Index

activation function, 277
adjacency matrix, 532, 561
Akaike information criterion, 53, 583
application studies
classifier combination, 399
clustering, 546
data fusion, 399
decision trees, 340, 350
feature selection and extraction, 493
MARS, 357
mixture models, 64, 547
neural networks, 548
nonparametric methods of density estimation, 213
normal-based linear and quadratic discriminant rule, 63
rule induction, 357
support vector machines, 315
approximate Bayesian computation, 137–144
back-propagation, 302–305
bagging, 385, 402
ball trees, 170
Bayes
decision rule, see decision rule
error, see error, Bayes
Bayesian information criterion, 53
Bayesian learning methods, 70–87
analytic, 73–87
conjugacy, 73
recursive, 72
Bayesian multinet, 189
Bayesian networks, 186–190
Bayesian sampling schemes, 87–126
importance sampling, 92–95
Markov chain Monte Carlo, 95–116
ratio of uniforms, 90
rejection sampling, 89
sequential Monte Carlo samplers, 119–126
boosting, 387
bootstrap
in cluster validity, 546
in error rate estimation, see error-rate estimation, bootstrap
Borda count, 379
branch and bound
in clustering, 516, 520
in density estimation, 203
in feature selection, 450–454, 496
in kd-trees, 163
in nearest-neighbour classification, 163–174
CART, see decision trees, CART
chain rule, 187
change of variables, 91
class-conditional probability density function, 7
classification trees, see decision trees
classifier combination methods, see ensemble methods
classifier fusion, 362
classifier generalisation, 6
classifier performance assessment, 404–432
clique percolation method, 573
clustering
agglomerative methods, 203
application studies, 546
cluster validity, 538–546
hierarchical methods, 502–510
agglomerative algorithm, 503
complete-link method, 506–507
divisive algorithm, 503, 510
general agglomerative algorithm, 508
inversions, 510, 547
nonuniqueness, 510, 547
single-link method, 503–506
sum-of-squares method, 507
mixture models, 511–513
quick partitions, 510–511
spectral clustering, 531–538
sum-of-squares methods, 513–531
clustering criteria, 514
complete search, 520
fuzzy k-means, 519–520
k-means, 49, 280, 516–517
nonlinear optimisation, 518
stochastic vector quantisation, 529–531
vector quantisation, 520–531
community detection, see complex networks, community detection
comparative studies, 30
approximation elimination search algorithms, 178
comparing performance, 424
decision trees, 357, 358
feature selection, 497
fuzzy c-means, 547
hierarchical clustering methods, 547
kernel bandwidth estimators, 200
kernel methods, 201, 203, 216
linear discriminant analysis – for small sample size, 248
MARS, 358
maximum weight dependence trees, 186
naive Bayes, 216
neural networks, 315
neural networks model selection, 582
nonlinear feature extraction, 493
nonlinear optimisation algorithms, 313
normal-based models, 64
number of mixture components, 53
principal components analysis, 469, 473
RBF learning, 290
regularised discriminant analysis, 45, 46
tree-pruning methods, 339
complete link, see clustering, hierarchical methods, complete-link method
complex networks, 555–580
community detection, 565–575
link prediction, 575–579
properties, 557
condensing of nearest-neighbour training set, 176
conditional risk, 14
confusion matrix, 5, 40, 406
conjugate gradients, 305
conjugate priors, 73
copulas, see nonparametric discrimination, copulas
covariance matrix
maximum likelihood estimate, 37
structures
common principal components, 45
proportional, 45
unbiased estimate, 80
cross-validation, 327, 582
error rate, see error-rate estimation, cross-validation
curse of dimensionality, 181
data
dimension, 4
head injury patient, 39
training set, 3
data fusion, 361, 366, 367, 370–376
data mining, 2, 315
data visualisation, 433, 463
INDEX 639

