

Introduction
to
Machine
Learning

Second
Edition

Adaptive Computation and Machine Learning

Thomas Dietterich, Editor

Christopher Bishop, David Heckerman, Michael Jordan, and Michael
Kearns, Associate Editors

A complete list of books published in The Adaptive Computation and
Machine Learning series appears at the back of this book.

Introduction
to
Machine
Learning

Second
Edition

Ethem Alpaydm

The MIT Press
Cambridge, Massachusetts
London, England

© 2010 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quantity discounts, please email
special_sales@mitpress.mit.edu.

Typeset in 10/13 Lucida Bright by the author using $\text{\LaTeX} 2\epsilon$.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Information

Alpaydin, Ethem.

Introduction to machine learning / Ethem Alpaydin. — 2nd ed.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-262-01243-0 (hardcover : alk. paper)

1. Machine learning. I. Title

Q325.5.A46 2010

006.3'1—dc22

2009013169

CIP

10 9 8 7 6 5 4 3 2 1

- Binding, 202
- Binomial test, 499
- Biometrics, 441
- Blocking, 482
- Bonferroni correction, 508
- Boosting, 431
- Bootstrap, 489

- C4.5, 191
- C4.5Rules, 197
- CART, 191, 203
- Cascade correlation, 264
- Cascading, 438
- Case-based reasoning, 180
- Causality, 396
 - causal graph, 388
- Central limit theorem, 526
- Class
 - confusion matrix, 493
 - likelihood, 50
- Classification, 5
 - likelihood- vs. discriminant-based, 209
- Classification tree, 188
- Clique, 411
- Cluster, 144
- Clustering, 11
 - agglomerative, 157
 - divisive, 157
 - hierarchical, 157
 - online, 281
- Code word, 146
- Codebook vector, 146
- Coefficient of determination (of regression), 76
- Color quantization, 145
- Common principal components, 119
- Competitive basis functions, 297
- Competitive learning, 280
- Complete-link clustering, 158
- Component density, 144

- Compression, 8, 146
- Condensed nearest neighbor, 173
- Conditional independence, 389
- Confidence interval
 - one-sided, 495
 - two-sided, 494
- Confidence of an association rule, 55
- Conjugate prior, 344
- Connection weight, 237
- Contingency table, 501
- Correlation, 89
- Cost-sensitive learning, 478
- Coupled HMM, 400
- Covariance function, 356
- Covariance matrix, 88
- Credit assignment, 448
- Critic, 448
- CRM, *see* Customer relationship management
- Cross-entropy, 221
- Cross-validation, 40, 80, 486
 - 5×2 , 488
 - K -fold, 487
- Curse of dimensionality, 170
- Customer relationship management, 155
- Customer segmentation, 155

- d -separation, 402
- Decision node, 185
- Decision region, 53
- Decision tree, 185
 - multivariate, 202
 - omnivariate, 205
 - soft, 305
 - univariate, 187
- Delve repository, 17
- Dendrogram, 158
- Density estimation, 11
- Dichotomizer, 53
- Diffusion kernel, 325

- Dimensionality reduction
 - nonlinear, 269
- Directed acyclic graph, 387
- Dirichlet distribution, 344
- Discount rate, 451
- Discriminant, 5
 - function, 53
 - linear, 97
 - quadratic, 95
- Discriminant adaptive nearest neighbor, 172
- Discriminant-based classification, 209
- Distributed vs. local representation, 156, 289
- Diversity, 420
- Divisive clustering, 157
- Document categorization, 102
- Doubt, 26
- Dual representation, 337, 352
- Dynamic classifier selection, 435
- Dynamic graphical models, 415
- Dynamic node creation, 264
- Dynamic programming, 453

- Early stopping, 223, 258
- ECOC, 327, *see* Error-correcting output codes
- Edit distance, 324
- Eigendigits, 118
- Eigenfaces, 118
- Eligibility trace, 459
- EM, *see* Expectation-Maximization
- Emission probability, 367
- Empirical error, 24
- Empirical kernel map, 324
- Ensemble, 424
- Ensemble selection, 437
- Entropy, 188
- Episode, 451
- Epoch, 251
- Error
 - type I, 497
 - type II, 497
- Error-correcting output codes, 427
- Euclidean distance, 98
- Evidence, 50
- Example, 87
- Expectation-Maximization, 150
 - supervised, 299
- Expected error, 476
- Expected utility, 54
- Experiment
 - design, 478
 - factorial, 481
 - strategies, 480
- Explaining away, 393
- Extrapolation, 35

