

# INDEX

- AdaBoost, 346-349  
 loss function, 351  
 generalization, 352-353  
 classification of MNIST data, 354  
 tuning parameters, 356
- Adaptive methods, 175. *See also* Nonlinear methods
- Additive modeling, 191
- Age of Reason, 79
- Aristotle, 78-79
- Backfitting, 191. *See also* Additive modeling, Projection pursuit  
 for regression, 191-193  
 for projection pursuit, 195  
 for AdaBoost, 351
- Backpropagation, 224-225  
 complexity control, 227-228, 231  
 early stopping, 228
- Bagging or bootstrap aggregation, 339-342  
 classification example, 342-344  
 regression example, 344
- Basis functions, 148, 175-176
- Bayes decision rule, 123
- Bayesian inference, *see also* Inductive principle
- Bayesian averaging, 12
- Beliefs, 11. *See also* Empirical knowledge
- Bias variance trade-off, 209, 213
- Binary tree, *see* Decision trees, CART
- Bin-based model, 444-445
- Boosting, 345-346
- Bootstrap sample, 340
- Centers, 43, 240. *See also* Prototype vectors
- Classification, 37-39  
 nearest neighbor, 51-52  
 via ERM, 124-128  
 via logistic regression, 169-173  
 via multiple-response regression, 178  
 via MLP networks, 229-232  
 via RBF networks, 237-239  
 with unequal costs, 123, 306, 439  
 via CART, 185-187  
 via AdaBoost, 345-347  
 via support vector machine, 284-292  
 via transduction, 367-368
- Classification and regression trees (CART), 180  
 for regression, 181-183  
 for classification, 185-187
- Clustering, 43, 240, 446. *See also* Vector quantization  
 of European languages, 256
- Cognitive bias, 415-416
- Combining methods, 327-330
- Committee of networks, 330-333
- Complexity control, 55-56. *See also* Model selection  
 analytic, 57  
 via resampling, 57-58
- Consistency of empirical risk minimization, 129-131
- Cross-validation, 58-59. *See also* Resampling  
 generalized cross-validation, 57
- Data piling, 305
- Data reduction, 52. *See also* Minimum Description Length
- Decision rule, 123, 277
- Decision trees, 180. *See also* CART
- Degrees of freedom, 57
- Delta rule, 223. *See also* Backpropagation, gradient-descent
- Dictionary structure, 147
- Dimensionality reduction, 247. *See also* SOM
- Discriminant function, 125-128. *See also* Decision rule
- Dual optimization formulation, 286  
 for SVM classification, 287, 296  
 for SVM regression, 310-311
- Early stopping, 228
- Ensemble methods, 327-330
- Epsilon-insensitive loss, 275, 308-310, 383
- Empirical risk minimization (ERM), *see* Inductive principle
- Entropy function, 186
- Explanation vs. prediction, 3-5
- Explanation bias, 415
- Empirical knowledge, 10-11  
 vs. first principle knowledge, 10-11  
 vs. beliefs, 10-11, 24-25  
 mathematical representation, 15-16
- Experimental procedure, 19-20, 414
- Explanation vs. prediction, 3-5, 18
- Face detection, 425
- False negative/ False positive, 123, 306-307
- Falsifiability  
 Popper's falsifiability, 91-96  
 VC-falsifiability, 154
- Feature selection, 21, 196, 419-425  
 forward, 196  
 in signal processing methods, 202  
 filter methods, 421  
 wrapper methods, 421  
 in Viola-Jones face detector, 426-428
- Feature space aka model space or  $\mathbf{z}$ -space 243, 247. *See also* SOM
- Filter methods, 421. *See also* Feature selection
- First principle knowledge, 10-11
- Fisher linear discriminant, 434
- Formalization (of application problem), 416-419, 450
- Fuzzy logic, 105, 113
- Generalization 21, 55-56. *See also* Complexity control  
 VC generalization bounds, 137-140  
 of optimal separating hyperplane, 288-289
- Generalized Lloyd algorithm, 243-244, 447. *See also* Clustering  
 batch version, 244  
 flow-through version, 246
- Gini function, 379, 380
- Gradient descent, 218, 223
- Greedy optimization strategy, 176, 181, 196, 351
- Heavisine signal, 207-208
- Hebbian rule, 220
- Hidden unit, 225, 227, 234
- Histogram of projections (for SVM)  
 for classification, 303-305  
 for transduction, 379-380  
 for Universum SVM, 385-388  
 for cost-sensitive SVM, 440
- Human modeling bias, 415
- Hyperplane  
 separable, 279-280, 284-285  
 nonseparable or soft-margin, 290-291
- Idealism, 75-77
- Instrumentalism, 75-77
- Indicator function, 16, 133
- Inductive model, 18
- Inductive learning formulation, 32-33, 364
- Inductive principle, 25, 40, 42, 45-55  
 Empirical risk minimization, 86  
 Maximum likelihood, 87  
 Occam's Razor, 88  
 Popper's falsifiability, 91

