2 General Inner Products	74
	3.2.3 Symmetric, Positive Definite Matrices	75
3.3	Lengths and Distances	77
3.4	Angles and Orthogonality	78
3.5	Orthonormal Basis	80
3.6	Inner Product of Functions	81
3.7	Orthogonal Projections	82
	3.7.1 Projection onto 1-Dimensional Subspaces (Lines)	84
	3.7.2 Projection onto General Subspaces	86
	3.7.3 Projection onto Affine Subspaces	89
3.8	Rotations	90
	3.8.1 Rotations in \mathbb{R}^2	91
	3.8.2 Rotations in \mathbb{R}^3	92
	3.8.3 Rotations in n Dimensions	93
	3.8.4 Properties of Rotations	94
3.9	Further Reading	94
	Exercises	95
4	Matrix Decompositions	96
4.1	Determinant and Trace	97
4.2	Eigenvalues and Eigenvectors	104
4.3	Cholesky Decomposition	112
4.4	Eigendecomposition and Diagonalization	114
4.5	Singular Value Decomposition	119
	4.5.1 Geometric Intuitions for the SVD	120
	4.5.2 Existence and Construction of the SVD	123
	4.5.3 Eigenvalue Decomposition vs Singular Value Decomposition	127
4.6	Matrix Approximation	130
4.7	Matrix Phylogeny	135
4.8	Further Reading	136
	Exercises	138
5	Vector Calculus	141
5.1	Differentiation of Univariate Functions	143
	5.1.1 Taylor Series	144
	5.1.2 Differentiation Rules	147
5.2	Partial Differentiation and Gradients	148
	5.2.1 Basic Rules of Partial Differentiation	149
	5.2.2 Chain Rule	150
5.3	Gradients of Vector-Valued Functions	151
5.4	Gradients of Matrices	156

<i>Contents</i>	iii
5.5 Useful Identities for Computing Gradients	160
5.6 Backpropagation and Automatic Differentiation	160
5.6.1 Gradients in a Deep Network	161
5.6.2 Automatic Differentiation	163
5.7 Higher-order Derivatives	166
5.8 Linearization and Multivariate Taylor Series	167
5.9 Further Reading	171
Exercises	172
6 Probability and Distributions	174
6.1 Construction of a Probability Space	174
6.1.1 Philosophical Issues	174
6.1.2 Probability and Random Variables	176
6.1.3 Statistics	179
6.2 Discrete and Continuous Probabilities	180
6.2.1 Discrete Probabilities	180
6.2.2 Continuous Probabilities	182
6.2.3 Contrasting Discrete and Continuous Distributions	183
6.3 Sum Rule, Product Rule and Bayes' Theorem	185
6.4 Summary Statistics and Independence	188
6.4.1 Means and Covariances	188
6.4.2 Empirical Means and Covariances	193
6.4.3 Three Expressions for the Variance	194
6.4.4 Sums and Transformations of Random Variables	195
6.4.5 Statistical Independence	196
6.4.6 Inner Products of Random Variables	197
6.5 Gaussian Distribution	199
6.5.1 Marginals and Conditionals of Gaussians are Gaussians	200
6.5.2 Product of Gaussian Densities	202
6.5.3 Sums and Linear Transformations	203
6.5.4 Sampling from Multivariate Gaussian Distributions	206
6.6 Conjugacy and the Exponential Family	206
6.6.1 Conjugacy	209
6.6.2 Sufficient Statistics	211
6.6.3 Exponential Family	212
6.7 Change of Variables/Inverse Transform	216
6.7.1 Distribution Function Technique	217
6.7.2 Change of Variables	219
6.8 Further Reading	223
Exercises	224
7 Continuous Optimization	227
7.1 Optimization using Gradient Descent	229
7.1.1 Stepsize	231
7.1.2 Gradient Descent with Momentum	232
7.1.3 Stochastic Gradient Descent	233
7.2 Constrained Optimization and Lagrange Multipliers	235
7.3 Convex Optimization	238
7.3.1 Linear Programming	241

