Index

A

A/B testing, 148
  beta-binomial hierarchical model for, 149, 151
  simple two-sample, 149
Activation function, 3
Additive law of probability, 13
Akaike information criterion (AIC), 152
AlexNet, 110
Amazon Web Services (AWS), 167
Analysis of Variance (ANOVA), 137
  MANOVA, 138
  mixed-design, 138
  one-way, 137
  two-way (multiple-way), 137
Ant colony optimization (ACO), 159–160
Arithmetic mean, 15
Asset price prediction, 171–172
  description of experiment, 173–175
  feature selection, 175–176
  supervised learning, 172–173
Associative property, 19
Autoencoders, 125–126, 195–199, 201–202
  linear autoencoders *vs.* PCA, 126–127
Axioms, 19
  associative property, 19
  commutative property, 19
  distributivity of scalar multiplication, 20
  identity element of addition, 19
  identity element of scalar multiplication, 20
  inverse elements of addition, 19

B

Back-propagation algorithm, 95–97, 107
Back-propagation through time (BPPT), 114–115
Backward selection, 151
Batch learning, 131
Bayesian classifier, 191–193, 196
Bayesian learning, 83
  50/25/25 cross-validation, 85–86
  limitation, 84
  Naïve Bayes classifier, 84
  tuning machine learning algorithms, 85
Bayesian statistics, 149
Bayes information criterion (BIC), 152
Bayes' theorem, 14
Beta-binomial hierarchical model, 149, 151
Beta distribution, 150
Bi-infinite sequence, 40
Binary classifier, 66
Binomial distribution, 150, 151
Blocking process, 145
BPTT. *See* Back-propagation through time (BPTT)

C

Canonical correlation analysis (CCA), 156
Central processing unit (CPU), 168, 169
Coefficient of determination (R squared), 17
Collaborative filtering, 214–218
Commutative property, 19
Complex cells, 101
Confusion matrix, 68–69
  for Bayesian classifier, 85
  for classification tree, 80

T. Beysolow II, *Introduction to Deep Learning Using R*, DOI 10.1007/978-1-4842-2734-3

Conjugate distribution, 150
Conjugate gradient algorithms, 98, 119
Continuous random variables, 14
Contrasting divergence (CD) learning, 129–131
Convergent sequence, 41
Convolutional layer
  convolving, 104
  feature maps, 104
  filtering, 104
Convolutional neural networks (CNNs), 5, 108, 202, 204
  AlexNet, 110
  convolutional layer, 103–104
  depth, 108
  FC layer, 106
  GoogLeNet, 109
  history, 101
  loss layer, 107
  pooling layer, 105
  preprocessing, 204–206
  regularization, 111
  ReLU layer, 106
  ResNet, 110
  structure and properties, 101–103
  tuning parameters, 108
  VGGnet, 110
Convolving, 104
Cook's distance, 142
Correlation coefficients, 16
Correlation matrix, 175
Cosine similarity, 215

D

Data science, 219
Decision tree learning, 78
  classification trees, 79–80
  limitations, 81
  regression trees, 80–81
Deep belief network (DBN), 6, 134–135
Deep learning, 219
  autoencoders, 195–199, 201–202
  CNNs, 202, 204
    preprocessing, 204–206
  model building and training, 206–214
    collaborative filtering, 214–218
  models, 3
    applied machine learning and, 7
    CNNs, 5
    DBNs, 6
    experimental design, 7
    feature selection, 7
    history, 8
    MLP, 4
    restricted Boltzmann machines, 6
    RNNs, 5
    SLP, 3–4
    structure of, 2
Deep neural networks, 2
Derivatives and differentiability, 42
Diagonal matrix, 22
Discrete random variables, 14
Discriminant, 43
Distributional prediction, 79
DropConnect, 111
DropOut, 111

