

# INDEX

## A

adjacency matrices  
centrality, 27, 30, 33–35  
directed and undirected  
networks, 27  
disease spread tracking, 67, 69  
persistent homology, 91–92  
spectral theory, 49–50  
weighted networks, 27  
Akaike information criterion (AIC),  
137, 140  
algebraic connectivity, 50, 51  
alpha (attenuation parameter), 35, 39  
alpha centrality. *See* Katz centrality  
authority centrality  
defined, 35  
measuring in social networks, 39–41  
averaged perceptron tagger, 183–184

## B

Barabási-Albert model, 52  
basis (Hamel basis), 133  
BERT (Bidirectional Encoder  
Representations from  
Transformers), 189–190  
beta, 70  
Betti numbers, 85–86  
defined, 85  
Euler characteristic, 87  
examples of, 85–86  
persistent homology, 88–89  
subgroup mining, 156–157  
validating measurement tools, 161  
betweenness of vertices  
applications of, 32–33  
bridges, 64  
community mining, 60  
disease spread tracking, 68

graph filtration, 79  
measuring in social networks, 37,  
38, 41, 42  
overview of, 32–33  
predictions with social media  
network metrics, 57, 59  
topological data analysis, 194  
Bidirectional Encoder Representations  
from Transformers (BERT),  
189–190  
binomial distributions  
dispersion, 137  
entropy, 107–110  
Bonacich centrality. *See* Katz centrality  
bridges  
betweenness, 38, 64, 68  
disease spread tracking, 68  
predictions with social media  
network metrics, 56–57  
walktrap algorithm, 61  
browseVignettes(), xxii

## C

calculate\_homology(), 157  
Canberra distance, 101–102, 103, 118  
Čech complexes, 82  
centrality, 29–42  
applications of, 30  
applying clustering to social media  
dataset, 60  
authority centrality, 35  
betweenness of vertices, 32–33  
closeness of vertices, 31  
defined, 30  
degree of vertices, 30–31  
disease spread tracking, 68  
distance and, 24  
eigenvector centrality, 33–34

- centrality (*continued*)
    - graph filtration, 77–79
    - hub centrality, 35
    - Katz centrality, 35
    - measuring in example social network, 36–42
    - PageRank centrality, 34–35
    - predictions with social media network metrics, 56–59
    - spectral theory, 27, 49
    - topological data analysis, 194
    - topological dimension, 82
  - Chebyshev distance, 101, 116
  - choice ranking comparison, 149–152
    - HodgeRank, 152
    - missing information, 150–151
    - no consistent preferences, 151
    - overview of, 149–150
    - preference loops, 150–151
  - circuit-centric quantum classifiers, 203–204
  - classification and classifiers
    - bot account detection, 64, 66
    - convolutional neural network classifiers, 110
    - curse of dimensionality, 16–17
    - decision boundaries, 10–11
    - defined, 2
    - homology, 85–86
    - homotopy, 167, 169
    - image classification, 18–20, 200–204
    - logistic regression classifiers, 16–17
    - metric geometry, 116–119
    - overfitting and underfitting, 13
    - overview of, 10–11
    - poetry analysis project, 186, 189–190
    - predicting edge formation, 59
    - quantum classifiers, 203–204
    - supervised classifiers, 59, 65
    - support vector machine classifiers, 103
  - closed triangles, 43, 49
  - closeness of vertices
    - measuring in social networks, 37, 38
    - overview of, 31
  - CNNs (convolutional neural networks), 18–19, 110, 202–204
  - cohomology, 141, 146
  - community mining (clustering vertices), 59–64
    - evaluating quality outcome of clusters, 61–62
    - exploring networks with random walks, 61
    - overview of, 59–60
    - running clustering algorithms, 62–64
    - springlass clustering, 62
  - conditional Fisher information, 134–135
  - conditional Rao score, 135
  - connected components
    - graph Laplacian, 50, 51
    - homology, 85–86, 89
    - random walk algorithms, 34
    - subgroup mining, 156
  - convex optimization problems, 148
  - convolutional neural networks (CNNs), 18–19, 110, 202–204
  - COVID-19 pandemic, 72–73, 127
  - curl flow, 152
  - curse of dimensionality, 7, 14, 95
    - geometric perspective, 17
    - overview of, 13–17
    - perturbed points in Euclidean space, 14–16
- D**
- data geometry, 1–21
    - machine learning, 2–4
      - matching algorithms, 4
      - supervised learning, 2–3
      - unsupervised learning, 3
    - structured data, 4–17
      - dummy variables, 5–7
      - numerical spreadsheets, 8–10
      - supervised learning, 10–17
    - unstructured data, 17–21
      - image data, 18–20
      - network data, 17–18
      - text data, 20–21
  - data integrity, 4
  - data points
    - in structured data, 4, 7–9
    - supervised learning, 9, 11, 14, 17
    - unsupervised learning, 3

