

Mathematics for Machine Learning

Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong

Contents

<i>List of illustrations</i>	vi
<i>List of tables</i>	x
<i>Preamble</i>	1
Part I Mathematics and Statistics	5
1 Introduction and Motivation	7
1.1 Why We Need Mathematics for Machine Learning	7
1.2 Finding Words for Intuitions	10
1.3 Two Ways To Read This Book	11
1.4 Exercises and Feedback	14
2 Linear Algebra	15
2.1 Systems of Linear Equations	17
2.2 Matrices	19
2.2.1 Matrix Multiplication	20
2.2.2 Inverse and Transpose	21
2.2.3 Multiplication by a Scalar	23
2.2.4 Compact Representations of Systems of Linear Equations	23
2.3 Solving Systems of Linear Equations	24
2.3.1 Particular and General Solution	24
2.3.2 Elementary Transformations	26
2.3.3 The Minus-1 Trick	29
2.3.4 Algorithms for Solving a System of Linear Equations	31
2.4 Vector Spaces	32
2.4.1 Groups	33
2.4.2 Vector Spaces	34
2.4.3 Vector Subspaces	36
2.5 Linear Independence	37
2.6 Basis and Rank	41
2.6.1 Generating Set and Basis	41
2.6.2 Rank	44
2.7 Linear Mappings	45
2.7.1 Matrix Representation of Linear Mappings	46
2.7.2 Basis Change	49
2.7.3 Image and Kernel	54
2.8 Affine Spaces	57

2.8.1	Affine Subspaces	57
2.8.2	Affine Mappings	58
	Exercises	59
3	Analytic Geometry	66
3.1	Norms	67
3.2	Inner Products	68
3.2.1	Dot Product	68
3.2.2	General Inner Products	68
3.2.3	Symmetric, Positive Definite Matrices	69
3.3	Lengths and Distances	71
3.4	Angles and Orthogonality	72
3.5	Inner Products of Functions	73
3.6	Orthogonal Projections	74
3.6.1	Projection onto 1-Dimensional Subspaces (Lines)	76
3.6.2	Projection onto General Subspaces	79
3.6.3	Projection onto Affine Subspaces	82
3.7	Orthonormal Basis	83
3.8	Rotations	84
3.8.1	Properties of Rotations	87
3.9	Further Reading	87
	Exercises	89
4	Matrix Decompositions	90
4.1	Determinants and Traces	91
4.2	Eigenvalues and Eigenvectors	98
4.3	Cholesky Decomposition	106
4.4	Eigendecomposition and Diagonalization	107
4.5	Singular Value Decomposition	111
4.5.1	Geometric Intuitions for the SVD	113
4.5.2	Existence and Construction of the SVD	116
4.5.3	Eigenvalue Decomposition vs Singular Value Decomposition	119
4.6	Matrix Approximation	121
4.7	Matrix Phylogeny	126
4.8	Further Reading	127
	Exercises	128
5	Vector Calculus	134
5.1	Differentiation of Univariate Functions	135
5.1.1	Taylor Series	136
5.1.2	Differentiation Rules	139
5.2	Partial Differentiation and Gradients	140
5.2.1	Basic Rules of Partial Differentiation	141
5.2.2	Chain Rule	142
5.3	Gradients of Vector-Valued Functions	143
5.4	Gradients of Matrices	148
5.5	Useful Identities for Computing Gradients	152
5.6	Backpropagation and Automatic Differentiation	152
5.6.1	Gradients in a Deep Network	153

<i>Contents</i>	iii
5.6.2 Automatic Differentiation	155
5.7 Higher-order Derivatives	158
5.8 Linearization and Multivariate Taylor Series	159
5.9 Further Reading	164
Exercises	164
6 Probability and Distributions	166
6.1 Construction of a Probability Space	166
6.1.1 Philosophical Issues	166
6.1.2 Probability and Random Variables	168
6.1.3 Statistics	169
6.2 Discrete and Continuous Probabilities	170
6.2.1 Discrete Probabilities	170
6.2.2 Continuous Probabilities	172
6.2.3 Contrasting Discrete and Continuous Distributions	173
6.3 Sum Rule, Product Rule and Bayes' Theorem	175
6.4 Summary Statistics and Independence	177
6.4.1 Means and Covariances	178
6.4.2 Three Expressions for the Variance	180
6.4.3 Statistical Independence	181
6.4.4 Sums and Transformations of Random Variables	183
6.4.5 Inner Products of Random Variables	183
6.5 Change of Variables/Inverse transform	185
6.5.1 Distribution Function Technique	186
6.5.2 Change of Variables	188
6.6 Gaussian Distribution	191
6.6.1 Marginals and Conditionals of Gaussians are Gaussians	193
6.6.2 Product of Gaussians	195
6.6.3 Sums and Linear Transformations	196
6.6.4 Sampling from Multivariate Gaussian Distributions	198
6.7 Conjugacy and the Exponential Family	199
6.7.1 Conjugacy	202
6.7.2 Sufficient Statistics	203
6.7.3 Exponential Family	204
6.8 Further Reading	206
Exercises	207
7 Continuous Optimization	209
7.1 Optimization using Gradient Descent	211
7.1.1 Stepsize	212
7.1.2 Gradient Descent with Momentum	213
7.1.3 Stochastic Gradient Descent	214
7.2 Constrained Optimization and Lagrange Multipliers	215
7.3 Convex Optimization	218
7.3.1 Linear Programming	220
7.3.2 Quadratic Programming	222
7.3.3 Legendre-Fenchel Transform and Convex Conjugate	223
7.4 Further Reading	227
Exercises	228