7.3.2	Quadratic Programming	243
7.3.3	Legendre-Fenchel Transform and Convex Conjugate	244
7.4	Further Reading	248
	Exercises	249
 Part II Central Machine Learning Problems		 251
8	When Models meet Data	253
8.1	Empirical Risk Minimization	260
8.1.1	Hypothesis Class of Functions	260
8.1.2	Loss Function for Training	261
8.1.3	Regularization to Reduce Overfitting	263
8.1.4	Cross Validation to Assess the Generalization Performance	264
8.2	Parameter Estimation	266
8.2.1	Maximum Likelihood Estimation	266
8.2.2	Maximum A Posteriori Estimation	269
8.3	Probabilistic Modeling and Inference	272
8.3.1	Probabilistic Models	272
8.3.2	Bayesian Inference	273
8.3.3	Latent Variable Models	275
8.4	Directed Graphical Models	278
8.4.1	Graph Semantics	279
8.4.2	Conditional Independence and d-Separation	281
8.5	Model Selection	283
8.5.1	Nested Cross Validation	284
8.5.2	Bayesian Model Selection	285
8.5.3	Bayes Factors for Model Comparison	287
9	Linear Regression	290
9.1	Problem Formulation	292
9.2	Parameter Estimation	293
9.2.1	Maximum Likelihood Estimation	293
9.2.2	Overfitting in Linear Regression	299
9.2.3	Maximum A Posteriori Estimation	301
9.2.4	MAP Estimation as Regularization	303
9.3	Bayesian Linear Regression	304
9.3.1	Model	305
9.3.2	Prior Predictions	305
9.3.3	Posterior Distribution	307
9.3.4	Posterior Predictions	309
9.3.5	Computing the Marginal Likelihood	312
9.4	Maximum Likelihood as Orthogonal Projection	314
9.5	Further Reading	316
10	Dimensionality Reduction with Principal Component Analysis	318
10.1	Problem Setting	319
10.2	Maximum Variance Perspective	321
10.2.1	Direction with Maximal Variance	322

<i>Contents</i>	v
10.2.2 M -dimensional Subspace with Maximal Variance	323
10.3 Projection Perspective	326
10.3.1 Setting and Objective	326
10.3.2 Finding Optimal Coordinates	328
10.3.3 Finding the Basis of the Principal Subspace	330
10.4 Eigenvector Computation and Low-Rank Approximations	334
10.4.1 PCA using Low-rank Matrix Approximations	335
10.4.2 Practical Aspects	335
10.5 PCA in High Dimensions	336
10.6 Key Steps of PCA in Practice	337
10.7 Latent Variable Perspective	341
10.7.1 Generative Process and Probabilistic Model	341
10.7.2 Likelihood and Joint Distribution	343
10.7.3 Posterior Distribution	344
10.8 Further Reading	345
11 Density Estimation with Gaussian Mixture Models	349
11.1 Gaussian Mixture Model	350
11.2 Parameter Learning via Maximum Likelihood	351
11.2.1 Responsibilities	353
11.2.2 Updating the Means	354
11.2.3 Updating the Covariances	357
11.2.4 Updating the Mixture Weights	359
11.3 EM Algorithm	361
11.4 Latent Variable Perspective	363
11.4.1 Generative Process and Probabilistic Model	363
11.4.2 Likelihood	366
11.4.3 Posterior Distribution	367
11.4.4 Extension to a Full Dataset	367
11.4.5 EM Algorithm Revisited	368
11.5 Further Reading	369
12 Classification with Support Vector Machines	371
12.1 Separating Hyperplanes	373
12.2 Primal Support Vector Machine	375
12.2.1 Concept Of The Margin	375
12.2.2 Traditional Derivation Of The Margin	377
12.2.3 Why We Can Set The Margin To 1	379
12.2.4 Soft Margin SVM: Geometric View	380
12.2.5 Soft Margin SVM: Loss Function View	381
12.3 Dual Support Vector Machine	383
12.3.1 Convex Duality Via Lagrange Multipliers	384
12.3.2 Soft Margin SVM: Convex Hull View	386
12.3.3 Kernels	389
12.3.4 Numerical Solution	391
12.4 Further Reading	393
<i>References</i>	395
<i>Index</i>	407