- data science geometry, 95–129
  - distance metrics, 96–116
    - entropy, 107–110
    - norm-based distance metrics, 99–105
    - shape comparison, 110–116
    - small dataset simulation, 98–99
    - Wasserstein distance, 105–107
  - fractals, 125–129
  - $k$ -nearest neighbors with metric geometry, 116–119
  - manifold learning, 119–125
    - Isomap, 121–122
    - locally linear embedding, 122–124
    - multidimensional scaling, 120–122
    - $t$ -distributed stochastic neighbor embedding, 124–125
- decision boundaries
  - classification, 10–11
  - overfitting, 13
- decision trees
  - decision boundaries, 10, 11
  - overfitting, 13
- deep learning
  - convolutional neural networks, 18, 202–203
  - defined, 4
  - geometric, 18
  - Riemannian manifolds, 18
  - vector embeddings, 20–21
- degree of networks, 48–49
- degree of vertices (degree centrality). *See also* Katz centrality
  - applications of, 31
  - community mining, 60
  - Forman–Ricci curvature, 46
  - graph filtration, 76, 78–79
  - graph Laplacian, 50
  - in-degree and out-degree, 30
  - $k$ -means clustering, 60, 61
  - limitations of, 31
  - measuring in social networks, 41
  - overview of, 30–31
  - scale-free graphs, 52
  - topological dimension, 82–84
  - triadic closure, 43
- dendrograms, 89, 107, 158–160
- density of networks
  - disease spread tracking, 69, 71
  - graph filtration, 77
  - overview of, 48–49
- dependent variables
  - defined, 2
  - dummy variables, 5
  - image classification, 203
  - link functions, 135
  - regression, 11, 12
  - supervised learning, 2–3
  - vertex centrality metrics, 56, 58
- dgLARS algorithm, 133–140
  - credit default prediction, 138–140
  - cross-validated vs. non-cross-validated, 136–140
  - depression prediction, 136–138
  - overview of, 133–136
  - poetry analysis project, 186
  - risk propensity measurement, 134–135
- dgLARS package, 136
- diameter of networks
  - graph filtration, 79–80
  - network comparison, 65–66
  - overview of, 49
- differential geometry, 88. *See also* dgLARS algorithm
- differential geometry least angle regression algorithm. *See* dgLARS algorithm
- Dijkstra’s algorithm, 199
- dimensionality
  - curse of, 7, 13–17, 95
  - defined, 14
  - reduction of, 95, 119–120, 184
  - unsupervised learning, 3
- directed networks, 19
  - applications of, 26
  - authority centrality, 39–40
  - converting undirected to, 39–40
  - defined, 26
  - degree of vertices, 30
  - edges, 28