decision rule, 7
Bayes
  for minimum error, 8–15, 18
  for minimum risk, 13–15
minimax, 18
Neyman–Pearson, 15, 373
decision surfaces, 7
decision theory, 7–20
decision trees, 323–342
  CART, 330, 332, 334
  construction, 326, 337–339
  definition, 333
  pruning algorithm, 330–332
  splitting rules, 327–330
deduction, 3
dendrogram, 502
density estimation
  nonparametric, 150–218
    expansion by basis functions, 204
    properties, 150
  parametric
    estimative, 34
    predictive, 35, 70
  semiparametric, 203
design set, see data, training set
detailed balance, 97, 105
discriminability, see performance
  assessment, discriminability
discriminant functions, 20–26
discrimination
  normal-based models, 35–46
    linear discriminant function, 41
    quadratic discriminant function, 36
  regularised discriminant analysis, see regularised discriminant analysis
distance measures
  binary variables, 443
distance between distributions
  Bhattacharyya, 452
  Chernoff, 452
divergence, 446, 452
  Kullback–Leibler, 126, 184
  Mahalanobis, 271, 452, 500
  multiclass measures, 447
  Patrick–Fischer, 452
  nominal and ordinal variables, 443
divergence, 446, 452
editing of nearest-neighbour training set, 174
EM algorithm, see estimation, maximum
  likelihood, EM algorithm
ensemble methods, 361–402
error rate, 406
  apparent, 406, 448
  Bayes, 10, 407
  estimation, see error-rate estimation
  expected, 406
  for feature selection, 448
  true, 406
error-rate estimation
  bootstrap, 411–413, 448
  cross-validation, 408
  holdout, 407
  jackknife, 410, 448
errors-in-variables models, 289
estimation
  maximum likelihood
    definition, 34
    EM algorithm, 55
Expectation Maximisation algorithm, 49
false negative rate, 16
false positive rate, 16
feature extraction, 433, 434, 463–495
  Karhunen–Loève transformation, 475–481
  Kittler–Young, 477
  SELFIC, 476
  multidimensional scaling, 484–493
    by transformation, 491
    classical scaling, 484
    metric multidimensional scaling, 486
    ordinal scaling, 487
principal components analysis, 98, 464–474
feature selection, 433, 435–463
algorithms, 449–463
  branch and bound, 450
  Markov blanket, 459
  suboptimal methods, 454–463
criteria, 440
  error rate, 448
  probabilistic distance, 446
  scatter matrices, 443
embedded methods, 439
INDEX

feature selection (Continued)
filter methods, 439
floating search methods, 457
redundancy, 436
relevance, 435
stability, 460
wrapper methods, 439
features, 2
forward propagation, 303
fuzzy $k$-means clustering, see clustering, sum-of-squares methods, fuzzy $k$-means
Gaussian classifier, see discrimination, normal-based models, 35–40
geometric methods, 433
Gibbs sampling, see Markov chain Monte Carlo algorithms, Gibbs sampling
Gini criterion, 328
graph, 188, 532–533, 561
graph centrality measures, 563–564
graph diameter, 563
Hermite polynomial, 206
imprecision, see performance assessment, imprecision
independent components analysis, 483
induction, 3
information gain, 329
intrinsic dimensionality, 4, 285
$k$-means clustering, see clustering, sum-of-squares methods, $k$-means
$k$-nearest-neighbour methods, see nonparametric discrimination, nearest-neighbour methods
Karhunen–Loève transformation, see feature extraction, Karhunen–Loève transformation
Karush–Kuhn–Tucker conditions, 252
kernel methods, 88, 194
Kullback–Leibler divergence, 126, 128, 129, 184, 438, 446, 460
Laplacian, 532, 535–536, 562
latent variables, 483
LBG algorithm, 522
learning vector quantisation, 528
likelihood ratio, 9
linear discriminant analysis, 221–249
error correction procedure
  multiclass, 237
two-class, 223
Fisher’s criterion
  multiclass, 238
two-class, 227
for feature extraction, see feature extraction, linear discriminant analysis
least mean-squared-error procedures, 228, 241–246
multiclass algorithms, 236–249
perceptron criterion, 223–227
support vector machines, see support vector machines
two-class algorithms, 222–236
linear discriminant function, see discrimination, normal models, normal-based linear discriminant function
generalised, 24
piecewise, 23
link prediction, see complex networks, link prediction
logistic discrimination, 263–268
loss matrix, 14
equal cost, 14
Luttrell algorithm, 528
machine learning, 322
Mahalanobis distance, see distance measures, distance between distributions, Mahalanobis
Markov blanket, 436, 459
Markov chain Monte Carlo methods, 95–116
Gibbs sampling, 95–103
Metropolis–Hastings, 103–107, 114
practical example, 111–114
reversible jump Markov chain Monte Carlo, 108
slice sampling, 109–111
MARS, see multivariate adaptive regression splines
MCMC algorithms, see Markov chain
Monte Carlo algorithms
Mercer’s condition, 293
Metropolis–Hasting, see Markov chain
Monte Carlo algorithms,
Metropolis–Hasting
minimum description length, 339, 345, 350,
462, 584
minimum-distance classifier, 22, 241
misclassification matrix, see confusion
matrix
missing data, 585–586
mixture models, 46–63
discriminant analysis, 48
EM algorithm, 55
in cluster analysis, see clustering, mixture
models
in discriminant analysis, see normal
mixture models
number of components, 52
sampling from, 48
mixture of experts model, 382–385
mixture sampling, 264
model averaging, 390–396
model selection, 6, 53, 135, 286, 293,
581–585
modularity, 569, 570–572
monotonicity property, 450
multidimensional scaling, see feature
extraction, multidimensional
scaling
multilayer perceptron, 26, 275, 298–314
multivariate adaptive regression splines,
351–356
mutual information, 441
naïve Bayes, 40, 181
nearest class mean classifier, 23, 42
neural networks, 274–321
model selection, 582
nonparametric discrimination
copulas, 207–213, 216
expansion by basis functions,
204
histogram approximations, 181
Bayesian networks, 186–190
independence, 181
Lancaster models, 182
maximum weight dependence trees,
183–190
histogram method, 180, 216
variable cell, 181
kernel methods, 194–204, 216
choice of kernel, 201
choice of smoothing parameter, 199
multivariate kernel, 198
product kernels, 198
variable kernel, 200
nearest-neighbour algorithms, 159–174
ball-trees, 170–174
kd-trees, 163–170
LAESA, 159–163
nearest-neighbour methods, 152–180, 216
choice of $k$, 180
choice of metric, 157
condensing, 176
discriminant adaptive nearest
neighbour classification, 177
editing techniques, 174
for RBF initialisation, 281
$k$-nearest-neighbour decision rule, 156
normal (Gaussian) distribution, 36
normal mixture models, 47, 67
cluster analysis, 511
discriminant analysis, 48
EM acceleration, 62
EM algorithm, 49, 55, 67
for RBF initialisation, 280
number of components, 52
normal-based quadratic discriminant
function, 36
oblique decision trees, 341
ordination methods, 433
outlier detection, 586
overfitting, 6
parametric density estimation, see density
estimation, parametric
particle filters, 93
pattern
definition, 2
perceptron, see linear discriminant analysis,
perceptron criterion
INDEX

