Index

A

accuracy, 162
  balanced accuracy, 164
activation function, 403, 418, 424
Adam (optimization method), 486
ADHD, 12, 377
application
  business and industrial applications, 10
  cancer classification, 173, 398
  cat versus dog classification, 1
  credit risk assessment, 167
  diagnosis of medical conditions, 12
  financial applications, 13
  genome-wide association, 377
  handwritten digit recognition, 204, 247
  object detection, 11
  object recognition, 10, 14
  predicting Automobile Miles-per-Gallon, 113
  predicting box office success, 9
  predicting house prices, 113
  sentiment analysis, 11, 242
  spam detection, 13, 243, 461
  student loan prediction, 8
ascent direction, 35
autoencoder
  linear, 215
  nonlinear, 294, 304, 403, 423
autograd, 55, 56, 89, 536
automatic differentiation, 55, 520, 526
  forward mode, 520
  reverse mode, 526
average
  cumulative average, 474
  exponential average, 474

B

backpropagation, 427, 526
backtracking line search, 493
Bag of Words (BoW), 240
bagging
  classification, 369
  regression, 367
batch normalization, 430
batch size, 489
bias weight, 100
biological neural networks, 418, 424
boosting
  cross-validation, 340
  feature selection, 258
Broyden–Fletcher–Goldfarb–Shanno (BFGS)
    method, 95, 509

C

capacity dial, 320, 328
categorical Cross Entropy cost, 158, 193
child node, 517
classification
  nonlinear, 286, 290, 304, 403, 443
  quality metrics, 160
  weighted classification, 167
classifier, 5
clustering, 14, 15
complexity dial, 306
computation graph, 517
confusion matrix, 165
constrained optimization, 156
contour plot, 29
convergence, 32, 63, 85, 497
convexity, 51, 57, 78, 81, 91, 103, 501
coordinate descent, 39, 50
coordinate search, 39
correlation value, 241
cost
  Cross Entropy, 133
  Least Absolute Deviations, 108
  Least Squares, 103
  Perceptron, 140
  Softmax, 135
cost function, 16
cost function history plot, 34, 61
Cross Entropy cost, 125
cross-validation

  boosting based, 340
  K-fold, 373
  leave-one-out, 375
  naive, 335
  regularization based, 350
curse of dimensionality, 26, 37
curvature, 75, 491

D

dataset
  Auto-MPG data, 113
  bacterial growth data, 298
  Boston Housing data, 113, 261, 267
  breast cancer data, 173, 398
  German credit data, 167, 261, 267
  Iris dataset, 202
  Kleiber’s law data, 9
  MNIST dataset, 204, 247, 430, 438
  Moore’s law data, 299
  Ohm’s law data, 300
  spam detection data, 166, 461
  student loan data, 8
Davidon–Fletcher–Powell (DFP) method, 509
decision boundary, 126
derivative, 511
descent direction, 29
diagonal matrix, 233, 257
differentiation
  automatic, 520
  numerical, 513
  symbolic, 525
differentiation rules, 56
dimension reduction, 14
discrete probability distribution, 193

E

early stopping
  boosting, 346
  regularization, 353
edge detection, 243
eigenvalue, 75, 77, 79
ensembling, 373, 446
epoch, 489
exact line search, 495
exponential average, 473

F

face detection, 11, 125
feature
  feature engineering, 2, 238, 275
  feature learning, 3, 304, 378
  feature scaling, 249
  feature selection, 113, 258, 264
  histogram features, 238
feature design, 2
feature selection, 113
first-order optimization, 45
first-order system of equations, 46
fMRI, 12, 377
Fourier basis, 384
Fourier kernel, 392
fully connected networks, 403
function approximation, 312
fundamental theorem of linear algebra, 387
fusion rule, 181, 184, 291