- Bayesian inference, 96-98
- Bayesian averaging, 102
- Principle of multiple explanations, 102-105
- Structural risk minimization, 145
- VC falsifiability, 154, 318
- Inference, 24, 80
  - logical, 80-82
  - inductive, 83-84, 450, 452
  - statistical, 85
  - nonstandard approaches, 364-366
- Interpretation (of predictive models), 22, 330, 414, 435, 451
  - of logistic regression, 173
- Kernel function (aka inner product kernel), 272, 294, 319
  - inner product, 293-294
  - properties, 294, 323
  - polynomial, 295
  - radial basis function, 295, 312
  - kernel complexity parameter, 297
- Kernel methods, *See* Support vector machines
- Knowledge, 1-5. *See also* First principle knowledge
  - historical perspective, 5-7
  - empirical, 10
  - vs. beliefs, 11
  - acquisition of, 80-81
- Lagrange coefficients, 287
- Lagrangian, 287
- Learning, 2
  - as function estimation, 15-18, 26
  - induction-deduction process, 32-33
  - goals, 38-39
  - the sign of a function, 45
- Learning methods aka learning algorithms
  - basic approaches, 47-54
  - statistical, 161-162
  - neural network, 215-216
- Learning problem setting, 364-367
  - standard inductive, 364-365
  - nonstandard, 365-367
  - philosophical interpretation, 403-404, 442-443
- Learning rate, 219, 222, 241, 248, 251
- Learning Using Privileged Information (LUPI), 389-392
- Leave-one-out cross-validation, 58
- Likelihood, 86, 97, 174
- Linear regression, 163-166
- Linearly separable, 284
- Load profile, 443
- Local estimation aka k-nearest neighbor methods, 50
- Logistic regression 169-173
- Loss functions, 37-38. *See also* MSE loss
  - for classification, 38, 186
  - for regression, 37
  - for unsupervised learning, 44, 240
  - in VC-theoretical problem setting, 121
  - negative log-likelihood, 174
  - epsilon-insensitive, 275, 308
  - exponential, 351. *See also* AdaBoost
- Margin, 135, 374, 291. *See also* Support Vector Machine
- Market timing of mutual funds, 432-434
  - daily loss function, 433
  - interpretation of trading models, 435
- Maximum a posteriori probability (MAP) 99
- Misclassification costs, 123, 306-307, 418
- Minimum Description Length (MDL), 54
- Model selection, 56. *See also* Complexity control
  - analytic approach, 56-57
  - resampling approach, 57-58
  - example, 62