642

performance assessment
  discriminability, 405, 406
  imprecision, 413
  reliability, 405, 413
population and sensor drift, 7, 202, 245, 289, 317, 419–421
primary monotone condition, 488
principal components analysis, see feature extraction, principal components analysis
principal coordinates analysis, 484
probabilistic expert systems, 189
probability
  a posteriori, 8
  a priori, 8
probability density function
  Dirichlet, 86
  gamma, 74
  normal, 36
  Poisson, 74
  student, 81
  Wishart, 80
projection pursuit, 26, 313
pseudo-distances, 489
quadratic discriminant function, see discrimination, normal models, quadratic discriminant function
radial basis function network, 26, 269, 275–290
random forests, 389
random graph, 564
Rao-Blackwellisation, 88
receiver operating characteristic, 17, 415–422
  area under the curve, 416–419
recursive partitioning, 322
regression, 21, 26–28
regularisation, 246, 284
regularised discriminant analysis, 42–67
reject option, 7, 12, 15
reliability, see performance assessment, reliability
representation space, 484
robust procedures, 586
ROC, see receiver operating characteristic rule induction, 342–351
scale-free networks, 559
scree test, 469, 485
secondary monotone condition, 488
self-organising feature maps, 525–531
Sherman–Morrison formula, 409
simulated annealing
  in clustering, 520
single-link, see clustering, hierarchical methods, single-link method
small-world effect, 557
softmax, 284
spectral clustering, 531–538
stacked generalisation, 382
stochastic vector quantisation, see clustering, sum-of-squares methods, stochastic vector quantisation
stress, 489
support vector machines, 249–263, 291–298, 388
  application studies, 268, 315
  canonical hyperplanes, 250
  linear
    multiclass algorithms, 247, 256
    two-class algorithms, 249–256, 262
  nonlinear, 291–298
  surrogate splits, 339
  Sylvester’s determinant theorem, 84
text classification, 190
training set, see data, training set, 7
tree-structured vector quantisation
ultrametric dissimilarity coefficient, 545
ultrametric inequality, 503
validation set, 582
variables of mixed type, 587–588
variational Bayes, 126–137
varimax rotation, 473
VC dimension, 589, 590
vector quantisation, see clustering, sum-of-squares methods, vector quantisation