INDEX

Symbols and Numbers

α (alpha), Q-learning blending, 629
β1 (beta 1), Adam parameter, 421
β2 (beta 2), Adam parameter, 421
δ (delta), proportion of error change, 357
Δ (Delta), proportion of error change, 357
ε (epsilon)
  numerical stability parameter, 419
  Q-learning policy, 630
η (eta), learning rate
  introduction, 377
  practical issues during optimization, 389
γ (gamma)
  Adagrad scaling parameter, 419
  discount factor, 611
  momentum scaling factor, 410
λ (lambda), regularization control, 204
σ (sigma), standard deviation, 24
θ (theta), coin bias, 103
P(A), simple probability, 50
P(A,B), joint probability, 53
P(A|B), conditional probability, 50
1-hot encoding, 226
1D convolution, 446
1×1 convolution, 447
20 Questions, 269
68-95-99.7 rule, 25

A

accuracy, 64
activation function
  exponential ReLU, 340
  gallery of functions, 344
  Heaviside step, 334
  identity, 331
  introduction, 318
  leaky ReLU, 337
  linear, 332
  logistic curve, 341
  maxout, 338
  noisy ReLU, 339
  parametric ReLU, 337
  piecewise linear, 336
  ReLU, 336
  role in preventing network collapse, 329
  shifted ReLU, 338
  sigmoid, 341
  sign function, 335
  sine wave, 343
  softmax, 345
  softplus, 340
  stairstep, 333
  step, 333
  swish, 341
  tanh, 341
  unit step, 334
actor, 602
acyclic graph, 324
Adaboost, 309
Adadelta, 418
Adagrad, 417
Adam, 420
adaptive code, 137
adaptive gradient learning, 417
adaptive moment estimation, 420
The Adventures of Huckleberry Finn, 146
adversarial perturbation, 491
adversary, 491

agent, 602
AI winter, 317
ancestor (in a neural network), 324
anchor point, 436
Animal Crackers, 559
Anscombe's quartet, 40
answer questions, 540
Apollo spacecraft, 297
arc (in a neural network), 323
archaeology coin-flipping game (scenario), 103
array, 329
artificial neuron. *See* neuron
atmospheric example, 205
atrous convolution, 457
attacks (with perturbations), 493
attention
  introduction, 574
  key, 575
  Q/KV, 579
  QKV, 575
  query, 575
  self-attention, 576
  value, 575
attention layer
  introduction, 578
  schematic symbol, 581
autoencoder
  basic structure, 495
  convolutional, 516
  latent layer, 501
automatic Bayes, 97
autoregression, 541
average
  common meaning, 16
  harmonic mean, 71
  mean, 16
  median, 16
  mode, 16
average pooling, 452

B

baby lengths (scenario), 33
backprop, 351
backpropagation, 351
backpropagation through time, 553
backward propagation, 376
bagging, 299
balanced decision tree, 270
base image, 680
batch gradient descent, 401
batchnorm
  discussion, 424
  schematic symbol, 424
Baxter, William, 139
Bayes' Rule
  discussion, 95
  evidence, 96
  hypothesis, 102
  likelihood, 96
  line fitting, 212
  multiple hypotheses, 109
  observation, 102
  posterior, 96
  prior, 96
  refining estimate, 101
  repeating, 101
  statement, 94
Bayes' Theorem. *See* Bayes' Rule
Bayes, Thomas, 83
Bayesian probability
  vs. frequentist, 85
  overview, 85
beam search, 595
bell curve, 22
Bernoulli distribution, 26
BERT, 590
bi-RNN, 559
bias
  in artificial neuron, 318
  bias trick, 320
  in families of curves, 204
  of a flipped coin, 86
bias in, bias out, 673
bias-variance tradeoff, 210
bidirectional RNN layer
  introduction, 559
  schematic symbol, 559
binary classifier network, 378
binary cross entropy, 150
binary relevance, 161
bind (to neuron), 314
bit, 136
blessing of non-uniformity, 173
blessing of structure, 173
block of numbers, 329