- VC-based, 143
  - for signal denoising, 204
  - for SVM classification, 297, 302
  - for SVM regression, 314-315
- Multilayer perceptron (MLP), 216, 219, 226
  - for dimensionality reduction, 263
  - for regression, 228
  - for classification, 229-230
  - generalization properties of MLP, 227
- Multiple model estimation, 46
- Multi-task learning (MTL), 396-398
  - SVM+MTL, 399-400
- Network growing algorithms, 131
- Neural networks, 23, 215. *See also* MLP, RBF, SOM
  - terminology, 216-217
  - history, 220
- Nonlinear methods aka adaptive methods, 175
  - taxonomy, 176-177
  - for classification, 177-178
- Occam's razor, 88-91
- Outliers, 21, 35
- Overfitting, 229, 239, 356. *See also* Model selection
- Parameter estimation, 48
- Penalization, 317-318. *See also* Ridge regression
- Penalization structure, 148
- Philosophy of science, 75, 442-443
- Polynomial regression model, 141
- Posterior probability, 97-98. *See also* MAP
  - for classification problem, 123
- Predictive estimator, 40-41
- Predictive learning methodology, 416-417
  - philosophy, 449-450
- Preprocessing, 21, 34-35
  - for MLP network, 227
  - for RBF network, 236
  - for SVM, 301
- Probabilistic modeling, 84, 122-124, 162
  - assumptions, 211
  - example, 125-128
- Predictive modeling (aka system imitation), 84, 122-124, 162
  - assumptions, 211
  - example, 125-128
- Prior probability, 97, 122, 125
- Probability, 6-7. *See also* Uncertainty
- Projection pursuit, 194-195
- Prototype vectors 43, 241-243
- Pruning, 184-185. *See also* CART
- Radial basis function networks, 234
  - training algorithm, 235-236
  - model complexity, 237
- Realism, 75-77
- Recycling, 226
- Risk minimization approach, 8-9. *See also* System imitation and VC-theory, 9
- Regression, 37, 39
  - nearest neighbor, 50-52
  - statistical methods, 162-164
  - taxonomy, 175-177
  - multiple-response, 178-179
  - neural network methods, 217
  - via SVM, 274-276, 308-310
  - via combining methods, 329
  - modeling energy consumption, 444
- Regularization, *see* Penalization, Model selection
- Regularization parameter, 167, 290, 317-318
- Regularization setting (for SVM regression), 309
- Reflexivity, 109-111
- Resampling, 57-59, 63. *See also* Model selection
- Ridge regression, 166-167
- Risk functional, 38-39, 129-130

- Scaling of data, *see* Preprocessing
- Self-serving data analysis, 25
- Self-organizing map (SOM), 247-249
  - neighborhood function, 248, 250
  - user-defined parameters, 250
  - for vector quantization, 255
  - clustering of European languages, 256
  - for supervised learning, 263
- Self-learning algorithm, 376-380
- Semi-supervised learning, 374, 376
  - assumptions, 378
- Separating hyperplane, 284
  - optimal, 285, 289
- Sigmoid (or logistic) function, 171, 217, 224, 230
- Shattering, 131
- Signal denoising, 202, 213
  - signal processing formulation, 203
  - procedure, 204
  - VC-based, 207
- Single class learning, 282
- Slack variables, 290
- Social systems and philosophy, 106-112
- Soft margin hyperplane, 291
- Statistical methodology, 162
- Stacking predictors, 333
- Statistical learning theory, *see* VC-theory
- Structural risk minimization (SRM), 145
- Structure, 145
  - dictionary structure, 147
  - penalization structure, 148
  - feature selection structure, 148
  - SVM structure, 281
  - margin-based, 309
- Support vector machine, 270. *See also* Margin
  - methodology, 272
  - linear SVM, 284
  - nonlinear SVM, 292
  - generalization bounds, 288-289
  - model selection, 297, 302, 314
  - cost-sensitive SVM 306, 307
  - regression, 308
  - SVM vs. penalization, 317
  - SVM+ or SVM-Plus, 391-394
  - tuning parameters, 394
- Support vectors 286-287, 292
- System identification vs. system imitation, 9
- Transductive inference, 373-374
- Transduction, 367-374
- Transplant-related mortality, 437-439
  - prediction accuracy, 441
- Unbalanced data, 307
- Uncertainty 5-7
- Units, *see* Prototype vectors
- Universum learning, 381-383
  - random averaging, 384
  - model selection, 384
  - histogram of projections, 385-388
  - and human culture, 389
- Unsupervised learning, 43, 240. *See also* Clustering, Dimensionality reduction
- VC-dimension, 131-132
  - for consistency of ERM, 132
  - and Popper's falsifiability, 132-133
  - examples of calculating, 133-136
  - for regression problem, 137
- VC theory, 118-119
  - philosophical and conceptual implications, 153-154
- Vector quantization, 242-244. *See also* Clustering
- Viola-Jones face detector, 426-427
  - computational aspects, 428-430
  - cascaded classifier, 430-431
- Wavelets, 205. *See also* Signal denoising
- Wrapper methods, 421
- Ying-Yang principle, 108