

# Contents

Preface

page xiii

## Part I Mathematical Foundations

### 1 Solving Equations

1.1	Linear Equation Systems	3
1.2	The Bisection and Secant Methods	6
1.2.1	The Bisection Search	6
1.2.2	The Secant Method	8
1.3	Fixed-Point Iteration	9
1.4	Newton–Raphson Method (Univariate)	20
1.5	Newton–Raphson Method (Multivariate)	24
	Problems	30
1.A	Order of Convergence of the Secant Method	31
1.B	Order of Convergence of the Newton–Raphson Method	32
1.C	Proofs of the Fixed-Point and Contraction Mapping Theorems	34

### 2 Unconstrained Optimization

2.1	Optimization by Solving Equations	36
2.2	Newton’s Method	37
2.3	Gradient Descent Method	43
2.4	Line Minimization	46
2.5	Quasi-Newton Methods	54
2.6	Conjugate Gradient Method	60
2.7	Issues of Local/Global Minimum	72
	Problems	73

### 3 Constrained Optimization

3.1	Optimization with Equality Constraints	76
3.2	Optimization with Inequality Constraints	79
3.3	Duality and KKT Conditions	84
3.4	Linear Programming (LP)	87
3.5	The Simplex Algorithm	94
3.6	Quadratic Programming (QP)	99
3.7	Interior Point Methods	101
	Problems	110

## Part II Regression

<b>4</b>	<b>Bias–Variance Tradeoff and Overfitting vs. Underfitting</b>	113
4.1	Supervised Learning as Optimization	113
4.2	Bias–Variance Tradeoff	117
4.3	Cross-Validation	121
4.4	Regularization and Ensemble Learning	122
<b>5</b>	<b>Linear Regression</b>	124
5.1	Linear Least Squares (LLS) Regression	124
5.2	Ridge Regression	135
5.3	Regression Based on Basis Functions	138
5.4	Bayesian Regression	148
	Problems	155
<b>6</b>	<b>Nonlinear Regression</b>	157
6.1	Nonlinear Least Squares Regression	157
6.2	Parameter Estimation by Natural Gradient Descent	164
	Problems	165
<b>7</b>	<b>Logistic and Softmax Regression</b>	166
7.1	Logistic Regression and Binary Classification	166
7.2	Softmax Regression for Multiclass Classification	174
	Problems	181
<b>8</b>	<b>Gaussian Process Regression and Classification</b>	183
8.1	Gaussian Process Regression	183
8.2	Gaussian Process Classifier – Binary	191
8.3	Gaussian Process Classifier – Multiclass	199
	Problems	207

## Part III Feature Extraction

<b>9</b>	<b>Feature Selection</b>	211
9.1	Distances and Separability Measurements	211
	Problems	215
<b>10</b>	<b>Principal Component Analysis</b>	218
10.1	Covariance and Correlation	218
10.2	Karhunen–Loève Transformation	220
10.3	Optimality of the KLT	223
10.4	Geometric Interpretation of the KLT	225
10.5	Computation of the KLT	228
10.6	Comparison with Other Orthogonal Transforms	229

10.7	Application to Image Data	232
10.8	PCA for Feature Extraction Problems	237
		240
<b>11</b>	<b>Variations of PCA</b>	243
11.1	Kernel Methods	243
11.2	Kernel PCA	247
11.3	Factor Analysis and Expectation Maximization	250
11.4	Probabilistic PCA	257
11.5	Classical Multidimensional Scaling	262
11.6	t-Distributed Stochastic Neighbor Embedding Problems	266
		274
<b>12</b>	<b>Independent Component Analysis</b>	276
12.1	Independence and Non-Gaussianity	276
12.2	Preprocessing and Whitening	278
12.3	Non-Gaussianity and FastICA	279
12.4	Likelihood and Independence Maximization Problems	283
		288
<b>Part IV Classification</b>		
<b>13</b>	<b>Statistic Classification</b>	293
13.1	Discriminative vs. Generative Methods for Classification	293
13.2	k-nearest Neighbors and Minimum Distance Classifiers	294
13.3	Naive Bayes Classification	297
13.4	Adaptive Boosting Problems	309
		320
<b>14</b>	<b>Support Vector Machine</b>	321
14.1	Maximum Margin and Support Vectors	321
14.2	Kernel SVM	330
14.3	Soft Margin SVM	331
14.4	Sequential Minimal Optimization Algorithm	333
14.5	Multiclass Classification	343
14.6	Kernelized Bayes Classifier Problems	348
		361
<b>15</b>	<b>Clustering Analysis</b>	363
15.1	k-Means Clustering	363
15.2	Gaussian Mixture Model	370
15.3	Bernoulli Mixture Model Problems	381
		385