BMI (body mass index), 243
bold driver method, 400
boosting, 302
bootstrap, 299
bootstrap aggregation, 299
bootstrapping. *See* bootstraps
bootstraps
  in bagging (for ensembles), 299
  in statistics, 31
  linear, 39
  multiple, 39
  negative, 38
  partial, 39
  positive, 38
  simple, 39
  strong, 38
  weak, 38
bottleneck, 501
bouncing around the minimum, 394
boundary
  classifier, 156
  decision boundary, 158
  method, 156
  simple, 58
box filter, 451
BPTT (backpropagation through time), 553
branch (in decision tree), 270
building a sundae (scenario), 548

C

C (support vector machine parameter), 285
candling, 157
Cartesian space, 213
categorical data, 226
categorical distribution, 27
categorical variable decision tree, 271
categorization. *See* classifier
categorizer. *See* classifier
category, 156
channel (in a tensor), 430
cheating
  at finding tanks, 185
  when overfitting, 197
child
  in a decision tree, 271
  in a neural network, 324
choosing store background music (scenario), 198
class, 156
classification. *See* classifier
classifier
  binary, 156
  binary relevance, 161
  multiclass, 160
  non-parametric, 264
  one-against-all, 161
  one-versus-all, 161
  one-versus-one, 163
  one-versus-rest, 161
  overview, 6
  parametric, 264
  training, 182
closure (training task), 591
cloze task, 591
clustering
  k-means, 166
  overview, 8
CNN (convolutional neural network), 431
CNN-LSTM, 557
code
  adaptive, 137
  constant-length, 141
  fixed-length, 141
  Huffman, 148
  variable-bitrate, 143
coin detection (scenario), 86
coin flipping
  basic idea, 86
  Bayesian approach, 87
coin-flipping game (scenario), 87
collapse, 329
Common Crawl, 597
compression, 496
compression ratio, 148
concealed data poisoning, 598
conditional probability, 50
confidence (in ensemble voting), 299
confidence interval, 32
confusion matrix
  with Bayes' Rule, 97
  correct analysis, 77
  definition, 60
  proper use, 74
  wrong use, 76

connectome, 315
consistency in data preparation, 223
constant distribution, 22
constant-length code, 141
contaminated data, 187
content blending, 498
content loss (for style transfer), 683
context (in information), 135
context vector, 562
contextualized word embeddings, 571
continuous probability distribution, 21
continuous variable decision tree, 272
convergence
    of GAN, 670
    of Q-learning, 633
convnet, 431
convolution
    introduction, 433
    layer, 446
    layer schematic symbol, 471
    multidimensional, 443
    striding, 453
    transposed convolution, 457
convolution with downsampling layer symbol, 521
convolution with upsampling layer symbol, 521
convolutional autoencoder, 516
convolutional neural network, 431
correctness, 56
correlation, 37
correlation coefficient, 38
cost (for neural network error), 352
covariance, 35
cpd (continuous probability distribution), 21
credit assignment problem, 608
cross entropy, 145
cross-validation
    basic, 190
    information leakage, 239
    k-fold, 192
curse of dimensionality, 168
cusp, 118
cycle (in a neural network), 323

D

DAG (directed acyclic graph), 324
dart throwing (scenario), 48
data
    augmentation, 255
    cleaning, 221
    contamination, 187
    hygiene, 187
    leakage, 187
    poisoning, 598
    preparation, 221
dataset enlargement, 255
DCGAN (deep convolutional GAN), 667
dead neuron, 345
decay parameter (for learning rate), 396
decay schedule, 398
decision boundary, 158
decision node (in decision tree), 270
decision region, 158
decision stump, 303
decision tree
    balanced, 270
    branch, 270
    children, 271
    decision node, 270
    depth, 270
    depth limiting, 280
    distant children, 271
    edge, 270
    ensemble, 299
    immediate children, 271
    internal node, 270
    leaf, 270
    link, 270
    node, 270
    overfitting, 275
    overview, 269
    parent, 271
    pruning, 280
    root, 270
    sibling, 271
    stump, 303
    subtree, 271
    terminal node, 270
    unbalanced, 270
decoder
    for autoencoder, 496
    for seq2seq, 561
decoder block (transformer), 587