Part II	Foundational Machine Learning Methods	229
8	When Models meet Data	231
8.1	Probabilistic Modeling and Latent Variables	236
8.2	Parameter Estimation	238
8.2.1	Maximum Likelihood Estimation	238
8.2.2	Maximum A Posteriori Estimation	241
8.2.3	Further Reading	242
8.3	Empirical Risk Minimization	242
8.3.1	Hypothesis Class	242
8.3.2	Loss Function	243
8.3.3	Regularization	244
8.3.4	Further Reading	245
8.4	Model Selection and Cross Validation	245
8.4.1	Cross-Validation to Assess the Generalization Performance	246
8.4.2	Bayesian Model Selection	247
8.4.3	Bayes Factors for Model Comparison	249
8.4.4	Further Reading	250
8.5	Directed Graphical Models	250
8.5.1	Graph Semantics	252
8.5.2	From Joint Distributions to Graphs	252
8.5.3	From Graphs to Joint Distributions	254
8.5.4	Conditional Independence and D-Separation	255
8.5.5	Further Reading	256
9	Linear Regression	257
9.1	Problem Formulation	259
9.2	Parameter Estimation	260
9.2.1	Maximum Likelihood Estimation	260
9.2.2	Overfitting in Linear Regression	265
9.2.3	Regularization and Maximum A Posteriori Estimation	267
9.3	Bayesian Linear Regression	270
9.3.1	Model	271
9.3.2	Prior Predictions	271
9.3.3	Posterior Distribution	272
9.3.4	Posterior Predictions	276
9.3.5	Computing the Marginal Likelihood	278
9.4	Maximum Likelihood as Orthogonal Projection	280
9.5	Further Reading	282
10	Dimensionality Reduction with Principal Component Analysis	285
10.1	Problem Setting	286
10.2	Maximum Variance Perspective	287
10.3	Projection Perspective	291
10.3.1	Setting and Objective	291
10.3.2	Optimization	293
10.4	Eigenvector Computation	298
10.5	PCA Algorithm	300
10.6	PCA in High Dimensions	302

<i>Contents</i>	v
10.7 Probabilistic Principal Component Analysis	303
10.7.1 Generative Process and Probabilistic Model	304
10.7.2 Likelihood and Joint Distribution	305
10.7.3 Posterior Distribution	306
10.8 Further Reading	307
11 Density Estimation with Gaussian Mixture Models	312
11.1 Gaussian Mixture Model	313
11.2 Parameter Learning via Maximum Likelihood	314
11.3 EM Algorithm	324
11.4 Latent Variable Perspective	326
11.4.1 Prior	327
11.4.2 Marginal	328
11.4.3 Posterior	328
11.4.4 Extension to a Full Dataset	328
11.4.5 EM Algorithm Revisited	329
11.5 Further Reading	329
12 Classification with Support Vector Machines	332
12.1 Separating Hyperplanes	334
12.2 Primal Support Vector Machine	335
12.2.1 Concept of the Margin	336
12.2.2 Traditional Derivation of the Margin	338
12.2.3 Why we can set the Margin to 1	340
12.2.4 Soft Margin SVM: Geometric View	341
12.2.5 Soft Margin SVM: Loss Function View	342
12.3 Dual Support Vector Machine	344
12.3.1 Convex Duality via Lagrange Multipliers	344
12.3.2 Convex Duality via the Convex Conjugate	346
12.3.3 Soft Margin SVM: Convex Hull View	349
12.3.4 Kernels	351
12.3.5 Numerical Solution	353
12.4 Further Reading	355
<i>References</i>	357
<i>Index</i>	367