<b>16</b>	<b>Hierarchical Classifiers</b>	387
16.1	Bottom-Up vs. Top-Down Methods	387
16.2	Binary Hierarchical Classification	389
16.3	Binary Hierarchical Clustering	394
	Problems	401

## Part V Neural Networks

<b>17</b>	<b>Biologically Inspired Networks</b>	405
17.1	Biological Neural Network Modeling	405
17.2	Hebbian Learning	409
17.3	Hopfield Network	411

<b>18</b>	<b>Perceptron-Based Networks</b>	416
18.1	Perceptron	416
18.2	BackPropagation	425
18.3	Autoencoder	435
18.4	Deep Learning	444
	Problems	449

<b>19</b>	<b>Competition-Based Networks</b>	451
19.1	Competitive Learning Network	451
19.2	Self-Organizing Map (SOM)	458
	Problems	467

## Part VI Reinforcement Learning

<b>20</b>	<b>Introduction to Reinforcement Learning</b>	471
20.1	Markov Decision Process	471
	20.1.1 Markov Chain	471
	20.1.2 Markov Reward Process	472
	20.1.3 Markov Decision Process	475
20.2	Model-Based Planning	478
20.3	Model-Free Evaluation and Control	483
	20.3.1 Monte Carlo (MC) Algorithms	487
	20.3.2 Temporal Difference (TD) Algorithms	490
	20.3.3 TD( $\lambda$ ) Algorithm	497
20.4	Value Function Approximation	502
20.5	Control Based on Function Approximation	507
20.6	Deep Q-Learning	509
20.7	Policy Gradient Methods	512
	Problems	519

## Part VII Large Language Models

<b>21 Large Language Models</b>	<b>523</b>
21.1 Natural Language Processing and Modeling	523
21.2 The Transformer Architecture	525
21.3 Attention Is All You Need	527
21.4 Backpropagation Learning	530
21.5 Transfer Learning in NLP	535
21.6 Reinforcement Learning with Human Feedback	536
21.7 Scaling Laws	537
21.8 Mathematical Operations in the Transformer Architecture	538
21.9 Emergence	539
21.10 Do LLMs Understand?	541
21.11 Future Directions	545
21.12 Applications of Transformers Beyond NLP	547
21.13 Generative Methods Beyond Transformers	550
Index	554

### A Different Pedagogical Philosophy

The idea of converting notes into a textbook came from the realization that my approach to teaching the subject is not quite the same as many of the existing textbooks on the subject, which can be roughly categorized as either theoretical or practical. Books of the theoretical type treat ML as a subject of applied mathematics, emphasizing the mathematical theory as the background of the various ML methods. Such books typically provide detailed mathematical background and rigorous derivations of the algorithms, but they pay little attention to the actual implementation of the algorithms discussed. Readers of these books may be left puzzled if they also want to know how the algorithms are actually implemented in code and applied to real-world problems. Conversely, books of the practical type treat the subject of ML as a computational toolbox, which is black or a gray at best, in the sense that they emphasize mostly the application of the tools in the toolbox while paying little attention to the mathematical background of the algorithms. Such books typically rely on certain off-the-shelf commercial software packages, and they discuss mostly how the functions in some built-in libraries are used, but there is very little on why the underneath algorithms work (or do not work under certain conditions). Some of these books may be in the style of manual books of cooking recipes, providing little more than a set of steps to follow. Although such books may serve the purpose for certain readers, others may feel unsatisfied