deconvolution (as transposed convolution), 457
deep convolutional GAN, 667
deep dreaming, 675
deep learning
- network structure, 326
- overview, 10

deep reinforcement learning, 647
delay step, 550
delayed exponential decay, 398
delta values (for neurons), 360
denoising, 519
dense layer. *See* fully connected layer
density, 169
dependent variable, 29
deployed system, 183
depth
- in decision tree, 270
- of a deep learning network, 327
- of a tensor, 430

depth limiting, 280
derivative
- definition, 119
- requirements, 123

descendent (in a neural network), 324
deterministic
- environment, 611
- function, 118

deterministic environment, 611
DFR (discounted future reward), 609
Dickens, Charles, 545
digits, 4
dilated convolution, 457
dimensionality reduction, 243
diminishing returns (in ensembles), 300
directed acyclic graph, 324
directed divergence, 151
directed graph, 323
discount factor, 611
discounted future reward, 609
discrete probability distribution, 20
discrimination information, 151
discriminator, 650
distant children (in decision tree), 271
distilling, 593
distinguishing cows and zebras (scenario), 224
dit, 137
domain, 158
downsampling, 451
downsizing, 457
downstream
- network, 542
- task, 542

Dr. Seuss, 137
dropout
- discussion, 422
- schematic symbol, 422

dual representation, 213
dummy variable, 226

E

E (neural network error), 359
early stopping, 202
edge
- in a decision tree, 270
- in a neural network, 323

eggs
- fertilized, 157
- quitter, 160
- unfertilized, 157
- winner, 160
- yolker, 160

eigendog, 256
eigenvector, 256
element
- in a neural network, 323
- in a tensor, 430

elementwise processing, 234
elevator scheduling (scenario), 9
ELMo, 571
ELU (exponential ReLU activation function), 340
embedding
- contextualized, 572
- sentence, 571
- space, 569
- token, 540
- word, 566

embedding layer schematic symbol, 574
encoder
- for autoencoder, 496
- for seq2seq, 561

encoder block (transformer), 587
encoder-decoder attention, 579

ensemble, 280
entropy, 143
environment (in reinforcement learning), 602
environmental state, 605
episode, 605
epoch, 182
epoch gradient descent, 401
epsilon-greedy policy, 630
epsilon-soft policy, 630
error (for neural network error), 352
error curve (for neural network error), 355
error-based decay, 399
event
  in information theory, 136
  in probability, 50
evidence, 96
expected value, 28
expert system, 4
explainability. *See* transparency
exploding gradient, 553
exploit, 608
explore or exploit dilemma, 608
exponential decay (of learning rate), 396
exponential ReLU activation function, 340
extra trees, 302
Extremely Randomized Trees, 302

F

f1 score, 71
fair coin, 86
false alarm, 73
false discovery rate, 73
false negative, 60
false negative rate, 73
false positive, 60
false positive rate, 73
FC. *See* fully connected layer
feature
  as part of a sample, 157
  a target of convolution filter, 438
feature bagging, 301
feature detector, 438
feature engineering, 5
feature filtering, 243
feature map, 438
feature selection, 243
featurewise processing, 233
feed-forward networks, 322
feedback, 602
feedback (in a neural network), 324
fertilized, 157
few-shot training, 594
fiber size (of a tensor), 430
fibre, 234
fidelty (of encoding), 497
fill a box with a sphere (scenario), 175
filter
  kernel, 433
  value, 433
finding animals for movie (scenario), 566
finding information, 4
fine-tuning, 542
finite distribution, 22
fist (sending Morse Code), 138
fitting wind speed data (scenario), 205
fixed-length code, 141
fixed-step decay, 399
flat distribution, 22
flatten layer
  example, 475
  illustration, 476
  schematic symbol, 474
Flippers game (scenario), 614
fold (in cross-validation), 192
footprint (of a convolution filter), 436
forest, 301
forging money (scenario), 650
fractional convolution, 457
frequentist probability
  vs. Bayesian, 85
  overview, 84
frozen weights
  deep dreaming, 677
  in a GAN, 660
full modal collapse, 671
fully connected layer
  introduction, 328
  schematic symbol, 328
fully connected network, 328
function
  argument, 118
  continuous, 118
  graphed, 124

