

# INDEX

## A

- Adadelta, 299
- Adagrad, 299
- Adam, 300
- affine transformation, 128
- arithmetic mean, 70
- AutoML, 283

## B

- backpropagation, 244
  - algorithm, 256
  - by hand
    - code, 249
    - derivatives, 247
  - computational graph, 267
  - error, 255
  - fully connected network, 255
    - implementation, 260
    - loss, 255
    - symbol-to-number, 268
    - symbol-to-symbol, 268
- batch training, 282
- Bayes' theorem, 21, 31
  - evidence, 59
  - likelihood, 59
  - Naive Bayes classifier, 62
    - independence assumption, 63
  - posterior probability, 59
  - prior probability, 60
  - uninformed prior, 62
  - updating the prior, 61
- Bayes, Thomas, 59
- beta distribution, 53
- binomial function, 47
- birthday paradox, 26
- block matrix, 125
- box plot, 80
  - fliers, 82
  - whiskers, 82

## C

- calculus
  - derivative
    - chain rule, 170, 184
    - constant, 167
    - definition, 165, 166
    - directional, 188
    - exponential, 175
    - first, 167
    - logarithm, 176
    - mixed partial, 183
    - notation, 166
    - partial, 181
    - power rule, 168
    - product rule, 169
    - quotient rule, 169
    - rules (table), 176
    - second, 167
    - trigonometric functions, 172
  - differential, 163
  - differentiation, 167
  - extrema (extremum), 177
  - global extrema, 178
  - gradient, 186
    - vector fields, 186
  - inflection point, 178
  - integral, 163
  - limit, 166
  - local extrema, 178
  - matrix calculus. *See* matrix calculus
  - maxima, 177
  - minima, 177
  - notation, 187
  - saddle point, 178
  - scalar field, 186
  - secant line, 165
  - slope, 164
  - stationary points, 165
  - tangent line, 165
- Cartesian product, 118
- central limit theorem, 55
- centroid, 149

- chain rule (derivatives), 170, 184
- chain rule (probability), 37
- Cohen's  $d$ , 97
- combinations (formula), 28
- computational graph, 267
- conditional probability, 31
- confidence interval (CI), 96
  - calculating, 97
  - interpreting, 97
  - critical value, 97
- confusion matrix, 254
- contingency tables, 254
- contour plot, 276
- convolution, 229
  - 1D, 230
  - 2D, 233
  - neural network (CNN), 234
  - cross-correlation, 232
  - filter, 235
  - kernel, 231
  - nonlinearity, 237
  - stride, 234
  - zero padding, 231
- correlation
  - Pearson, 86
  - Spearman, 90
- covariance matrix, 147
- cross product
  - right-hand rule, 119

## D

- data
  - box plot, 80
  - channels, 223
  - CIFAR-10, 224
  - dataset as matrix, 222
  - Fashion-MNIST, 290
  - feature space, 105
  - features, 105
  - interval, 68
  - missing data, 83
  - MNIST, 239
  - nominal, 68
  - not a number (NaN), 83
  - one-hot encoding, 69
  - ordinal, 68
  - outliers, 82
  - ratio, 68

- shape, 225
- summary statistics, 70
- degrees of freedom, 95
- derivative
  - chain rule, 170, 184
  - constant, 167
  - differentiation, 167
  - directional, 188
  - exponential, 175
  - first, 167
  - gradient
    - vector field, 186
  - logarithm, 176
  - mixed partial, 183
  - notation, 166
  - partial, 181
  - power rule, 168
  - product rule, 169
  - quotient rule, 169
  - rules (table), 176
  - scalar field, 186
  - second, 167
  - trigonometric functions, 172
- determinant
  - properties, 134
- deviation
  - mean, 74
  - sample, 75
- differential equation
  - autonomous system, 206
  - critical points, 207
- distribution (probability), 41
  - Bernoulli, 48
  - beta, 53
  - binomial, 46
  - continuous, 51
  - Fast Loaded Dice Roller (FLDR), 49
  - gamma, 53
  - Gaussian, 53
  - lognormal, 83
  - normal, 53, 74, 83
  - Poisson, 48
  - probability density function, 53
  - uniform, 45, 51
- dot product, 114

## E

- effect size, 97
- eigenvalue, 141

eigenvector, 141  
Euclidean distance, 146  
event, 18  
evidence (Bayesian), 59

## F

F1 score, 72  
false negative (FN), 251  
false positive (FP), 251  
Fashion-MNIST, 290  
Fast Loaded Dice Roller (FLDR), 49  
feature space, 105  
features, 105  
floating-point numbers, 5  
fully connected layer, 239

