

Introduction	1
---------------------------	---

Part I Basics

1 Elements of data organization	11
1.1 Conceptual data models	12
1.1.1 Entity-relationship models	12
1.2 Logical data models	13
1.2.1 Relational data models	14
1.2.2 Graph-based data models	17
1.2.3 Hierarchical data models	19
1.3 Data quality	21
1.3.1 Data quality dimensions	22
1.4 Data cleaning	23
1.4.1 Validation	23
1.4.2 Standardization	24
1.4.3 Imputation	24
1.4.4 Augmentation	26
1.4.5 Deduplication	27
References	33
2 Descriptive statistics	35
2.1 Samples	35
2.2 Statistical charts	37
2.2.1 Bar charts and histograms	38
2.2.2 Scatter plots	39
2.2.3 Pie charts, grouped and stacked bar charts, heatmaps ...	41
2.3 Measures of central tendency	43
2.3.1 Arithmetic mean and sample median	43
2.3.2 Sample quantiles	48
2.3.3 Geometric and harmonic mean	49

2.4	Measures of variation	50
2.4.1	Deviation around the mean or the median	51
2.4.2	Shannon index	53
2.5	Measures of association	54
2.5.1	Sample covariance and Pearson's correlation coefficient ..	55
2.5.2	Rank correlation coefficients	56
2.5.3	Sample mutual information and Jaccard index	58
	References	64

Part II Stochastics

3	Probability theory	67
3.1	Probability measures	68
3.1.1	Conditional probability	72
3.1.2	Bayes' theorem	76
3.2	Random variables	79
3.2.1	Discrete and continuous random variables	79
3.2.2	Probability mass and density functions	82
3.2.3	Transformations of random variables	85
3.3	Joint distribution of random variables	88
3.3.1	Joint probability mass and density functions	88
3.3.2	Conditional probability mass and density functions	90
3.3.3	Independent random variables	91
3.4	Characteristic measures of random variables	92
3.4.1	Median, expected value, and variance	92
3.4.2	Covariance and correlation	98
3.4.3	Chebyshev's inequality	102
3.5	Sums and products of random variables	103
3.5.1	Chi-squared and Student's t -distribution	107
	References	110
4	Inferential statistics	111
4.1	Statistical models	112
4.1.1	Models of discrete random variables	112
4.1.2	Models of continuous random variables	115
4.2	Laws of large numbers	118
4.2.1	Bernoulli's law of large numbers	118
4.2.2	Chebyshev's law of large numbers	123
4.2.3	Variance estimation and Bessel correction	125
4.2.4	Lindeberg–Lévy central limit theorem	126
4.3	Interval estimation and hypothesis testing	129
4.3.1	Interval estimation	129
4.3.2	Z -test	131
4.3.3	Student's t -test	134
4.3.4	Effect size	136
4.4	Parameter and density estimation	137

4.4.1	Maximum likelihood estimation	139
4.4.2	Bayesian parameter estimation	145
4.4.3	Kernel density estimation	149
4.5	Regression analysis	151
4.5.1	Simple linear regression	151
4.5.2	Theil–Sen regression	156
4.5.3	Simple logistic regression	157
	References	159
5	Multivariate statistics	161
5.1	Data matrices	161
5.2	Distance and similarity measures	163
5.2.1	Distance and similarity measures for numeric variables ..	164
5.2.2	Distance and similarity measures for categorical variables	166
5.2.3	Distance and similarity matrices	168
5.3	Multivariate measures of central tendency and variation	170
5.3.1	Centroid and geometric median, medoid	171
5.3.2	Sample covariance and correlation matrix	173
5.4	Random vectors and matrices	174
5.4.1	Expectation vector and covariance matrix	174
5.4.2	Multivariate normal distributions	176
5.4.3	Multinomial distributions	179
	References	181

Part III Machine learning

6	Supervised machine learning	185
6.1	Elements of supervised learning	187
6.1.1	Loss functions and empirical risk minimization	190
6.1.2	Overfitting and underfitting	192
6.1.3	Training, model validation, and testing	196
6.1.4	Numerical optimization	202
6.2	Regression algorithms	206
6.2.1	Linear regression	206
6.2.2	Gaussian process regression	212
6.3	Classification algorithms	216
6.3.1	Logistic regression	216
6.3.2	K -nearest neighbors classification	220
6.3.3	Bayesian classification algorithms	221
6.4	Artificial neural networks	229
6.4.1	Regression and classification with neural networks	231
6.4.2	Training neural networks by backpropagation of error ...	235
6.4.3	Convolutional neural networks	238
	References	246

7	Unsupervised machine learning	249
7.1	Elements of unsupervised learning	249
7.1.1	Intrinsic dimensionality of data	250
7.1.2	Topological characteristics of data	251
7.2	Dimensionality reduction	254
7.2.1	Principal component analysis	254
7.2.2	Autoencoders	258
7.2.3	Multidimensional scaling	259
7.2.4	t -distributed stochastic neighbor embedding (t -SNE)	260
7.3	Cluster analysis	264
7.3.1	K -means algorithm	266
7.3.2	Hierarchical cluster analysis	269
	References	275
8	Applications of machine learning	277
8.1	Supervised learning in practice	278
8.1.1	MNIST: handwritten text recognition	278
8.1.2	CIFAR-10: object recognition	280
8.1.3	Large Movie Review Dataset: sentiment analysis	284
8.2	Unsupervised learning in practice	287
8.2.1	Text mining: topic modelling	287
8.2.2	Network analysis: community structure	288
	References	293

Appendix

A	Exercises with answers	301
A.1	Exercises	301
A.2	Answers	308
	References	324
B	Mathematical preliminaries	327
B.1	Basic concepts	327
B.1.1	Numbers and sets	327
B.1.2	Maps and functions	329
B.1.3	Families, sequences and tuples	331
B.1.4	Minimum/maximum and infimum/supremum	332
B.2	Linear algebra	332
B.2.1	Vectors and points	332
B.2.2	Matrices	333
B.2.3	Subspaces and linear maps	335
B.2.4	Eigenvectors and eigenvalues	336
B.3	Multivariate calculus	337
B.3.1	Limits	337
B.3.2	Continuous functions	338
B.3.3	Differentiable functions	340

B.3.4 Integrals	342
References	347
Supplementary literature	349
Index	353