- mathematical, 118
- return values, 118
- single-valued, 119
- smooth, 118

G

g (gradient), 410
game theory, 656
GAN (generative adversarial network), 649
garbage in, garbage out, 223
gated recurrent unit, 554
Gaussian distribution, 22
GD. *See* gradient descent
generalization
- accuracy, 197
- error, 197
- loss, 197

generated image (for style transfer), 680
generative adversarial network, 649
generator, 650
generator-discriminator, 664
Gini impurity, 281
Glorot Normal initialization, 325
Glorot Uniform initialization, 325
GPT-2, 593
GPT-3
- discussion, 596
- performance, 597

gradient
- definition, 126
- vanishing, 129
- zero, 129

gradient ascent, 481
gradient descent
- Adadelta, 418
- Adagrad, 417
- Adam, 420
- batch, 401
- epoch, 401
- mini-batch, 405
- momentum, 409
- Nesterov, 414
- RMSprop, 418
- stochastic, 403
- typical meaning, 407

Gram matrix, 681
graph
- acyclic, 324
- directed, 323
- introduction, 323
- weighted graph, 324

graph theory, 323
Great Dane, 184
greedy algorithm (for decision trees), 274
Green Eggs and Ham, 137
grid, 329
ground truth, 57
grouping. *See* clustering
GRU (gated recurrent unit), 554
guitar data, 228

H

He Normal initialization, 325
He Uniform initialization, 325
head (in attention layer), 580
Heaviside step activation function, 334
hidden layer, 327
hidden state, 549
hierarchy
- of features, 440
- of filters, 461
- of scales, 471

high-dimensional spaces, 42
high-dimensional weirdness
- hypercube, 175
- hyperorange, 178
- hypersphere, 175
- packing hyperspheres in hypercubes, 177
- volume of a hypersphere in a hypercube, 175

hill (3D surface), 129
hit rate, 66
Holmes, Sherlock (stories)
- generating with GPT2, 596
- generating with RNN, 555

Hopper, Edward, 685
Huffman code, 148
husky, 185
hypercube, 175
hyperorange, 178
hyperparameter
- number of clusters, 166
- tuning, 168

hypersphere, 175
hypothesis (in Bayes' Rule), 102

I

i.i.d. (independent and identically distributed), 29
ice-cream shop (scenario), 55
idealized curve, 205
identically distributed, 29
identifying dog breeds (scenario), 183
identity activation function, 331
IG (information gain), 281
image classifier, 6
image matrix (for style transfer), 681
ImageNet, 478
immediate children (in decision tree), 271
impurity, 280
inceptionism, 679
independent and identically distributed, 29
independent variable, 29
inertia, 408
infinite distribution, 24
information
- definition, 137
- density, 143
- divergence, 151
- finding, 4
- gain, 281
- properties, 136

information exchange
- global context, 135
- local context, 135
- receiver, 134
- sender, 134
- surprise, 134

information leakage
- basic idea, 187
- in cross validation, 239

information theory, 133
initial state (in reinforcement learning), 605
initialization (of a neural network), 325
input layer, 327
input tensor, 329
instant reward, 610
internal node (in decision tree), 270
interval decay, 399
inverse transformation, 234
inversion (of encoding), 497

J

joint probability, 53
JPG encoder, 495
junkyard (scenario), 17

K

k
- dimensions left after PCA, 249
- number of clusters, 166
- number of folds, 192
- number of nearest neighbors, 264

k-means clustering, 166
k-nearest neighbors, 264
Kandinsky, Wassily, 685
kernel (of a convolution filter), 433
kernel trick, 287
key (for transformer), 575
kissing line (tangent), 122
KL divergence, 151
kNN (k-nearest neighbors), 264
Kullback-Leibler divergence, 151

