

INDEX

- AdaBoost, 392
- Adaptive methods, 40–54, 272, 277–279
- Additive models, 279
- Algorithmic complexity, 51
- Annealed VC entropy, 106

- Backfitting, 154, 279
 - algorithm, 279
 - complexity control, 283
- Backpropagation, 156–161
 - complexity control, 289–291
 - online versus batch implementations, 287
 - regularization effect of initialization, 291
 - with momentum term, 286
- Bagging, 390
- Basis functions, 250
 - equivalent kernel, 257, 261, 428
 - nonadaptive and adaptive, 250
- Bayes decision rule, 350–351
- Bayes theorem, 342
- Bayesian interpretation of probability, 12
- Bayesian inference, 47–51
 - empirical Bayesian, 74
 - marginalization, 49
 - maximum a posteriori probability, 49
 - relationship to penalization inductive principle, 50
- Bias, 80
- Bias variance trade-off, 80
- Binary tree, *see* Classification and regression trees
- Biological systems, 2
- Bivariate Gaussian, 188, 207, 217
- Blocks signal, 305
- Boosting, 390
 - algorithm, 392
 - for classification, 391
 - for regression, 400
- Bridge penalty, 72

- Causality, 8
- Centers, *see* Prototype vectors
- Characteristic polynomial of a matrix, 516
- Classification, 340
 - classical, 32
 - compensating for different prior probabilities, 370
 - mixture of Gaussians, 385, 435
 - problem statement, 25
 - tree-based methods, 378

- Classification (*Continued*)
 via AdaBoost, 392
 via constrained topological mapping, 374
 via multilayer perceptrons, 372
 via nearest-neighbors, 382
 via radial basis functions, 373
 via support vector machine, 430
 with unequal costs of misclassification, 370
- Classification and regression trees (CART), 170, 378
 algorithm, 381
- Clustering, 29, 191. *See also* Vector quantization
 fuzzy approaches, 195
- Combining methods, 390
- Combining predictive models, 332
- Committee of networks, 333
- Competitive learning, 189
- Complexity, *see also* VC-dimension
 characterization of, 66
 control of, *see* Model selection
 estimates of, 263–269
- Conscience mechanism, 191
- Consistency of empirical risk minimization, 103
 distribution-independent conditions, 106
 nontrivial, 104
- Constrained topological mapping (CTM), 314
 algorithm for classification, 374–378
 algorithm for regression, 316–319
 complexity control, 317
- Convergence of the empirical risk, 106
- Cross-entropy loss, 368
- Cross-validation, 78
 analytic form for linear estimators, 262
- Cubic spline, 273
- Curse of dimensionality, 62, 97, 383
 consequences of, 64
- Data compression, 183. *See also* Minimum description length
- Data mining, 15, 18
- Data piling, 469, 470–472
- Decision boundary, 342, 359
- Decision rule, 342, 355
- Degrees of freedom, *see* Effective degrees of freedom
- Delta rule, 156, 286
- Density estimation, 30
 expectation-maximization algorithm, 161
 mixture of normals, 31
 nonparametric, 36
 problem statement, 28
- Designed experiment, 5, 29
- Dictionary methods, 68, 251, 277
- Dictionary representation, 124, 174
- Dimensionality reduction, 29, 178, 201, 214, 283, 315
- Direct ad hoc inference (DAHI), 466
- Discriminant functions, 346, 355–357.
See also Fisher Linear Discriminant
- Duality
 in the least squares problem, 260
 in optimization theory, 407
 of kernel and basis function representations, 255
- Early stopping, 40, 46, 128, 289
- Effective degrees of freedom, 75–77, 128–132, 264–267. *See also* VC-dimension
- Eigenvalues of the smoother matrix, 261, 264
- Eigenvectors of a matrix, 516
- Electrocardiogram (ECG), 308
- Empirical inference science, 100, 501
- Empirical risk minimization, 30, 31, 45, 100, 152. *See also* Inductive principles
- Entropy function, 379
- Epistemology, 503
- Epochs, 40, 160
- Equivalence classes, 406, 477, 482, 497
- Exclusive-or problem, 432
- Expectation-maximization, 153, 161
 algorithm for clustering, 192
- Experimental procedure, 15, 182
- Factor analysis, 232
- False negative, 351
- False positive, 351
- Falsifiability, 110, 146
 degree of, 147
 VC falsifiability, 147
 and Occam's razor, 148
 and simplicity, 148
- Feature selection, 6, 125, 173, 405
 nonlinear, 174