## G

gamma distribution, 53  
Gaussian distribution, 53  
geometric mean, 71  
gradient  
    vector field, 186  
gradient descent  
    Adadelta, 299  
    Adagrad, 297, 299  
    Adam, 297, 300  
    batch training, 282  
    effect of multiple minima, 281  
    in 1D, 272  
    in 2D, 276  
    minibatch, 283, 284  
    momentum, 285  
        neural network, 289  
    Nesterov momentum, 294  
    online learning, 283  
    RMSprop, 297  
    stochastic, 282  
    vanilla, 272

## H

Hadamard product, 110  
harmonic mean, 72  
Hessian matrix, 211  
    as Jacobian of gradient, 212  
    Cholesky decomposition, 217  
    critical points, 213  
    optimization, 214  
    quadratic approximation, 215

Hinton, Geoffrey, 297  
histogram  
    converting to probabilities, 43  
    definition, 43  
hypothesis testing, 92  
     $p$ -value, 95  
    alpha, 96  
    alternative hypothesis, 94  
    assumptions, 95  
    calculate CI, 97  
    confidence interval (CI), 96  
        interpreting, 97  
    critical value, 97  
    degrees of freedom, 95  
    hypothesis, 94  
    interpretation, 94  
    Mann-Whitney U, 93, 99  
    nonparametric, 93  
    null hypothesis, 94  
    one-sided, 94  
    parametric, 93, 95  
    statistically significant, 96  
    t-test, 93  
        assumptions, 95  
        two-sided, 94  
        warning, 96  
    Welch's t-test, 95  
    Wilcoxon rank sum test, 99

## I

identity matrix, 132  
indefinite matrix, 140  
inertia, 285  
inner product, 114  
interquartile range, 82  
inverse matrix, 138

## J

joint probability, 37  
    table, 33

## K

$k$ -nearest neighbors, 105  
Kronecker product, 125  
Kullback-Leiber divergence, 151

## L

- L1-norm, 145
- L2-norm, 145
- law of large numbers, 58
- Let's Make a Deal*, 46
- likelihood (Bayesian), 59
- linear algebra
  - definition, 103
  - Hadamard product, 110
  - Kullback-Leibler divergence, 151
  - matrix, 105
    - affine transformation, 128
    - block, 125
    - characteristic equation, 142
    - characteristic polynomial, 142
    - cofactor, 136
    - cofactor expansion, 136
    - conjugate transpose, 140
    - covariance, 147
    - determinant, 134
    - determinant properties, 134
    - direct product, 125
    - eigenvalue, 141
    - eigenvector, 141
    - Hermitian, 140
    - Hermitian adjoint, 140
    - identity, 132
    - indefinite, 140
    - inverse, 138
    - Kronecker product, 125
    - minor, 135
    - Moore-Penrose pseudoinverse, 160
    - multiplication, 120, 121
    - negative definite, 140
    - negative semidefinite, 140
    - nondegenerate, 138
    - nonsingular, 138
    - ones, 131
    - order, 106
    - orthogonal, 139
    - positive definite, 140
    - positive semidefinite, 140
    - rotation, 128
    - singular, 138
    - square, 123
    - symmetric, 139
    - trace, 130
    - transpose, 130
    - triangular, 133
    - zero, 131
  - principal component analysis (PCA), 154
  - relative entropy, 151
  - scalar, 104
  - singular value decomposition (SVD), 157
  - tensor
    - arithmetic, 109
    - array operations, 109
    - order, 106
  - vector
    - boxcar distance, 146
    - centroid, 149
    - Chebyshev distance, 146
    - city block distance, 146
    - column, 104
    - cross product, 119
    - dot product, 114
    - Euclidean distance, 146
    - infinite norm, 145
    - inner product, 114
    - magnitude, 112
    - Mahalanobis distance, 148
    - Manhattan distance, 146
    - norm, 144
    - orthogonal, 115
    - outer product, 117
    - projection, 116
    - right-hand rule, 119
    - row, 104
    - Taxicab distance, 146
    - transpose, 113
    - unit, 112
- logistic function, 229

## M

- Mahalanobis distance, 148
- Manhattan distance, 146
- Mann-Whitney U, 99
  - null hypothesis, 100
- marginal probability, 33
- Matplotlib, 12
- matrix, 105
  - affine transformation, 128
  - block, 125
  - calculus. *See* matrix calculus
  - characteristic equation, 142