L

L-Learning, 616
L-table, 617
L-value, 617
label, 156
language model, 541
last-minute stopping, 202
latent layer, 501
latent space, 509
latent variable, 501
law of diminishing returns in ensemble construction, 300
layer normalization, 583
layer schematic symbol
- attention, 581
- batchnorm, 425
- bidirectional recurrent, 559
- convolution, 471
- convolution with downsampling, 521
- convolution with upsampling, 521
- dropout, 423
- embedding, 574

flatten, 474
fully connected, 328
masked attention, 588
multi-head attention, 581
norm-add, 583
pooling, 451
positional embedding, 584
recurrent, 555
recurrent cell, 551
reshape, 667
self-attention, 581
layers
introduction, 322
overview, 10
lazy algorithm (kNN), 265
leaf (in decision tree), 270
leaky ReLU activation function, 337
learning rate
in neural networks, 377
in Q-learning, 629
learning rate adjustment
constant, 391
delayed exponential decay, 398
error-based decay, 399
exponential decay, 396
fixed-step decay, 399
interval decay, 399
learning the slow way (scenario), 354
LeCun Normal initialization, 325
LeCun Uniform initialization, 325
letter frequencies in English, 139
life-seeking probes (scenario), 97
likelihood, 96
line (in a neural network), 323
line fitting with Bayes' Rule (scenario), 212
linear activation function, 332
linear function, 331
linear layer. *See* fully connected layer
link (in decision tree), 270
list, 329
local receptive field, 436
logical flow, 540
logistic curve activation function, 341
long short-term memory, 553
long-term dependency problem, 563
loop (in a neural network), 324
loss, 352
lossless encoding, 496
lossy encoding, 496
lousy learning, 616
low-dimensional dog breed descriptions (scenario), 255
low-pass filter, 451
LSTM (long short-term memory), 553

M

m
modification of neuron value, 357
momentum (in gradient descent), 410
machine learning, 3
magnitude (of a gradient), 128
The Manchurian Candidate, 598
mapping (transformation), 236
marginal probability, 55
Marx, Groucho, 559
masked attention layer schematic symbol, 588
masking (attention decoder), 588
matching a face mask (scenario), 462
matrix, 329
max pooling, 452
maximum ascent, 126
maximum descent, 126
maximum of a function
finding with derivative, 125
finding with gradient, 128
global maximum, 119
local maximum, 121
maxout activation function, 338
McGlassface, Glasses, 61
mean, 16
mean normalization, 228
mean of Gaussian, 24
mean squared error, 365
mean subtraction, 228
median, 16
mellowmax, 632
messaging with mirrors (scenario), 497
mini-batch, 405
mini-batch gradient descent, 405
minimum of a function
finding with derivative, 125
finding with gradient, 128
global minimum, 119
local minimum, 121

mixing colors of paint (scenario), 574
MLP (multilayer perceptron), 328
MNIST
    used for an autoencoder, 505
    used for convolution, 473
    used for GAN, 667
modal collapse, 671
mode, 16
momentum, 408
momentum gradient descent, 409
Monet, Claude, 685
morbus pollicus (scenario), 63
Morse code, 137
Morse, Samuel, 139
MP (*morbus pollicus*), 63
MP3 encoder, 495
MSE (mean squared error), 365
multi-head attention
    introduction, 580
    schematic symbol, 581
multiclass classification, 160
multidimensional convolution, 443
multifilter loss, 676
multilayer loss, 676
multilayer perceptron, 328
multinoulli distribution, 27
multiple correlation, 39
multivariate transformation, 231
Munch, Edvard, 685
Muppets
    BERT, 590
    ELMo, 571
music system
    collecting data, 199
    good fit, 201
    overfitting, 199
    underfitting, 200