- FA, *see* Factor analysis
- Factor analysis, 120
- Factor graph, 412
- Factorial HMM, 400
- Feature, 87
 - extraction, 110
 - selection, 110
- Finite-horizon, 451
- First-order rule, 201
- Fisher kernel, 325
- Fisher's linear discriminant, 129
- Flexible discriminant analysis, 120
- Floating search, 112
- Foil, 199
- Forward selection, 110
- Forward variable, 370
- Forward-backward procedure, 370
- Fuzzy k -means, 160
- Fuzzy membership function, 295
- Fuzzy rule, 295

- Gamma distribution, 347
- Gamma function, 344
- Gaussian prior, 349
- Generalization, 24, 39

- Generalized linear models, 230
- Generative model, 342, 397
- Generative topographic mapping, 306
- Geodesic distance, 133
- Gini index, 189
- Gradient descent, 219
 - stochastic, 241
- Gradient vector, 219
- Gram matrix, 321
- Graphical models, 387
- Group, 144
- GTM, *see* Generative topographic mapping
- Hamming distance, 171
- Hebbian learning, 283
- Hidden layer, 246
- Hidden Markov model, 367, 398
 - coupled, 400
 - factorial, 400
 - input-output, 379, 400
 - left-to-right, 380
 - switching, 400
- Hidden variables, 57, 396
- Hierarchical clustering, 157
- Hierarchical cone, 260
- Hierarchical mixture of experts, 304
- Higher-order term, 211
- Hinge loss, 317
- Hint, 261
- Histogram, 165
- HMM, *see* Hidden Markov model
- Hybrid learning, 291
- Hypothesis, 23
 - class, 23
 - most general, 24
 - most specific, 24
- Hypothesis testing, 496
- ID3, 191
- IF-THEN rules, 197
- lid (independent and identically distributed), 41
- Ill-posed problem, 38
- Impurity measure, 188
- Imputation, 89
- Independence, 388
- Inductive bias, 38
- Inductive logic programming, 202
- Infinite-horizon, 451
- Influence diagrams, 414
- Information retrieval, 491
- Initial probability, 364
- Input, 87
- Input representation, 21
- Input-output HMM, 379, 399
- Instance, 87
- Instance-based learning, 164
- Interest of an association rule, 55
- Interpolation, 35
- Interpretability, 197
- Interval estimation, 493
- Irep, 199
- Isometric feature mapping, 133
- Job shop scheduling, 471
- Junction tree, 410
- K -armed bandit, 449
- K -fold
 - cross-validation, 487
 - cv paired t test, 502
- k -means clustering, 147
 - fuzzy, 160
 - online, 281
- k -nearest neighbor
 - classifier, 172
 - density estimate, 169
 - smoother, 177
- k -nn, *see* k -nearest neighbor
- Kalman filter, 400
- Karhunen-Loève expansion, 119
- Kernel estimator, 167

- Kernel function, 167, 320, 353
- Kernel PCA, 336
- Kernel smoother, 176
- kernelization, 321
- Knowledge extraction, 8, 198, 295
- Kolmogorov complexity, 82
- Kruskal-Wallis test, 511

- Laplace approximation, 354
- Laplacian prior, 350
- lasso, 352
- Latent factors, 120
- Lateral inhibition, 282
- LDA, *see* Linear discriminant analysis
- Leader cluster algorithm, 148
- Leaf node, 186
- Learning automata, 471
- Learning vector quantization, 300
- Least square difference test, 507
- Least squares estimate, 74
- Leave-one-out, 487
- Left-to-right HMM, 380
- Level of significance, 497
- Levels of analysis, 234
- Lift of an association rule, 55
- Likelihood, 62
- Likelihood ratio, 58
- Likelihood-based classification, 209
- Linear classifier, 97, 216
- Linear discriminant, 97, 210
- Linear discriminant analysis, 128
- Linear dynamical system, 400
- Linear opinion pool, 424
- Linear regression, 74
 - multivariate, 103
- Linear separability, 215
- Local representation, 288
- Locally linear embedding, 135
- Locally weighted running line smoother, 177