- Feature space, 201, 215, 406, 426
- Final prediction error, 76
- First-principle models, 2
- Fisher linear discriminant, 358, 362–366
- Formal problem statement, 16. *See also*
 - Learning problem setting/
 - formulation
- Fourier
 - series, 298
 - transform, 69, 73
- Frequentist, 11
- Function approximation, 17, 24, 88. *See also*
 - System identification
- Function complexity, 68
- Fuzzy, 11
 - Another Fuzzy Clustering (AFC)
 - algorithm, 200
 - c-means, 196
 - clustering, 195
 - membership function, 12
 - set, 11, 12
- Gauss-Newton method, 512
- Generalization
 - distribution-independent bounds, 115, 116, 118
 - false, 110
 - of optimal separating hyperplane, 419
- Generalized cross validation, 77, 296
- Generalized inverse, 164, 518
- Generalized Lloyd algorithm, 187, 196
- Generalized memory-based learning, 313
- Gini function, 379, 380
- Gradient descent, 153–158. *See also*
 - Stochastic approximation
- Greedy optimization, 154, 169, 174, 279, 294, 378
- Growth function, 105, 107
- Hat matrix, 260
- Heavisine signal, 305
- Hebbian rule, 156
- Hessian matrix, 508
- Hidden layer, 157, 160, 284
- High-dimensional distributions, 63
- Hilbert space, 428
- Hints, 60, 260
- Histogram, 36
- Hyperplane
 - separable, 410, 411, 418
 - nonseparable, 411, 412, 424
- Independent component analysis, 242
- Indicator functions, 26, 111–114, 341
- Inductive principles, 25, 40, 42, 45–55
 - properties of, 54
- Inference through contradictions, 481–485
- Information theory, 183
- Interpretation of predictive models, 7, 259, 382, 494, 505–506
- Inverse of a matrix, 514–518
- Jensen’s inequality, 106
- Kernel function, 254–255
 - equivalent basis function, 257, 260–261
 - inner product, 426–430
 - properties, 22–23
 - span (width), 23, 310
- Kernel methods, 309. *See also* Local risk
 - minimization
 - classification, 382–385
 - density estimation, 38–39
 - regression, 22, 254
- Key theorem of learning theory, 104
- Knowledge, 503
 - empirical, 504
 - instrumental, 505
 - provisional (relativistic), 504
- Kolmogorov’s theorem, 66
- Kuhn-Tucker theorem, 421–423
- Kullback-Leibler criterion, 369
- Lagrange multipliers, 422
- Lagrangian, 352, 422
- Learning, 21
- Learning imperative, 503. *See also* Learning
 - problem formulation. *See also* Noninductive inference
- Learning machine, 21
- Learning methods
 - empirical comparisons, 326–331, 385–389
- Learning problem formulation, 57, 58, 467. *See also* Formal (learning) problem
 - statement Learning rate, 39, 156, 222–224, 286, 509