G

Galileo, 279
  ramp experiment, 114
genome-wide association study (GWAS), 377
global optimization, 24
gradient, 516
gradient boosting, 346, 458
gradient computation, 56, 536
gradient descent
  component-wise normalization, 481
  convergence criteria, 63
  gradient descent algorithm, 56, 63, 87
  illustration, 57
  mini-batch gradient descent, 203
  momentum acceleration, 473
  normalized gradient descent, 478
  Python implementation, 63
  slow-crawling behavior, 69, 478
  weaknesses, 65
  zig-zagging behavior, 65, 69, 473
greedy algorithm, 262

H

Hadamard product, 549
handwritten digit recognition, 10, 247
hard-margin SVMs, 155
harmonic series, 37
Hessian matrix, 91, 134, 495
Hessian-free optimization
  quasi-Newton, 503
  secant method, 504
  subsampling, 502
hidden layer, 406

hinge cost, 143
histogram features
  for audio data, 248
  for categorical data, 239
  for image data, 243
  for text data, 240
human interpretability, 258, 348, 366, 373, 446
hyperbolic tangent function, 136
hyperplane, 52

I
image compression, 243
imbalanced data, 164, 167
imputation, 254
information gain, 456
inner-product rule, 54

J
JAX, 56, 536

K
K-fold cross-validation, 373
K-means, 221
kernel
  cross-validation, 398
  Fourier kernel, 392
  kernels as measures of similarity, 396
  optimization, 397
  polynomial kernel, 391
  radial basis function (RBF) kernel, 394
kernel trick, 386
Kernelized models
  multi-class classification, 390
  regression, 388
  two-class classification, 389
Kleiber's law, 9, 121

L
learning rate, 31
Least Absolute Deviations, 108
Least Squares cost function, 101
  minimization, 103
  susceptibility to outliers, 108
leave-one-out cross-validation, 375
line search, 88
  backtracking, 493
  exact, 495
linear regression
  dealing with duplicates, 114
  Least Absolute Deviations cost, 108
  Least Squares cost, 101
  notation and modeling, 99
  Python implementation, 105
  root mean squared error, 111
linear separation, 141
linear two-class classification
  notation and modeling, 126
Lipschitz constant, 495
local optimization, 27
log-loss SVM, 157
logistic regression, 131
logistic sigmoid function, 130
low-memory quasi-Newton methods, 509

M
machine precision, 87
majority vote, 454
Manhattan distance, 560
manifold, 10, 15, 423
margin, 151
Margin-Perceptron, 151
matrix
  arithmetic, 554
  norms, 562
matrix factorization problem, 227
maximum
  global, 21
  local, 23
maximum margin classifier, 153
maxout activation, 427
mean imputation, 254
Mean Squared Error (MSE), 112
median (statistics), 367
metrics
  accuracy, 198
  mean squared error, 112
mini-batch optimization, 203, 487
minimum
  global, 21
  local, 23, 46
missing data, 254
mode (statistics), 453
momentum acceleration, 473
Moore's law, 299
multi-class classification, 10
multi-output regression, 282
  cost functions, 118
  notation and modeling, 116
  Python implementation, 119

N
naive cross-validation, 335
  weaknesses, 339
neural networks, 317, 403
  activation functions, 424
  backpropagation, 427
  batch normalization, 430
  biological perspective, 418
  compact representation, 407
  cross-Validation via early stopping, 438
  graphical representation, 419
  multi-hidden-layer, 413
  nonconvexity, 428
  optimization, 428
  Python implementation, 418
  single-hidden-layer, 403
  two-hidden-layer, 408
neuron, 419
Newton’s method
  algorithm, 81
  comparison to gradient descent, 86
  connection to gradient descent, 84
  convergence, 85
  illustration, 85
  numerical stability, 87
  Python implementation, 89
  scaling limitations, 90
  zero-finding perspective, 88
nonlinear autoencoder, 294
nonlinear multi-output regression, 282
nonlinear regression, 275
nonnegative matrix factorization, 230
normalized exponential function, 194
normalized gradient descent, 478
norms (vector/matrix), 559
numerical differentiation, 513