- characteristic polynomial, 142
- cofactor, 136
- cofactor expansion, 136
- conjugate transpose, 140
- covariance, 147
- determinant
  - properties, 134
- direct product, 125
- eigenvalue, 141
- eigenvector, 141
- Hermitian, 140
- Hermitian adjoint, 140
- identity, 132
- indefinite, 140
- inverse, 138
- Kronecker product, 125
- minor, 135
- Moore-Penrose pseudoinverse, 160
- multiplication, 121
  - properties, 120
- negative definite, 140
- negative semidefinite, 140
- nondegenerate, 138
- nonsingular, 138
- ones, 131
- order, 106
- orthogonal, 139
- positive definite, 140
- positive semidefinite, 140
- rotation, 128
- singular, 138
- square, 123
- symmetric, 139
- trace, 130
- transpose, 130
- triangular, 133
- zero, 131

matrix calculus, 193

- chain rule
  - scalar function by matrix, 203
  - scalar function by vector, 200
  - vector function by scalar, 202
  - vector function by vector, 203
- comparing Jacobians, gradients, slopes, 205
- denominator layout, 194
- derivative
  - activation function, 219
  - element-wise operations, 217
- Hessian matrix, 205, 211
  - as Jacobian of gradient, 212
- identities
  - scalar function by matrix, 203
  - scalar function by vector, 199
  - vector function by scalar, 202
  - vector function by vector, 203
- Jacobian matrix, 205
- matrix function by scalar, 198
- numerator layout, 194
- scalar function by matrix, 198
- scalar function by vector, 196
- table of derivatives, 194
- tangent vector, 196, 198
- vector function by vector, 197
- vector-valued function, 195

matrix direct product, 125

matrix multiplication, 121

- properties, 120

mean

- arithmetic, 70
- geometric, 71
- harmonic, 72

mean deviation, 74

median, 72

median absolute deviation, 76

minibatch, 283, 284

missing data, 83

momentum, Nesterov, 294

Monty Hall dilemma, 19, 46

mutually exclusive events, 24

## N

Naive Bayes classifier, 62

- independence assumption, 63

nearest centroid classifier, 149

negative definite matrix, 140

negative semidefinite matrix, 140

neural network

- bias trick, 129
- bias vector, 225
- convolution, 229
  - 1D, 230
- convolutional layer, 234
  - filter, 235
- dataset shape, 225
- embedding, 119
- features, 105

- neural network, *continued*
  - features space, 105
  - feedforward, 225
  - fully connected layer, 239
  - hyperparameter, 283
  - initialization, 42
  - logistic function, 229
  - minibatch, 224, 264
  - momentum, 289
  - pooling layer, 237
  - rectified linear unit (ReLU), 226
  - sigmoid function, 229
  - training, 259
  - weight matrix, 225
- neuroevolution, 217
- Newton's method, 208
  - Hessian matrix, 216
  - Jacobian, 209
  - Taylor series approximation, 215
- normal distribution, 53
- not a number (NaN), 83
- NumPy, 4
  - array indexing, 8
  - arrays on disk, 10
  - broadcasting, 110
  - colon, 9
  - data types, 6
  - defining arrays, 5
  - ellipsis, 9
  - matrix multiplication, 123
  - special arrays, 7
- O**
- one-hot encoding, 69
- online learning, 283
- optimization
  - first-order, 214, 215
  - neuroevolution, 217
  - second-order, 214, 215
  - intractable, 217
- orthogonal matrix, 139
- outer product, 117
- outliers, 82
- P**
- p*-value, 95
- principal component analysis (PCA), 154
- pooling layer
  - average, 238
  - information loss, 238
  - maximum, 237
- positive definite matrix, 140
- positive semidefinite matrix, 140
- posterior probability, 59
- precision, 72
- principal component analysis (PCA), 154
- prior probability, 60
  - updating, 61
- probability
  - Bayes' theorem, 21, 31
    - evidence, 59
    - likelihood, 59
    - posterior probability, 59
    - prior probability, 60
    - uninformed prior, 62
    - updating the prior, 61
  - birthday paradox, 26
  - central limit theorem, 55
  - chain rule, 37
  - conditional, 31
  - definition, 18
  - distribution, 41
    - Bernoulli, 48
    - beta, 53, 83
    - binomial, 46
    - continuous, 51
    - discrete, 45
    - Fast Loaded Dice Roller (FLDR), 49
    - from histogram, 44
    - gamma, 53
    - Gaussian, 53
    - lognormal, 83
    - normal, 53, 83
    - Poisson, 48
    - probability density function, 53
    - uniform, 45, 51
  - enumerating the sample space, 23
  - event, 18
  - joint, 17, 33, 37
  - joint probability table, 33
  - law of large numbers, 58
  - marginal, 17, 33
  - mutually exclusive events, 24

