

## Index

- %GOFLOGIT SAS macro
  - for goodness-of-fit 12, 103
- 632+ method 207
- accuracy 95, 114, 117, 134, 215
- Agresti, Alan 10
- Akaike information criterion 22
- analysis of covariance 170
- analysis of variance 10, 124
- artificial neural networks 53
  
- backward stepwise elimination
  - 181, 182
- balancing 82
- Barthel Index 62
- base classifier 54
- Bayes classifiers 45
- Bayes consistency 31, 32, 33, 34, 50, 152, 253, 256
- Bayes error 222
- Bayes optimal decision rule 38
- Bayes risk 32
- Bayesian information criterion 22
- Bayesian model averaging 251
- beer consumption example 176
- benchmarking 32, 56
- bias 109, 220, 226, 233
- bimodal 183
- Boltzmann learning machines 29
- BoosTexter 77
- boosting 12, 34, 38, 54, 68, 70, 74, 84
  
- bootstrap resampling 65, 112, 134, 137, 198, 202, 218, 227
- Breiman, Leo 10, 137
  
- C language 12
- calibration 106, 107, 112
- cardiac functioning example 173
- case-to-case distance 211, 213
- cell cultures experiment 207
- cholesterol study example 11, 58, 66, 94, 107, 165, 228
- classification 14, 50, 113, 240
- classification tree 22
- classifier 225, 251
- clustering 17, 143, 261
- coefficient of determination 177, 193
- coefficients 93, 101, 180
- committee 34, 117, 249, 250
- composite clinical score 9
- conditional inference forest 10, 13, 154
- conditional probability 45, 46
- confidence interval 11, 108, 215, 234
- confusion matrix 217
- constraints 21
- continuous outcomes 9, 36, 239
- convergence 101
- correlation 10, 44, 101, 115, 136, 157, 165, 168, 177, 201, 216, 234, 248
- correlation coefficient 158, 161

## 282 Index

- crop yield 174
- cross-validation 112, 198, 222, 224
- cubic histogram 34
- Cutler, Adele 10, 137
- data integrity 231
- data mining 15, 185
- decision boundary 52, 59, 66, 87, 98, 102, 125, 133
- decision rule 94, 95, 102, 129, 132
- decision trees 10, 29, 67, 84, 112, 118, 120
- dependent variable 16
- Devroye, Luc (DGL) 14
- discriminant analysis 37
- discriminant function 43
- distance function 143
- distance measure 49
- down-sampling 142, 219
- dropping down a tree 118
- empirical risk minimization 225
- ensemble methods 57, 117, 247, 249
- error estimation 217, 219, 221, 226, 232
- error rate 64, 67, 70, 71, 74, 76, 94, 113, 141, 198, 223, 248
- Euclidean distance 49, 143, 211
- evolutionary algorithm 55
- evolutionary computing 29
- false negative 67
- false positive 67
- Farrington statistic 104
- features (variables) 16, 47, 62, 70, 74, 112, 114, 120, 132, 137, 153, 212
- functional data estimation 203
- gene expression transcripts 17
- generalized additive model 44, 53
- genetic algorithm 55
- Gini concentration measure 105
- Gini error 140
- Gini index 124, 133
- goodness-of-fit 92, 103, 116
- group prediction 43
- hidden layer 53
- high-order correlation 8
- Hopfield neural networks 29
- Hosmer–Lemeshow test 104
- Hothorn, Torsten 13
- image classification 260
- imbalanced data 63, 205
- important features (variables) 7
- independence 229, 248
- independent variable 16
- interactions 7, 92, 95, 99, 132
- interpretable model 7
- k*-nearest neighbor 22, 34, 48, 68, 72, 76, 83, 220
- kernel functions 69
- kernel methods 29, 50, 51
- knowledge discovery 15
- Kruskal–Wallis 189
- König, Inke 11
- lasso 21
- layered nearest neighbor 152
- learning machine 14, 60, 70, 115, 130, 137, 224, 234, 249
- leave-one-out method 198, 226, 241
- likelihood function 22, 100
- linear discriminant 31, 43, 66, 80, 83, 115, 228, 242
- linear model 53, 56
- linear regression 4, 22, 37, 41, 53, 157
- linear support vector machine 51
- local 35, 49, 52, 67, 73, 128, 162, 256

**283**      Index

- logic regression 16, 47, 51
- logistic regression 4, 11, 25, 27, 42, 66, 91, 106, 114
- logitboost 27, 38, 55
- longitudinal data 260
- lupus (SLE) study example 9, 11, 61, 70, 79, 81, 112, 129
  