- Loess, *see* Locally weighted running line smoother
- Log likelihood, 62
- Log odds, 58, 218
- Logistic discrimination, 220
- Logistic function, 218
- Logit, 218
- Loss function, 51
- LSD, *see* Least square difference test
- LVQ, *see* Learning vector quantization

- Mahalanobis distance, 90
- Margin, 25, 311, 433
- Markov decision process, 451
- Markov mixture of experts, 379
- Markov model, 364
 - hidden, 367
 - learning, 366, 375
 - observable, 365
- Markov random field, 410
- Max-product algorithm, 413
- Maximum a posteriori (MAP)
 - estimate, 68, 343
- Maximum likelihood estimation, 62
- McNemar's test, 501
- MDP, *see* Markov decision process
- MDS, *see* Multidimensional scaling
- Mean square error, 65
- Mean vector, 88
- Memory-based learning, 164
- Minimum description length, 82
- Mixture components, 144
- Mixture density, 144
- Mixture of experts, 301, 434
 - competitive, 304
 - cooperative, 303
 - hierarchical, 305
 - Markov, 379, 400
- Mixture of factor analyzers, 155
- Mixture of mixtures, 156

- Mixture of probabilistic principal component analyzers, 155
- Mixture proportion, 144
- MLE, *see* Maximum likelihood estimation
- Model combination
 - multiexpert, 423
 - multistage, 423
- Model selection, 38
- MoE, *see* Mixture of experts
- Momentum, 257
- Moralization, 411
- Multidimensional scaling, 125
 - nonlinear, 287
 - using MLP, 269
- Multilayer perceptrons, 246
- Multiple comparisons, 507
- Multiple kernel learning, 326, 442
- Multivariate linear regression, 103
- Multivariate polynomial regression, 104
- Multivariate tree, 202

- Naive Bayes' classifier, 397
 - discrete inputs, 102
 - numeric inputs, 97
- Naive estimator, 166
- Nearest mean classifier, 98
- Nearest neighbor classifier, 172
 - condensed, 173
- Negative examples, 21
- Neuron, 233
- No Free Lunch Theorem, 477
- Noise, 30
- Noisy OR, 409
- Nonparametric estimation, 163
- Nonparametric tests, 508
- Null hypothesis, 497

- Observable Markov model, 365
- Observable variable, 48
- Observation, 87
- Observation probability, 367

- OC1, 203
- Occam's razor, 32
- Off-policy, 458
- Omnivariate decision tree, 205
- On-policy, 458
- One-class classification, 333
- One-sided confidence interval, 495
- One-sided test, 498
- Online k -means, 281
- Online learning, 241
- Optimal policy, 452
- Optimal separating hyperplane, 311
- Outlier detection, 9, 333
- Overfitting, 39, 79
- Overtraining, 258

- PAC, *see* Probably approximately correct
- Paired test, 501
- Pairing, 482
- Pairwise separation, 216, 428
- Parallel processing, 236
- Partially observable Markov decision process, 464
- Parzen windows, 167
- Pattern recognition, 6
- PCA, *see* Principal components analysis
- Pedigree, 400
- Perceptron, 237
- Phone, 381
- Phylogenetic tree, 398
- Piecewise approximation
 - constant, 248, 300
 - linear, 301
- Policy, 451
- Polychotomizer, 53
- Polynomial regression, 75
 - multivariate, 104
- Polytree, 407
- POMDP, *see* Partially observable Markov decision process