- of an event, 22
- product rule, 25
- product rule (conditional), 31
- random variable
  - continuous, 19
  - discrete, 19
- sample, 18
- sample space, 18
- sum rule (dependent events), 25
- sum rule (independent events), 24
- Titanic*, 35
- total, 32
- two dice, 23
- probability (distribution)
  - beta, 83
- probability density function, 53
- product rule (probability)
  - conditional events, 31
  - independent events, 25
- PyTorch, 267

**Q**

- quantiles, 78
- quartiles, 78

**R**

- random variable
  - continuous, 19
  - discrete, 19
- recall, 72
- recursion, 135
- reinforcement learning, 298
- relative entropy, 151
- RMSprop, 297
- rotation matrix, 128

**S**

- saddle point, 178
- sample (probability), 18
- sample space, 18
- scalar, 104
- scikit-learn, 14
- SciPy, 11
- sigmoid function, 229
- singular matrix, 138
- singular value decomposition, 157
- square matrix, 123
- standard error (of the mean), 77
- statistically significant, 12, 96
- statistics
  - correlation, 86
    - Pearson, 86
    - Spearman, 90
  - definition, 67
  - degrees of freedom, 95
  - deviation
    - biased sample, 75
    - mean, 74
    - median absolute, 76
    - standard error, 77
    - unbiased sample, 75
  - F1 score, 72
  - hypothesis testing, 92
    - alpha, 96
    - alternative hypothesis, 94
    - assumptions, 95
    - calculate CI, 97
    - Cohen's *d*, 97
    - confidence interval (CI), 96
    - critical value, 97
    - effect size, 97
    - hypothesis, 94
    - interpretation, 94
    - Mann-Whitney U, 93, 99
    - nonparametric, 93
    - null hypothesis, 94
    - one-sided, 94
    - p*-value, 95
    - parametric, 93, 95
    - statistically significant, 96
    - t-test, 93, 95
    - t-test assumptions, 95
    - two-sided, 94
    - warning, 96
    - Welch's t-test, 95
    - Wilcoxon rank sum test, 99
  - interquartile range, 82
  - Mann-Whitney U, 292
  - mean
    - arithmetic, 70
    - geometric, 71
    - harmonic, 72
  - median, 70, 72
  - nonstationary, 298
  - one-hot encoding, 69
  - precision, 72
  - quantiles, 78

- statistics, *continued*
  - quartiles, 78
  - recall, 72
  - standard deviation, 70
  - standard error, 70
  - stationary, 298
  - summary, 70
  - t-test, 292
  - types of data
    - interval, 68
    - nominal, 68
    - ordinal, 68
    - ratio, 68
  - variance, 70
    - biased sample, 75
    - median absolute, 76
    - standard error, 77
    - unbiased sample, 75
- stochastic gradient descent, 282
- sum rule (probability)
  - dependent events, 25
  - independent events, 24
- Sumerian cities, 32
- summary statistics, 70
- swarm optimization, 217
- symmetric matrix, 139

**T**

- t-test
  - assumptions, 95
  - confidence interval, 97
  - Welch's, 95
- Taylor series, 214
- tensor, 106
  - arithmetic
    - array operations, 109
  - order, 106
- TensorFlow, 267
- toolkits, 2
- total probability, 32
- trace, 130
- training a network, 259
- transcendental function, 214
- triangular matrix, 133
- true negative (TN), 254
- true positive (TP), 254
- $2 \times 2$  tables, 254

**U**

- uninformed prior, 62

**V**

- variance
  - biased sample, 75
  - median absolute, 76
  - standard error, 77
  - unbiased sample, 75
- vector
  - boxcar distance, 146
  - centroid, 149
  - Chebyshev distance, 146
  - city block distance, 146
  - column, 104
  - cross product, 119
  - dot product, 114
  - Euclidean distance, 146
  - infinite norm, 145
  - inner product, 114
    - matrix notation, 123
  - L1-norm, 145
  - L2-norm, 145
  - magnitude, 112
  - Mahalanobis distance, 148
  - Manhattan distance, 146
  - norm, 144
  - orthogonal, 115
  - outer product, 117
    - matrix notation, 123
  - projection, 116
  - right-hand rule, 119
  - row, 104
  - Taxicab distance, 146
  - transpose, 113
  - unit, 112

**W**

- Welch's t-test, 95
- Wilcoxon rank sum test, 99