N

n-gram, 595
naive Bayes classifier, 290
Nash equilibrium, 657
natural language generation, 541
natural language processing, 540
natural language understanding, 540
naughts and crosses, 603
negative correlation, 38
negative covariance, 35
negative gradient, 128
negative predictive value, 73
neighbor (for kNN), 265
neighborhood, 120
Nesterov momentum, 414
network collapse, 329
neural network
    introduction, 315
    simple example, 321
neural style transfer, 680
neuron
    artificial, 317
    dead, 345
    overview, 11
    real, 314
neurotransmitter, 314
New Horizons spacecraft, 190
next sentence prediction, 591
NLG (natural language generation), 541
NLP (natural language processing), 540
NLU (natural language understanding), 540
No Free Lunch Theorem, 422
node
    in a decision tree, 270
    in a neural network, 323
node splitting
    Gini impurity, 281
    impurity, 280
    information gain, 281
    overview, 272
    purity, 280
noisy curve, 205
noisy ReLU activation function, 339
nominal data, 226
non-deterministic (variational autoencoder), 521
non-parametric classifier, 264
nonlinear function, 331
nonlinearity. *See* activation function
norm-add
    definition, 583
    schematic symbol, 583
normal deviate, 24
normal distribution, 22
normalization, 228
NSP (next sentence prediction), 591
numerical data, 226

O

OAA (one-against-all), 161
observability, 608
observation (in Bayes' Rule), 102
offline algorithm, 403
on-demand algorithm (kNN), 265
one-against-all, 161
one-by-one convolution, 447
one-hot encoding, 226
one-versus-all, 161
one-versus-one, 153
one-versus-rest, 161
online algorithm, 405
optimizers, 387
ordinal data, 226
outlier
    when cleaning data, 223
    when making a boundary, 201
output layer, 327
output tensor, 329
OvA (one-versus-all), 161
overfitting
    decision trees, 275
    definition, 196
    dog breeds, 185
overshoot (during gradient descent), 396
OvO (one-versus-one), 163
OvR (one-versus-rest), 161

P

padding, 440
parameter space (for autoencoder), 508
parametric blending, 498
parametric classifier, 264
parametric ReLU activation function, 337
paraphrase (text), 540
parent
    in a decision tree, 271
    in a neural network, 324
partial correlation, 39
partial modal collapse, 671
PCA. *See* principal component analysis
pdf (probability density function), 21
penalty (for neural network error), 352
perceptron
    introduction, 316
    Mark I Perceptron, 317
perfect precision, 69
perfect recall, 69
performance metrics
    illustrated, 75
    summary, 73
perturbation, 491
Picasso, Pablo, 681
piecewise linear activation function, 336
pixel, 430
planet mining (scenario), 97
plateau (3D surface), 129
plurality voting, 299
Pluto images, 190
pmf (probability mass function), 20
pointwise feed-forward layer, 587
policy, 606
polysemy, 559
poodle, 183
pooling
    average, 452
    discussion, 449
    layer, 452
    max, 452
    schematic symbol, 451
population (in statistics), 32
positional
    encoding, 584
    schematic symbol, 584
positional embedding, 585
positive correlation, 38
positive covariance, 35
positive predictive value, 64
posterior, 96
precision
    definition, 64
    perfect, 69
precision-recall tradeoff, 67
predicted value, 57
predicting traffic from temperature (scenario), 234
prediction, 6
pretraining, 542
principal component analysis
    description, 244
    line of maximum variance, 246
    projection, 244
prior, 96

probability
conditional, 50
dependence, 50
event, 50
independence, 50
introduction, 19
joint, 53
marginal, 55
simple, 50
probability density function, 21
probability distribution
bell curve, 22
Bernoulli, 26
categorical, 27
continuous, 20
example in junkyard, 19
finite, 22
Gaussian, 22
infinite, 24
introduction, 17
multinoulli, 27
normal, 22
uniform, 21
probability mass function, 20
ProGAN, 670
projection, 244
prompt (for text generation), 541
protect an orange in transit (scenario), 177
pruning, 280
purity, 280