- Positive examples, 21
- Posterior probability distribution, 341
- Posterior probability of a class, 50
- Posterior probability of a parameter, 67
- Posthoc testing, 507
- Postpruning, 194
- Potential function, 212, 411
- Power function, 498
- Precision
 - in information retrieval, 492
 - reciprocal of variance, 347
- Predicate, 201
- Prediction, 5
- Prepruning, 194
- Principal components analysis, 113
- Prior knowledge, 294
- Prior probability distribution, 341
- Prior probability of a class, 50
- Prior probability of a parameter, 67
- Probabilistic networks, 387
- Probabilistic PCA, 123
- Probably approximately correct learning, 29
- Probit function, 355
- Product term, 211
- Projection pursuit, 274
- Proportion of variance, 116
- Propositional rule, 201
- Pruning
 - postpruning, 194
 - prepruning, 194
 - set, 194
- Q learning, 458
- Quadratic discriminant, 95, 211
- Quantization, 146
- Radial basis function, 290
- Random Subspace, 421
- Randomization, 482
- RBF, *see* Radial basis function
- Real time recurrent learning, 272
- Recall, 492
- Receiver operating characteristics, 490
- Receptive field, 288
- Reconstruction error, 119, 146
- Recurrent network, 271
- Reference vector, 146
- Regression, 9, 35
 - linear, 74
 - polynomial, 75
 - polynomial multivariate, 104
 - robust, 329
- Regression tree, 192
- Regressogram, 175
- Regularization, 80, 266
- Regularized discriminant analysis, 100
- Reinforcement learning, 13
- Reject, 34, 52
- Relative square error, 76
- Replication, 482
- Representation, 21
 - distributed vs. local, 288
- Response surface design, 481
- Ridge regression, 266, 350
- Ripper, 199
- Risk function, 51
- Robust regression, 329
- ROC, *see* Receiver operating characteristics
- RSE, *see* Relative square error
- Rule
 - extraction, 295
 - induction, 198
 - pruning, 198
- Rule support, 198
- Rule value metric, 199
- Running smoother
 - line, 177
 - mean, 175

- Sammon mapping, 128
 - using MLP, 269
- Sammon stress, 128
- Sample, 48
 - correlation, 89
 - covariance, 89
 - mean, 89
- Sarsa, 458
 - Sarsa(λ), 461
- Scatter, 129
- Scree graph, 116
- Self-organizing map, 286
- Semiparametric density estimation, 144
- Sensitivity, 493
- Sensor fusion, 421
- Sequential covering, 199
- Sigmoid, 218
- Sign test, 509
- Single-link clustering, 157
- Slack variable, 315
- Smoother, 174
- Smoothing splines, 178
- Soft count, 376
- Soft error, 315
- Soft weight sharing, 267
- Softmax, 224
- SOM, *see* Self-organizing map
- Spam filtering, 103
- Specificity, 493
- Spectral decomposition, 115
- Speech recognition, 380
- Sphere node, 203
- Stability-plasticity dilemma, 281
- Stacked generalization, 435
- Statlib repository, 17
- Stochastic automaton, 364
- Stochastic gradient descent, 241
- Stratification, 487
- Strong learner, 431
- Structural adaptation, 263
- Structural risk minimization, 82
- Subset selection, 110
- Sum-product algorithm, 412
- Supervised learning, 9
- Support of an association rule, 55
- Support vector machine, 313
- SVM, *see* Support vector machine
- Switching HMM, 400
- Synapse, 234
- Synaptic weight, 237

- t distribution, 495
- t test, 498
- Tangent prop, 263
- TD, *see* Temporal difference
- Template matching, 98
- Temporal difference, 455
 - learning, 458
 - TD(0), 459
 - TD-Gammon, 471
- Test set, 40
- Threshold, 212
 - function, 238
- Time delay neural network, 270
- Topographical map, 287
- Transition probability, 364
- Traveling salesman problem, 306
- Triple trade-off, 39
- Tukey's test, 512
- Two-sided confidence interval, 494
- Two-sided test, 497
- Type 2 maximum likelihood
 - procedure, 360
- Type I error, 497
- Type II error, 497

- UCI repository, 17
- Unbiased estimator, 65
- Underfitting, 39, 79
- Unfolding in time, 272
- Unit normal distribution, 493
- Univariate tree, 187
- Universal approximation, 248

- Unobservable variable, 48
- Unstable algorithm, 430
- Utility function, 54
- Utility theory, 54

- Validation set, 40
- Value iteration, 453
- Value of information, 464, 469
- Vapnik-Chervonenkis (VC)
 - dimension, 27
- Variance, 66
- Vector quantization, 146
 - supervised, 300
- Version space, 24
- Vigilance, 285
- Virtual example, 262
- Viterbi algorithm, 374
- Voronoi tessellation, 172
- Voting, 424

- Weak learner, 431
- Weight
 - decay, 263
 - sharing, 260
 - sharing soft, 267
 - vector, 212
- Wilcoxon signed rank test, 511
- Winner-take-all, 280
- Within-class scatter matrix, 130
- Wrappers, 138

- z , *see* Unit normal distribution
- z -normalization, 91, 526
- Zero-one loss, 51