- directed networks (*continued*)
  - eigenvector centrality, 34
  - hub centrality, 39–40
  - interconnectivity of networks, 48
  - networks in R, 26–27
  - PageRank algorithm, 33
  - Twitter, 17, 26
- disaster logistics planning, 142–146
- discrete exterior derivatives, 140–146
  - cohomology, 141, 146
  - differential forms, 141
  - disaster logistics planning, 142–146
  - engineering problems, 146
  - overview of, 140, 141
  - social network analysis, 141–142
- `dist()`, 99, 101, 104
- distance metrics, 96–116
  - entropy, 107–110
  - norm-based distance metrics, 99–105
  - overview of, 96–98
  - shape comparison, 110–116
  - small dataset simulation, 98–99
  - Wasserstein distance, 105–107
- distributed computing, 194–195
- diversity of vertices, 42
- Dow Jones Industrial Average (DJIA), 127–128
- dummy variables, 5–7
  - categorical variables, 5–6
  - geometry of, 5–7
  - multicollinearity, 7
- D-Wave, 197
- E**
- Ebola outbreak, 29
- eccentricity of vertices
  - diameter and, 49
  - graph filtration, 80
  - overview of, 45
  - radius and, 49
- edge lists, 26–27
- edges
  - adjacent, 28
  - closeness of vertices, 31
  - degree of vertices, 30
  - density of networks, 48–49
  - depiction of, 25
  - directed and undirected networks, 17, 26
  - disease spread tracking, 67–69, 70
  - diversity of vertices, 42
  - Erdős-Renyi graphs, 51–52
  - Euler characteristic, 87
  - Forman–Ricci curvature, 46–47
  - graph filtration, 76–78, 80
  - intracommunity and intercommunity edges, 61
  - link prediction in social media, 58–59
  - network comparison, 65
  - overview of, 25
  - path length, 28
  - weighted and unweighted networks, 28–29
- efficiency of networks, 49
- efficiency of vertices, 44–45
- `eigen()`, 50
- eigenvalues, 33, 49–50
- eigenvectors, 33, 49–50
  - eigenvector centrality, 33–39
    - authority and hubness, 35
    - Katz centrality and, 36
    - measuring in social networks, 38–39
    - overview of, 33–34
    - PageRank centrality and, 34–35
- elastic net regression, 101
- EM algorithm, 171
- entities
  - named entity recognition, 180
  - spread of, 66–68
  - vertices and edges, 25
- entropy, 107–110
  - diversity of vertices, 42
  - relative, 135
  - Shannon entropy, 42
- epidemiology
  - centrality, 30
  - disease spread tracking, 67–74
  - spectral radius, 50
- Erdős-Renyi graphs, 51–52
  - network comparison, 65–66
  - persistent homology, 90–93

- Euclidean distance
  - course of dimensionality, 14, 16
  - $k$ -nearest neighbors, 118–119, 186
  - multidimensional scaling, 121, 122
  - network distance and, 29
  - norm-based distance metrics, 99–103
  - spreadsheet geometry, 9
- Euclidean vector space
  - course of dimensionality, 15, 16
  - defined, 8
  - manifolds, 119
  - multidimensional scaling, 120–122
  - shape comparison, 113–116
  - spreadsheet geometry, 8
  - tangent spaces, 133
  - vector embeddings, 21
- Euler characteristic, 87–88
  - Betti numbers, 87
  - Gauss-Bonnet theorem, 88
  - maximal cliques, 87
  - negative, 87
  - simplicial complexes, 87
- expectation-maximization (EM)
  - algorithm, 172
- F**
- Facebook
  - bot account detection, 18, 24
  - degree of vertices, 30
  - global network metrics, 47
  - link prediction, 58
  - network distance, 24
  - text search, 20
  - undirected networks, 17, 26
- fast greedy clustering, 61–64
- feature importance, 3
- filtration
  - graph filtration, 76–81
  - network filtration, 75–94
- Fisher information, 134–135
- flag complexes, 82–83
- fMRI. *See* functional magnetic resonance imaging
- Forman–Ricci curvature
  - differential geometry, 88
  - disrupting communication and disease spread, 72–73
  - overview of, 45–47
  - stock market change point detection, 129
- Forman–Ricci flow, 73–74
- fractals, 125–129
- Fréchet distance, 111–112
- functional magnetic resonance imaging
  - network comparison, 64, 66
  - persistent homology, 90–93
- G**
- gamma, 70
- gate-based circuits, 196–197
- Gauss-Bonnet theorem, 88
- Gaussian distribution, 172–173
- Gaussian noise, 14
- gcd(), 199–200
- genomics, 17, 88, 136
  - datasets, 86, 101, 119
- geodesics, 33, 44, 121
  - tangent spaces and, 96–97
- geometric deep learning, 18
- geospatial data, 8–9, 9
- gerrymandering, 24
- global network metrics, 47–51
  - graph filtration, 79
  - interconnectivity of networks, 48–49
  - network comparison, 93
  - spectral measures of networks, 49–51
  - spreading processes on networks, 49
- Google
  - image search, 202–203
  - PageRank algorithm, 33
  - PageRank centrality, 34–35
  - text search, 20
- GPT-3, 189
- gradient descent, 169–171
- gradient flow, 152
- graph diameter, 79–80
- graph filtration, 76–81
  - brain imaging studies, 80
  - degree centrality, 78–79
  - graph diameter, 79–80
- graph Laplacian, 50–51
- graph theory, 24, 195, 198–199

greatest common denominator, 199–200  
greedy algorithms, 61–64  
Gromov-Hausdorff distance,  
    113–116, 160  
gromovlab package, 114