O
object detection, 11
object recognition, 14
Ohm’s law, 300
one-hot encoding, 193, 239
One-versus-All
  algorithm, 182
  notation and modeling, 174
online learning, 203
optimality condition
  first-order, 45, 46
  second-order, 75
  zero-order, 23
optimization
  constrained, 155
  first-order, 45
  global, 24
  local, 27
  second-order, 75
  zero-order, 21
optimization dial, 320, 328
orthonormal basis, 211
outliers, 109
overfitting, 323, 333

P
parent node, 517
PCA-sphering, 255
perceptron, 140
Perceptron cost, 140
polynomial kernel, 391
positive (semi)definite, 75
Principal Component Analysis (PCA), 213
pruning, 464
pseudo-inverse, 84
purity, 456
Python (programming language), 56, 63, 89, 105, 119, 134, 138
Python implementation
  function flattening, 543
  gradient descent, 63
  linear regression, 105
  multi-class classification, 190
  multi-output regression, 119
  neural networks, 418
  Newton’s method, 89
  nonlinear regression, 281
  nonlinear two-class classification, 290
  PCA, 218

Q
quadratic function, 42, 51
quadratic geometry, 76
quality metrics
  classification, 160
  regression, 111
quantization, 114
quasi-Newton methods, 503, 505

R
radial basis function, 394
random forests, 462
random local search, 37
random search, 31
Rayleigh quotient, 48, 72, 235

recommender systems, 219
rectified linear unit, 143, 426
recursively built neural networks, 403
recursively built trees, 450
regression
  linear regression, 101
  multi-output regression, 116
  nonlinear, 275, 282, 304, 403, 443
  quality metrics, 111
  weighted, 114
regularization, 87, 500
  cross-validation, 350
  feature selection, 264
residual, 263, 349, 460
revenue forecasting, 9
RMSProp, 430, 487

S
saddle point, 47, 71, 76, 478
sampling
  random, 25
  uniform, 24
scree plot, 226
secant method, 504
second-order optimization, 75
sentiment analysis, 12, 242
signed distance, 195
similarity measure, 396
soft-margin SVMs, 155
Softmax cost, 135
spam detection, 13, 243
spanning set, 208, 310, 546
sparse coding, 230
spectrogram, 249
speech recognition, 10, 248
standard basis vector, 39
standard normalization, 249
stationary point, 47, 63, 76, 81, 478
steepest ascent, 52
steepest descent, 52
stemming, 241
steplength
  adjustable steplength rule, 35
  conservative steplength rules, 490
  diminishing steplength rule, 35, 36, 60
  fixed steplength rule, 31, 35, 60, 495
  steplength issues, 30, 60, 63
  steplength parameter, 31, 88, 490
stochastic optimization, 489
stop words, 241
stump (trees), 443
supervised learning, 7
  classification, 10
  regression, 7
support vector machines, 150
  hard-margin, 155
symmetric linear system, 50

T
Taylor series, 52, 79, 81, 500, 504, 531
testing error, 361
testing set, 364
time series, 474
training model, 3
training set, 2
trees, 318, 443
  classification trees, 452
  creating deep trees via addition, 445
  creating deep trees via recursion, 444
  cross-validation, 464
  gradient boosting, 458
  pruning, 466
  random forests, 462
  regression trees, 446
triangle inequality, 561

U
underfitting, 323, 333
uniform sampling, 24
universal approximation, 307
universal approximators
  fixed-shape, 316
  neural networks, 317
  trees, 318
unsupervised learning, 7

V
validation error, 323, 333
validation model, 3
validation set, 3
vanishing gradient problem, 425
vector
  arithmetic, 548
  norms, 559

W
weighted regression, 114
weighted two-class classification, 167
whitening, 255

Z

zero-order optimization, 21