

José Unpingco

Python for Probability, Statistics, and Machine Learning

Second Edition

 Springer

Contents

1	Getting Started with Scientific Python	1
1.1	Installation and Setup	2
1.2	Numpy	4
1.2.1	Numpy Arrays and Memory	6
1.2.2	Numpy Matrices	9
1.2.3	Numpy Broadcasting	10
1.2.4	Numpy Masked Arrays	13
1.2.5	Floating-Point Numbers	13
1.2.6	Numpy Optimizations and Prospectus	17
1.3	Matplotlib	17
1.3.1	Alternatives to Matplotlib	19
1.3.2	Extensions to Matplotlib	20
1.4	IPython	20
1.5	Jupyter Notebook	22
1.6	Scipy	24
1.7	Pandas	25
1.7.1	Series	25
1.7.2	Dataframe	27
1.8	Sympy	30
1.9	Interfacing with Compiled Libraries	32
1.10	Integrated Development Environments	33
1.11	Quick Guide to Performance and Parallel Programming	34
1.12	Other Resources	37
	References	38
2	Probability	39
2.1	Introduction	39
2.1.1	Understanding Probability Density	40
2.1.2	Random Variables	41
2.1.3	Continuous Random Variables	46

2.1.4	Transformation of Variables Beyond Calculus	49
2.1.5	Independent Random Variables	51
2.1.6	Classic Broken Rod Example	53
2.2	Projection Methods	55
2.2.1	Weighted Distance	57
2.3	Conditional Expectation as Projection	58
2.3.1	Appendix	64
2.4	Conditional Expectation and Mean Squared Error	65
2.5	Worked Examples of Conditional Expectation and Mean Square Error Optimization	68
2.5.1	Example	69
2.5.2	Example	72
2.5.3	Example	75
2.5.4	Example	78
2.5.5	Example	79
2.5.6	Example	82
2.6	Useful Distributions	83
2.6.1	Normal Distribution	83
2.6.2	Multinomial Distribution	84
2.6.3	Chi-square Distribution	86
2.6.4	Poisson and Exponential Distributions	89
2.6.5	Gamma Distribution	90
2.6.6	Beta Distribution	91
2.6.7	Dirichlet-Multinomial Distribution	93
2.7	Information Entropy	95
2.7.1	Information Theory Concepts	96
2.7.2	Properties of Information Entropy	98
2.7.3	Kullback–Leibler Divergence	99
2.7.4	Cross-Entropy as Maximum Likelihood	100
2.8	Moment Generating Functions	101
2.9	Monte Carlo Sampling Methods	104
2.9.1	Inverse CDF Method for Discrete Variables	105
2.9.2	Inverse CDF Method for Continuous Variables	107
2.9.3	Rejection Method	108
2.10	Sampling Importance Resampling	113
2.11	Useful Inequalities	115
2.11.1	Markov’s Inequality	115
2.11.2	Chebyshev’s Inequality	116
2.11.3	Hoeffding’s Inequality	118
	References	120

- 3 Statistics** 123
 - 3.1 Introduction 123
 - 3.2 Python Modules for Statistics 124
 - 3.2.1 Scipy Statistics Module 124
 - 3.2.2 Sympy Statistics Module 125
 - 3.2.3 Other Python Modules for Statistics 126
 - 3.3 Types of Convergence 126
 - 3.3.1 Almost Sure Convergence 126
 - 3.3.2 Convergence in Probability 129
 - 3.3.3 Convergence in Distribution 131
 - 3.3.4 Limit Theorems 132
 - 3.4 Estimation Using Maximum Likelihood 133
 - 3.4.1 Setting Up the Coin-Flipping Experiment 135
 - 3.4.2 Delta Method 145
 - 3.5 Hypothesis Testing and P-Values 147
 - 3.5.1 Back to the Coin-Flipping Example 149
 - 3.5.2 Receiver Operating Characteristic 152
 - 3.5.3 P-Values 154
 - 3.5.4 Test Statistics 155
 - 3.5.5 Testing Multiple Hypotheses 163
 - 3.5.6 Fisher Exact Test 163
 - 3.6 Confidence Intervals 166
 - 3.7 Linear Regression 169
 - 3.7.1 Extensions to Multiple Covariates 178
 - 3.8 Maximum A-Posteriori 183
 - 3.9 Robust Statistics 188
 - 3.10 Bootstrapping 195
 - 3.10.1 Parametric Bootstrap 200
 - 3.11 Gauss–Markov 201
 - 3.12 Nonparametric Methods 205
 - 3.12.1 Kernel Density Estimation 205
 - 3.12.2 Kernel Smoothing 207
 - 3.12.3 Nonparametric Regression Estimators 213
 - 3.12.4 Nearest Neighbors Regression 214
 - 3.12.5 Kernel Regression 218
 - 3.12.6 Curse of Dimensionality 219
 - 3.12.7 Nonparametric Tests 221
 - 3.13 Survival Analysis 228
 - 3.13.1 Example 231
- References 236

- 4 Machine Learning** 237
 - 4.1 Introduction 237
 - 4.2 Python Machine Learning Modules 237
 - 4.3 Theory of Learning 241
 - 4.3.1 Introduction to Theory of Machine Learning 244
 - 4.3.2 Theory of Generalization 249
 - 4.3.3 Worked Example for Generalization/Approximation Complexity 250
 - 4.3.4 Cross-Validation 256
 - 4.3.5 Bias and Variance 260
 - 4.3.6 Learning Noise 265
 - 4.4 Decision Trees 268
 - 4.4.1 Random Forests 275
 - 4.4.2 Boosting Trees 277
 - 4.5 Boosting Trees 281
 - 4.5.1 Boosting Trees 281
 - 4.6 Logistic Regression 285
 - 4.7 Generalized Linear Models 295
 - 4.8 Regularization 300
 - 4.8.1 Ridge Regression 304
 - 4.8.2 Lasso Regression 309
 - 4.9 Support Vector Machines 311
 - 4.9.1 Kernel Tricks 315
 - 4.10 Dimensionality Reduction 317
 - 4.10.1 Independent Component Analysis 321
 - 4.11 Clustering 325
 - 4.12 Ensemble Methods 329
 - 4.12.1 Bagging 329
 - 4.12.2 Boosting 331
 - 4.13 Deep Learning 334
 - 4.13.1 Introduction to Tensorflow 343
 - 4.13.2 Understanding Gradient Descent 350
 - 4.13.3 Image Processing Using Convolutional Neural Networks 363
 - References 379
- Notation** 381
- Index** 383