## H

Hamel basis, 133  
Hamming distance, 163–164  
harmonic flow, 152  
Hausdorff distance, 113, 160  
hclust(), 156  
heatmaps, 11, 89  
help(), xxii  
hierarchical clustering, 3, 89,  
    156–158, 163  
Hodge-Helmholtz decomposition, 152  
HodgeRank, 152  
homology, 85–94  
    Betti numbers, 85–86  
    cohomology, 140–141, 146  
    defined, 85  
    differential geometry, 88  
    Euler characteristic, 87–88  
    persistent homology, 88–89, 129,  
        159, 195  
    measurement validation  
        and, 160–161  
    network comparison and,  
        89–94, 155–156  
    subgroup mining and,  
        156–157, 159, 162  
homotopic Fréchet distance, 111  
homotopy algorithms, 167–177  
    comparing, 173  
    homotopic, defined, 167  
    homotopy-based regression, 169–174  
    logistic regression vs. homotopy-  
        based regression, 174–176  
    overview of, 167–168, 169  
hub centrality  
    community mining, 60  
    defined, 35  
    graph filtration, 77  
    measuring in social networks,  
        39–42  
    unsupervised learning, 60

hyperparameters  
    classification, 11  
    defined, 3  
    overfitting, 13  
    regression, 12

## I

IBM, 197  
igraph library, 27, 29–30, 35, 43, 52, 85,  
    87, 90, 165  
    cluster\_edge\_betweenness(), 64  
    default value, 39  
    eccentricity(), 49  
    edge\_density(), 48, 69  
    efficiency(), 49  
    sample\_gnp(), 51  
    sample\_pa(), 52  
    sample\_smallworld(), 52  
    sir(), 70  
    spectrum(), 50  
    transitivity(), 49  
image classification  
    convolutional neural networks,  
        18–20  
    quantum computing approaches,  
        200–204  
image data  
    convolutional neural networks,  
        18–19, 20  
    Forman–Ricci flow, 73–74  
    overview of, 18–20  
    persistent homology, 88–89  
in-degree, 30  
independent variables  
    decision trees, 11  
    defined, 2  
    dgLARS algorithm, 135  
    dimensionality, 14, 101  
    dummy variables, 5, 7  
    geometric deep learning, 18  
    image classification, 19, 203  
    multicollinearity, 7  
    supervised learning, 2–3  
    unsupervised learning, 3  
    vertex centrality metrics, 57–58  
instances. *See* data points  
interconnectivity of networks, 40, 44,  
    47–49, 52

inverse Hamming distance, 164  
Isomap, 121–122  
isometric embedding, 113, 116

## K

Katz centrality  
  eigenvector centrality and, 35  
  measuring in social networks, 39  
  overview of, 35

k-means clustering, 3  
  community mining, 59–60  
  vs. Mapper algorithm, 161, 163

*k*-nearest neighbors (*k*-NN), 2–3  
  decision boundaries, 10, 11  
  dummy variables, 7  
  metric geometry, 116–119  
  overfitting, 13  
  poetry analysis project, 186  
  regression, 11–12

knnGarden package, 117

KONECT Windsurfer Network, 69–71

Kullback-Leibler divergence  
  dgLARS algorithm, 135  
  entropy, 108–110  
  t-distributed stochastic neighbor  
    embedding, 124

## L

Lasso algorithm  
  homotopy-based optimization,  
    172, 174–176  
  Lasso regression, 101  
  poetry analysis project, 190

lasso2 package, 172

linear dependence, 7

linear regression, 2–3  
  dgLARS algorithm, 136, 138  
  making predictions with social  
    media network metrics, 57  
  multicollinearity, 7  
  supervised regression, 12  
  vs. homotopy-based regression,  
    173–174, 176

link functions, 136

link prediction, 58–59

locally linear embedding (LLE), 122–124

local optima, 169–172, 174, 176

logistic regression, 2  
  curse of dimensionality, 16  
  decision boundaries, 10, 11  
  dgLARS algorithm, 140  
  link functions, 136  
  multicollinearity, 7  
  overfitting, 13  
  vs. homotopy-based regression,  
    174–176