E

Eigenvalues, 34–36
Eigenvectors, 34–36
Elman neural networks, 115
Embedded algorithms,157. *See also* Wrappers, Filters, and Embedded (WFE) algorithms
Ensemble methods, 82
  gradient boosting algorithm, 82–83
  random forest, 83
    limitations, 83
Euclidean function
  Euclidean loss, 107
  softmax loss function, 107
  softmax normalization, 107
Euclidean norm, 29
Expectation maximization (EM) algorithm, 76
  expectation step, 77
  maximization step, 77–78

F

Factor analysis, 154–155
  limitations, 155
Factor loadings, 155
Fast learning algorithm, 135
  steps, 136
Feature maps, 104
Feature/variable selection techniques, 151
  backwards and forward selection, 151–152

factor analysis, 154–155
limitations, 155
PCA, 152–154
FeedForward() function, 206
Fisher's principles, 144–145
Fixed tabu search, 163–164
F-statistic and F-distribution, 138–145
Full factorial, 147
Fully connected (FC) layer, 102, 106
Fully recurrent networks, 113–114

G

Genetic algorithms (GAs), 158
Geometric mean, 15
Gibbs sampling, 129, 135
Global minimizers, 47–48
Global optimum, 95
Google Finance API, 172
GoogLeNet, 109
Gradient, 42
Gradient boosting algorithm, 82–83
Gradient descent algorithm, 53–54
Graphics processing unit (GPU), 168

H

Hadamard matrix, 146
Halton, Faure, and Sobol sequences, 148
Hamming distance, 164
Handling categorical data, 155
categorical label problems, 156
CCA, 156
encoding factor levels, 156
Hard drive disk (HDD), 167
Hardware and software suggestions
CPU, 169
GPU, 168
motherboard, 169
optimizing machine learning software, 170
processing data with standard hardware, 167
PSU, 170
RAM, 169
solid state drive and HDD, 167
Having memory, 113, 116
Hessian-free optimization, 48
Hessian matrix, 43, 49
Hidden layers, 99
Hill climbing search methods, 158

I

Identity matrix, 22
ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), 109–110
Inception architecture, 109
Instantaneous algorithm, 90
Intelligent optimization, 161
Interpretation, 145

J

Jacobian matrix, 49

K

Kernels, 72
K-means clustering, 74, 156
limitations, 75–76
K-nearest neighbors (KNN), 165, 189–191

L

Learning rate, 54–55, 209
choosing, 55–56, 58–60
Levenberg-Marquardt heuristic, 61
Newton's method, 60–61
Least Absolute Shrinkage and Selection Operator (LASSO), 63
ridge regression and, 64
LeNet, 108
Levenberg-Marquardt (LM) algorithm, 61
Leverage, 142
Linear algebra, 17
axioms, 19
associative property, 19
commutative property, 19
distributivity of scalar multiplication, 20
identity element of addition, 19
identity element of scalar multiplication, 20
inverse elements of addition, 19
matrices, 20
addition, 21
column vector and square matrix, 24
derivatives and differentiability, 42
distributive over matrix addition, 27

Linear algebra(*cont.*)
eigenvalues and eigenvectors, 34–36
Euclidean norm, 29
Hessian, 43
hyperplanes, 39–40
inner products, 32
L1 norm, 29–30
L2 norm, 29
limits, 41
linear transformations, 36–37
matrix by matrix multiplication, 23
multiplication, 22
multiplication properties, 26
norms, 29, 31
norms on inner product spaces, 32–33
nullspace, 39
orthogonality, 34
orthogonal projections, 38
outer product, 34
partial derivatives and gradients, 42
P-norm, 30–31
proofs, 33–34
properties, 21
quadratic forms, 37
range, 38
rectangular, 26
row and column vector multiplication, 24
row vector, square matrix, and column vector, 25
scalar multiplication, 21, 23, 27
sequences, 40
sequences, properties, 40
square, 25
Sylvester's criterion, 37–38
trace, 28
transpose, 28
transposition, 21
types, 21–22
scalars and vectors, 17
subspaces, 20
vectors, properties, 18
addition, 18
element wise multiplication, 19
subtraction, 18
Linear autoencoders *vs.* PCA, 126–127
Linear regression, 51
gradient descent algorithm, 53–54
learning rate, 54–55
multiple linear regression via gradient descent, 54
OLS, 51–53
Linear transformation, 36–37
Local minimizers, 47
conditions for, 48–49
Local search methods, 157
ACO, 159–160
genetic algorithms (GAs), 158
hill climbing, 158
simulated annealing (SA), 159
VNS, 160–161
Logistic function, 66
Logistic regression, 66–67, 186–189
limitations, 69–70
Long short-term memory (LSTM)
applications, 117
distinguishing factor, 117
forget gate, 117
overview, 116
traditional, 118
training, 118
visualization, 117
Loss layer, 107