- Learning problem formulation (*Continued*)
 - for self-organizing map, 222–224
 - for backpropagation, 286
 - for learning vector quantization, 385
- Learning vector quantization, 384–385
- Least squares estimation, 258
 - linear, 259
 - nonlinear, 151
 - penalized linear, 259
- Least squares solution of a linear system, 515
 - via singular value decomposition, 516
- Left inverse, 515, 516
- LeNet, 1 436
- Levenberg-Marquadt method, 512
- Likelihood function, 30, 47–48
- Likelihood ratio, 350, 352
- Linear discriminant analysis, 358, 362–366.
 - See also* Fisher Linear Discriminant
- Linear estimators, 75, 256
- Linear matrix equations, 514
- Linear regression, 258, 259, 321, 322.
 - See also* Least squares estimation
- Linear subset regression, 140
- Linearly separable, 418
- Lloyd-Max conditions, 183, 185–186
- Local risk minimization, 309–313
 - for classification, 382–385
 - practical complexity control, 312–313
- Locally weighted linear approximation, 313
- Log-likelihood, 28
- Loss, 25, 27
 - for classification, 26
 - for regression, 26
 - for density estimation, 28
 - for vector quantization, 28
 - for multiple model estimation, 488
 - margin-based, 408–414
 - exponential, 399
- Mahalanobis distances, 356
- Margin, 406, 408, 454
- Margin-based loss, 408, 414
- Maximum likelihood, 30. *See also*
 - Empirical risk minimization
- Mercer's conditions, 428–429
- Minimum spanning tree, 225
- Minimum description length (MDL), 51, 116, 420
- Model selection, 73
 - analytical approach, 75, 262,
 - case studies, 128, 267
 - resampling approach, 78
- Multidimensional scaling, 209
- Multilayer perceptron (MLP), 156–157
 - for dimensionality reduction, 230
 - for classification, 346
 - for regression, 253, 284
 - support vector machine implementation, 429, 436
- Multiple model estimation (MME), 469, 486
 - double-SVM method, 491, 494
 - greedy procedure for MME 489
 - for classification, 491
 - for regression, 494
- Multivariate adaptive regression splines (MARS), 293
 - algorithm, 296–297
 - complexity control, 296
 - interpretation via anova decomposition, 297
 - relationship with CART, 293
- Mutual fund, 321
- Net Asset Value (NAV) of a mutual fund, 320
- Network growing algorithms, 169
- Neural networks, 154. *See also* Multilayer perceptron
 - construction, 169
- Newton methods for optimization, 510
- Noninductive inference, 467, 469, 501, 506
- Nonparametric methods, 36
- Normal equations, 518
- Occam's razor, 43, 146. *See also*
 - Model selection
- Optimal brain damage, 291
- Optimization, 151
 - direct search methods, 509
 - for minimizing classification error, 341, 346
 - nonlinear, 507
 - nonlinear least-squares, 511
 - second order methods, 510
 - steepest descent methods, 509
- Orthonormal basis functions, 298

- Outliers, 5, 197, 297, 417, 486
- Overfitting, 82, 124. *See also* Model selection
- Parameter estimation, 154
- Partial Least Squares, 283
- Pedagogical pattern selection, 29
- Penalization, *see* Regularization
- Perceptron algorithm, 345
- Philosophy of natural science, 109
- Phoneme clustering, 224–225
- Polynomial basis, 429
- Polynomial decision boundary, 427
- Polynomial estimators, 125, 269, 405
- Popper, 109
 - dimension, 148
 - falsifiability, 110, 148
- Porcupine, 63, 470
- Postal zipcodes, 435
- Posterior probability, 47
 - estimated via conditional expectation, 358
 - estimation for classification, 347
- Precision-recall tradeoff, 354
- Prediction vs. approximation, 24, 88, 453
- Predictive learning, 15, 17. *See also* System imitation
- Premature saturation, 287
- Preprocessing, 5, 126, 182, 603
- Principal component analysis, 202, 212, 234, 242
 - properties of, 203
- Principal curves, 205. *See also* Self-organizing map
 - self-consistency conditions, 206
- Principle of VC-falsifiability, 148
- Prior probabilities, 33
- Projection pursuit, 204, 279
 - algorithm, 281–282
 - relationship with multilayer perceptron, 284
- Prototype vectors, 178, 382
- Pruning, 154, 282, 378
- Pseudoinverse, 515, 516. *See also* Generalized inverse
- Radial basis function networks, 73, 182, 275, 373, 429, 476
 - algorithm, 276–277
 - selection of centers and widths, 277
- Random entropy *see* VC-entropy
- Rank of a matrix, 514
- Receiver operating characteristic (ROC), 352
- Recycling, 40
- Regression, 249
 - classical, 34
 - estimation of posterior probabilities, 357
 - estimation of principal curves, 207
 - kernel representation, 260
 - problem statement, 26
 - taxonomy of methods, 250
 - via support vector machines, 439
 - within self-organizing map algorithm, 219
- Regularization, 46, 61, 88, 90, 91, 127, 497, 503. *See also* Inductive principles, *See also* Function approximation, *See also* System identification.
 - effects of backpropagation, 288
 - in splines, 271
 - least squares, 259
 - nonparametric penalties, 73
 - parametric penalties, 72
 - related to support vector machine, 453
- Resampling, 78
- Ridge penalty, 72, 126, 259, 365
- Right inverse, 515
- Risk functional, 25
- Robust regression, 415, 469
- Saddle point, 508
- Sammon Mapping, 213
- Sampling theorem, 69, 120
- Scaling, of data, 182, 277, 317, 362, Schwartz' criterion, 77
- Self-organization, 218
- Self-organizing map, 214, 314, 374
 - controlling complexity, 220
 - neighborhood, 218
- Semisupervised learning, 469, 476, 480
- Separating hyperplane, 418
- Shape skeleton, 228
- Shattering, 108
- Shibata's model selector, 77, 129
- Sigmoid function, 125, 164, 253, 346, 429
- Single-class SVM, 460
- Singular value decomposition, 517
- Slack variables, 412, 478