Q

Q-Learning, 626
Q-table, 627
Q-value, 627
Q/KV (query / key, value), 579
QKV (query, key, value), 575
quadratic cost function, 365
quality control for dolls (scenario), 61
quality learning, 626
quantitative data, 226
query (for transformer), 575
quitter, 160

R

random forest, 301
random labeling, 302
random variable
description, 20
drawing, 20
introduction, 17
randomness, 16
real-world data, 183
recall
definition, 66
perfect, 69
receptor site (neuron), 314
recipes, 6
reconstruction (of encoded signal), 496
recurrent cell
introduction, 550
schematic symbol, 551
recurrent layer schematic symbol, 555
recurrent neural network
bidirectional, 558
deep, 557
deep bidirectional, 560
introduction, 548
reference point (of a convolution filter), 436
regularization
batchnorm, 424
dropout, 422
introduction, 203
layer normalization, 583
reinforcement learning, 602
rejection, 73
relative entropy, 151
release data, 183
released system, 183
ReLU activation function, 336
remembering people's names (scenario), 196
reparameterization trick, 527
representation blending, 498
reshape layer
introduction, 667
schematic symbol, 667
residual connection, 582
resulting reward, 610
reward
discounted future, 609
final, 604
instant, 610
introduction, 602

resulting, 610
total, 609
total future, 609
ultimate, 604
reward signal. *See* reward
rigged coin, 86
RL (reinforcement learning), 602
RMSprop, 418
RNN. *See* recurrent neural network
rolled-up diagram, 550
root (in decision tree), 270
rotation validation. *See* cross-validation
rules, 4

S

saddle (3D surface), 130
sample, 156
sample set, 32
samplewise processing, 232
SARSA, 639
scenarios
archaeology clustering, 8
archaeology coin-flipping game, 103
baby lengths, 33
building a sundae, 548
choosing store background music, 198
coin detection, 86
coin-flipping game, 87
dart throwing, 48
distinguishing cows and zebras, 224
elevator scheduling, 9
fill a box with a sphere, 175
finding animals for movie, 566
fitting wind speed data, 205
Flippers game, 614
forging money, 650
ice-cream shop, 55
identifying dog breeds, 183
junkyard, 17
learning the slow way, 354
life-seeking probes, 97
line fitting with Bayes' Rule, 212
low-dimensional dog breed descriptions, 255
matching a face mask, 462
messaging with mirrors, 497
mixing colors of paint, 574
morbus pollicus, 63
planet mining, 97
predicting traffic from temperature, 234
protect an orange in transit, 177
quality control for dolls, 61
remembering people's names, 196
sorting eggs, 156
testing for disease, 63
text messaging, 135
transmitting books, 146
scientific notation, 222
seed (for text generation), 541
selection
sampling with replacement (SWR), 30
sampling without replacement (SWOR), 30
with replacement, 30
without replacement, 30
selection with replacement, 30
selection without replacement, 30
selectively persistent short-term memory, 553
self-attention, 576
self-attention layer
introduction, 578
schematic symbol, 581
semantics, 546
semi-supervised learning, 501
sensitivity, 66
sentence embedding, 571
sentiment analysis
introduction, 540
with BERT, 592
seq2seq, 561
sequence, 539
SGD (stochastic gradient descent), 401
Shannon, Claude, 133
shift invariance, 453
shifted ReLU activation function, 338
SI (slope-intercept) space, 213
Siberian husky, 185
sibling (in decision tree), 271
sigmoid activation function, 341
sign activation function, 334