M

Machine learning, 1, 50, 219
algorithms, 51
asset price prediction, 171–172
description of experiment, 173–175
feature selection, 175–176
supervised learning, 172–173
feature selection, 185–186
history, 50
model evaluation, 176
ridge regression, 176–178
SVR, 178–180
model training and evaluation, 186
Bayesian classifier, 191–193
KNN, 189–191
logistic regression, 186–189
proliferation, 219
speed dating, 180
classification, 181–182
data cleaning and imputation, 182–185
unsupervised learning, 74
assignment step, 74

K-means clustering, 74
K-means clustering, limitations, 75–76
update step, 75
Markov process, 87
Matrices, 20
addition, 21
column vector and square matrix, 24
derivatives and differentiability, 42
distributive over matrix addition, 27
eigenvalues and eigenvectors, 34–36
Euclidean norm, 29
Hessian, 43
hyperplanes, 39–40
inner products, 32
L1 norm, 29–30
L2 norm, 29
limits, 41
linear transformations, 36–37
matrix by matrix multiplication, 23
multiplication, 22
multiplication properties, 26
norms, 29, 31
norms on inner product spaces, 32–33
nullspace, 39
orthogonality, 34
orthogonal projections, 38
outer product, 34
partial derivatives and gradients, 42
P-norm, 30–31
proofs, 33–34
properties, 21
quadratic forms, 37
range, 38
rectangular, 26
row and column vector multiplication, 24
row vector, square matrix, and column vector, 25
scalar multiplication, 21, 23, 27
sequences, 40
properties, 40
square, 25
Sylvester's criterion, 37–38
trace, 28
transpose, 21, 28
types, 21–22
Mean squared error (MSE), 17, 65
Mixed-design ANOVA, 138
mlp() function, 100
MLP. *See* Multilayer perceptron (MLP) model
Momentum within RBMs, 132
Motherboard, 169
Multicollinearity, 62
confusion matrix, 68–69
logistic regression, 69–70
regression models, 64–67
ridge regression, 62–64
ROC curve, 67–68
SVM, 70–73
testing, 62
VIF, 62
Multilayer perceptron (MLP) model, 4
back-propagation algorithm, 95–97
considerations, 97–99
distinguishing factor from SLPs, 94
global optimum, 95
limitations, 97–99
Multiple linear regression via gradient descent, 54
Multiplicative law of probability, 13
Multivariate ANOVA (MANOVA), 138
Mxnet, 99

N

Naïve Bayes classifier, 84
Neighborhoods, concept, 49
interior and boundary points, 50
Netflix, 214
Neural history compressor, 116
Newton's method, 60–61
Non-parametric bootstrapping, 81
Norms, 29
Null hypothesis, 145

O

One-way ANOVA, 137
Online learning, 131
Optimization, 45
unconstrained, 45–46
global minimizers, 47–48
local minimizers, 47
local minimizers, conditions, 48–49
Ordinary least squares (OLS), 51–53
Orthogonality, 34
Orthogonal projections, 38

P

Parameter tuning, 173
Partial derivative, 42
Perceptron model, training, 90
Plackett-Burman designs, 146
Point prediction, 79
Pooling layer, 105
Positive semi-definite matrix, 22
Posterior distribution, 149
Power supply unit (PSU), 170
Principal components, 152
Principal components analysis (PCA), 36, 126–127, 152–154, 176
Prior distribution, 149
Probability, 11–12
Probability theory, 86
Pseudo-random numbers, 148