- machine learning 15
- Mahalanobis distance 242
- matched case-control
- MatLab 12
- mean squared error 240
- merging models 186
- misclassification 67, 97, 141, 205, 237
- missing data 9
- model averaging 196
- model richness 22, 53, 128
- Mojirsheibani, Majid 252
- Monte Carlo 87, 213, 231
- multidimensional scaling (MDS) 18, 143
- multilevel models 170
- multiple models 186
- multiple response permutation procedures (MRPP) 210
- m*XCV cross-validation 223
  
- naïve Bayes machine 26, 46, 201
- nearest neighbors 34
- neighborhood 49, 52, 256
- neural networks 29, 39, 53, 68, 71, 76, 83
- Newcombe, Robert 11
- node 120, 140
- node purity 133
- nonlinear model 53
- nonparametric 16, 22
- nonparametric density estimation 203
- null hypothesis 209, 214
  
- Occam's Razor 187
- optimization 79, 69
- out-of-bag (OOB) sampling 65, 71, 79, 134, 137, 218, 227, 232
- outcome 14, 63
- outliers 44, 159, 166
- overall error rate 64, 65, 217
- overfitting 21, 54, 128, 130
- oversampling 206
  
- parameter 92
- parametric bootstrap methods 203
- partial correlation 175
- pattern recognition 15
- Pearson statistic 104
- penalized maximum likelihood 112
- performance 65, 74, 80, 215, 228
- permutation procedure for testing group differences 210
- permutation tests 207
- pooling lists 186
- pooling machines 251, 252
- population incidence 219
- precision 64, 217
- prediction 14, 54, 58, 92, 118, 222, 236, 256
- prediction engine 4, 15, 117, 213, 225, 248
- predictive accuracy 11
- predictors 16, 43, 47, 61, 62, 63, 70, 143, 178, 247
- principal components analysis 19
- priors 45
- probabilistic learning theory 5
- probabilistic pattern recognition 49
- probability 92, 248
- probability estimates 102
- probability function 22
- probability of group membership 43, 45, 92, 103, 106, 113, 128

## 284 Index

- Proc IML 12
- Proc Logistic 12
- proximity 19
- pruning 128, 138
- purity 122, 123
  
- quadratic discriminant function 44
  
- R language 12
- R-squared 105, 177, 193
- Random Forests 10, 12, 26, 34, 50, 62, 63, 75, 77, 112, , 116, 137, 152
- Random Jungles 10, 13, 34, 50, 112, 116, 152, 153, 214
- randomization 223
- ranking features (variables) 86
- rate of convergence 31
- ratio sampling 219
- receiver operator curves 103
- reference model 113
- regression 4, 239
- regression tree 163
- repeated measures 260
- resampling 111, 112, 198, 218, 225, 227
- robustness of the inference 169
- ROC curves 236
- rotation method 240
  
- sample size 8, 136, 142, 204, 251
- sample variance 124, 223
- sampling with replacement 199, 218
- SAS 12
- Schwarz, Daniel 13
- sensitivity 64, 68, 81, 84, 94, 114, 141, 205, 217, 231, 236
- sequential scheme 34
- shrinkage methods 112
- Simpson's paradox 170
  
- simulated annealing 29
- simulated data 8
- simulated study 58
- single nucleotide polymorphisms (SNPs) 7, 17, 46, 136, 231
- small sample sizes 206
- software 12
- sparse 104
- specificity 64, 82, 84, 94, 114, 205, 217, 231, 236
- spinal curvature example 172, 184
- splitting 120
- splitting to purity 35, 152
- stacking 117
- statistically balanced blocks 34
- stepwise regression 181
- stopping rules 127
- stroke study example 11, 62, 75, 79, 103, 144, 234
- structural risk minimization 225
- subsets 9
- super classifiers 32, 57, 79
- supervised learning 17, 144
- support vector machine (SVM) 11, 16, 33, 44, 50, 69, 83
- support vectors 52
- survival 15, 62
- survival forest 259
- survival machines 259
- SVMLight 12
- systemic lupus erythematosus (SLE) 61
  
- Tango, Toshiro 11, 257
- terminal node 119, 130, 143, 152, 253
- testing 60
- thresholds 58, 95, 237
- total balancing 219
- training 60, 77, 100, 121, 134, 224, 226, 229

**285**      Index

---

- transfer function 53
- transition node 119
- tree complexity 128, 138
- tuning 70, 141, 257
- two-fold cross-validation (2XCV) 222
  
- unbalanced 63, 82, 94
- unbalanced bootstrap sampling 205
- unbalanced data 205, 218,  
231, 233
- unbalanced groups 8
- under-sampling 142, 219
- unequal group sizes 212, 218
  
- unequal variances 101
- unsupervised learning 17, 37, 144
  
- validation 6, 149
- Vapnik–Chernovenkis dimension  
23, 54, 128
- variable importance 116, 132
- variable permutation 134
- variable selection 102
- vitamin E experiment 207
  
- Ward, Michael 11
- Wilcoxon rank sum test 237