simple correlation, 39
simple probability, 50
sine wave activation function, 343
skip connection, 582
sleuth (of bears), 568
slice processing
    elementwise, 234
    featurewise, 233
    samplewise, 232
sliding window, 543
slope (tangent), 122
softmax
    as activation function, 345
    in Q-learning, 630
softplus activation function, 340
sorting dolls (scenario), 61
sorting eggs (scenario), 156
spatial filter, 436
specificity, 73
splitting. *See* node splitting
stairstep activation function, 332
standard deviation, 24
standardization, 229
Starship *Theseus*, 97
state
    agent (reinforcement learning), 605
    environmental (reinforcement learning), 605
    hidden (RNN), 549
    of an RNN, 548
    starting (RNN), 548
    variable (reinforcement learning), 605
state variable, 605
statistics, 15
step activation function, 332
Stevenson, Robert Louis, 137
stochastic environment, 611
stochastic gradient descent, 401
striding, 453
strong learner, 303
style image, 680
style loss, 683
style matrix, 681
style transfer, 680
subjective Bayes, 97
subtree (of decision tree), 271
summarize (text), 540
supervised learning
    for classification, 156
    overview, 6
support layer, 327
support vector, 284
support vector machine
    kernel trick, 287
    overview, 282
    strictness parameter C, 285
    support vector, 284
surprise, 134
SVM. *See* support vector machine
sweeping (a filter), 433
swish activation function, 340
SWOR (selection without replacement), 30
SWR (selection with replacement), 30

T

A Tale of Two Cities
    with embedding, 579
    prediction, 545
tangent, 122
tangent line, 122
target (one-hot encoded label), 361
temperature (selecting outputs), 595
tensor, 328
terminal node (in decision tree), 270
test data
    definition, 186
    never learn from, 186
testing for disease (scenario), 63
text messaging (scenario), 135
TFR (total future reward), 609
threshold
    for artificial neuron, 316
    for real neuron, 314
tic-tac-toe, 603
time step, 549
toad visual system
    hierarchy, 461
    introduction, 437
token
    embedding, 566
    as a word, 540
total future reward, 609

total reward, 609
training
  accuracy, 196
  basic idea, 182
  deep network overview, 12
  error, 196
  flowchart, 189
  loss, 196
training set, 182
transfer function. *See* activation function
transfer learning, 542
transformation
  inverse, 234
  multivariate, 231
  overview, 223
  univariate, 231
transformer
  introduction, 586
  structure, 588
translation, 540
transmitting books (scenario), 146
transparency
  decision trees, 274
  one-versus-one classifiers, 164
transposed convolution, 457
Treasure Island, 137
triple modular redundancy, 297
true discovery rate, 73
true negative, 60
true positive, 60
true positive rate, 66
Turner, J.M.W., 685
Twain, Mark, 146
two-moons dataset, 275
type I error, 73
type II error, 73

U

unbalanced decision tree, 270
uncertainty, 143
underfitting, 197
unfair coin, 86
unfertilized, 157
uniform distribution, 21
unit (for artificial neuron), 315
unit step activation function, 334
univariate transformation, 231
Universal Language Model Fine-Tuning, 574
universal perturbation, 491
unrolled diagram, 550
unsupervised learning, 8
update rule, 617
upsampling, 457
upsizing, 457
user data, 183

V

VAE (variational autoencoder), 521
Vail, Alfred, 139
validation data. *See* validation set
validation error, 197
validation set
  definition, 187
  estimating error from, 189
valley (3D surface), 129
value (for transformer), 575
van Gogh, Vincent, 685
vanishing gradient
  on surfaces, 128
  training an RNN, 553
variable-bitrate code, 143
variance
  with respect to bias, 204
  in statistics, 26
variance normalization, 229
variational autoencoder, 521
vector, 329
vertex (in a neural network), 323
VGG16
  for creative applications, 676
  introduction, 478
visualizing filters, 480
vocabulary, 541
volume, 329
voting (ensembles)
  confidence, 299
  overview, 298
  plurality voting, 299
  weighted plurality voting, 299

W

weak learner, 303

weight
- naming convention for neurons, 322
- for neurons, 316
- overview in deep network, 11

weight sharing, 433
weighted coin, 86
weighted graph, 324
weighted plurality voting, 299
weirdness (high-dimensional), 175
width (of a recurrent cell), 551
winner (egg), 160
wire (in a neural network), 323
word embedding, 566

X

Xavier Normal initialization, 325
Xavier Uniform initialization, 325

Y

yolker, 160
Yorkshire terrier, 184

Z

zero gradient, 128
zero padding, 442
zero point, 436
zero-dimensional array, 328
zero-shot training, 594