Q

Quadratic forms, 37
Quantitative finance, 171

R

Random access memory (RAM), 169
Random forest, 83
  limitations, 83
Randomization, 145
Random sampling, 14
Random variables, 14–15
Reactive search optimization (RSO), 161
  fixed tabu search, 163–164
  KNN, 165
  reactive prohibitions, 162–163
  RTS, 164
  WalkSAT algorithm, 165
Reactive tabu search (RTS), 164
Receiver Operating Characteristic (ROC) curve, 67–68
Rectangular matrices, 26
Rectified linear units (ReLU) layer, 106
Recurrent neural networks (RNNs), 5
  architecture, 114
  BPPT, 114–115
  Elman, 115
  example, 120–124
  fully, 113–114
  LSTM, 116–118
  neural history compressor, 116
  parameter update algorithm, 119–120
  structural damping within, 119
Regression, 172–173
Regression models, 51
  evaluating, 64–65
    classification, 65
    coefficient of determination, 65
    logistic regression, 66–67
    MSE, 65
    SE, 65
  linear regression, 51
    gradient descent algorithm, 53–54
    learning rate, 54–55
    multiple linear regression via gradient descent, 54
    OLS, 51–53
Regularization
  DropConnect, 111
  DropOut, 111
  L1 and L2, 111
  negative effect, 111
  stochastic pooling, 111
Reinforcement learning, 86–87
Relief algorithm, 157
ResNet, 110
Restricted Boltzmann machines (RBMs), 6, 125, 127
  energy function, 127
  Hopfield networks, 128
  implementations, 129
  individual activation probabilities, 129
  momentum within, 132
  probability distributions, 128
  standard, 127
  visualization, 135
Ridge regression, 62, 63, 176–178
  and LASSO, 64
Robust tabu search, 163

S

Simple cells, 101
Simulated annealing (SA), 159
Single layer perceptron (SLP) model, 3–4
  activation function, 90
  architecture, 89
  distinguishing factor from MLP, 95
  limitations, 91–93
  perceptron model, 90
  statistics, 94
  WH algorithm, 90

Singular value decomposition (SVD), 215–216
SLP. *See* Single layer perceptron (SLP) model
Solid state drives, 168
Space filling, 147
Sparsity, 133
Speed dating, 180
  classification, 181–182
  data cleaning and imputation, 182–185
Standard deviation, 16
Standard error (SE), 65
Statement of experiment, 144
Statistical concepts, 11
  and *vs.* or, 12–13
  Bayes' theorem, 14
  coefficient of determination (R squared), 17
  MSE, 17
  probability, 11–12
  random variables, 14–15
  standard deviation, 16
  variance, 15
Statistical replication, 145
Stochastic pooling, 111
Stride, 108
Structural damping, 119
Subspaces, 20
Supervised learning, 50
  regression, 172–173
Support vector machine (SVM), 70–72
  extensions, 73
  kernels, 72
  limitations, 73
  sub-gradient method applied to, 72
Support vector regression (SVR), 178–180
Sylvester's criterion, 37–38

T

Test of significance, 145
Transposition, 18
Two-way infinite sequence, 40
Two-way (multiple-way) ANOVA, 137

U

Unconstrained optimization, 45–46
  global minimizers, 47–48
  local minimizer, conditions, 48–49
  local minimizers, 47
Unsupervised learning, 74
  assignment step, 74
  K-means clustering, 74
    limitations, 75–76
  update step, 75

V

Vanishing gradient, 116
Variable neighborhood search (VNS), 160–161
Variance, 15
Variance inflation factor (VIF), 62
VGGnet, 110

W, X

Wake-sleep algorithm, 8
WalkSAT algorithm, 165
Weight decay, 133
Widrow-Hoff (WH) algorithm, 90
Wrappers, Filters, and Embedded (WFE) algorithms, 157
  relief algorithm, 157

Y

Yahoo! Finance API, 172

Z

Zero-padding, 108