- Smooth multiple additive regression
 - technique (SMART), 282
- Soft margin hyperplane, 425
- Sparse feature selection, 174
- Sparse high-dimensional data, 470–474
- Spline, 271
 - basis for support vector machines, 429
 - knot selection strategies, 272
 - multivariate, 294
- Squared error distortion, 180
- Stacking predictors, 333
- Statistical decision theory, 348, 358
- Statistical dependency, 7
- Statistical learning theory (also known as VC-theory), 99, 343
- Statistical model estimation, 14, 17. *See also*
 - System identification
- Stochastic approximation, 39, 153, 154
- Structural risk minimization, 47, 89, 122, 478. *See also* Inductive principles
 - for classification, 341
 - in the support vector machine, 406
- Structure, 122
 - dictionary, 124
 - feature selection, 125
 - penalization, 126
 - input preprocessing, 126
 - initialization, 127
- Sufficient statistic, 355
- Superposition principle for linear
 - estimators, 256
- Supervised learning, 3
- Support vector machine, 404
 - inner product kernels, 426
 - Fourier kernel, 430
 - polynomial kernel, 429
 - radial basis function kernel, 429
 - spline kernel, 429–430
 - optimization problem statement, 430–431, 441–442
 - vs regularization, 453
 - model selection, 445, 454–459
- Support vector data description (SVDD), 460
- Support vectors, 407, 417, 419, 422
- SVM-Plus, 466
- System identification, 89, 453. *See also*
 - Function approximation
- System imitation, 57, 89, 453, 502. *See also*
 - Predictive learning
- Tangent distance, 436
- Taxonomy of regression methods, 250
- Tensor product splines, 274–275
- Time series prediction, 29
- Trace of a matrix, 517
- Transduction, 41, 474, 502
- Tree-structured self-organizing map, 224
- Uncertainty, 11
- Units, *see* Prototype vectors
- Universal approximators, 67
 - examples of, 67
- Universum data, 469, 481–485. *See also*
 - Inference through contradictions
- Unsupervised learning, 3–4, 178
- Vapnik’s imperative, 502
- Variables, 10
- Variance, 82
 - estimates for linear estimators, 84
- VC entropy, 105
- VC-dimension, 107, 147, 266–267, 304, 408, 420, 501
 - measuring, 143
 - for classification and regression, 110
 - of a set of indicator functions, 108
 - of a set of linear indicator functions, 111
 - of a set of real-valued functions, 108
- VC-theory, *see* Statistical learning theory
- Vector quantization, 178, 183
 - problem statement, 28
- Virtual SV method, 474
- Voronoi regions, 185
- Waveform data, 388
- Wavelets, 298
 - complexity control, 303
- Weight decay, 290
- Weight decay penalty, 72–73
- Working set, 469, 475, 477, 480
- Worst-case analysis, 104