Louvain clustering, 62–64

## M

machine learning categories, 2–4  
  matching algorithms, 4  
  supervised learning, 2–3  
  unsupervised learning, 3

mahalanobis(), 103

Mahalanobis distance, 103–104, 105

Manhattan distance, 100–102

*k*-nearest neighbors, 116, 118–119  
  multidimensional scaling, 121, 122  
  subgroup mining, 156–157

manifold hypothesis, 95

manifold learning, 119–125  
  Isomap, 121–122  
  locally linear embedding, 122–124  
  multidimensional scaling, 120–122  
  vs. principal component  
    analysis, 119

  t-distributed stochastic neighbor  
    embedding, 124–125

manifolds

  defined, 8  
  distance metrics, 96, 98, 194  
  Gauss-Bonnet theorem, 88  
  homology, 85, 88  
  Riemannian manifolds, 18  
  tangent spaces, 132–133

Mapper algorithm, 161–166

  stepping through, 162–163  
  using TDAmapper to find cluster  
    structures in data, 163–166

matching algorithms, 4

Matlab, 152

maximal cliques, 82–84

  disaster logistics planning, 144  
  Euler characteristic, 87  
  quantum network algorithms, 197

- MDS (multidimensional scaling), 120–122
- median(), 70
- metric geometry, 98
  - fractals, 125–129
  - $k$ -nearest neighbors, 116–119
  - manifold learning, 119–125
- Minkowski distance, 101–102, 121
- MNIST dataset, 204
- model fit
  - dgLARS algorithm, 137–138
  - homotopy-based regression, 172
  - nonlinearity, 147, 149
- modularity, 61–64
- Morse functions, 162
- multicollinearity, 7, 133
- multicore approaches, 193–195
- multidimensional scaling (MDS), 120–122

## N

- named entity recognition, 180
- natural language processing
  - pipelines, 180–181
  - topology-based tools, 188–191
- network analysis, 55–74
  - spread analysis, 66–74
    - disease spread tracking
      - between towns, 67–69
    - disease spread tracking between windsurfers, 69–72
    - disrupting communication and disease spread, 72–74
  - supervised learning, 56–59
    - diary entry prediction in social media, 56–58
    - link prediction in social media, 58–59
  - unsupervised learning, 59–64
    - applying clustering to the social media dataset, 59–60
    - community mining, 61–64
- network comparison, 64–66
- network depth, 31, 33
- network distance, 28–29
  - applications of, 24, 28–29
  - defined, 28

- link prediction in social media, 59
- persistent homology, 91–93
  - weighted and unweighted networks, 28
- network filtration, 75–94
  - graph filtration, 76–81
  - homology, 85–94
    - Betti numbers, 85–86
    - Euler characteristic, 87–88
  - persistent homology, 88–89
    - comparison with, 90–94
  - simplicial complexes, 81–85
- network geometry, 23–54
  - directed and undirected networks, 18
  - global network metrics, 47–51
    - interconnectivity of a network, 48–49
    - spectral measures of a network, 49–51
    - spreading processes on a network, 49
- models for real-world behavior, 51–53
  - Erdős-Renyi graphs, 51–52
  - scale-free graphs, 51–52
  - Watts-Strogatz graphs, 52–53
- network science, 24–25
- network theory, 25–29
  - directed networks, 26
  - networks in  $\mathbb{R}$ , 26–27
  - paths and network distance, 28–29
- overview of, 17–18
- Riemannian manifolds, 18
- vertex metrics, 30–47
  - centrality, 30–42
  - diversity of vertices, 42
  - eccentricity of vertices, 45
  - efficiency of vertices, 44–45
  - Forman–Ricci curvature, 45–47
  - triadic closure, 43–44
- networks, defined, 8
- neural networks, 2
  - convolutional, 18–19, 202–204
  - decision boundaries, 10, 11, 13
  - homotopy-based optimization, 172



- quantum, 203–204
- topology-based NLP tools, 189
- neuroscience and brain imaging
  - graph filtration, 80
  - network comparison, 64
  - persistent homology, 90–93
- NLP. *See* natural language processing
- NLTK toolkit, 181, 183–184
- nodes. *See* vertices
- nonconvex objects, 147–148
- nonlinear algebra, 146–149
  - convex optimization problems, 148
  - nonconvex objects, 147–148
  - numerical algebraic geometry, 147–149
  - vs. linear algebra, 146–147
- nonlinear spaces, 132–140
  - dgLARS algorithm, 133–140
    - credit default prediction, 138–140
    - depression prediction, 136–138
    - overview of, 133–136
  - tangent spaces, 132–133, 135
- nonselected predictors, 135
- norms, defined, 99
- numerical algebraic geometry, 147–149
- numerical spreadsheets. *See* spreadsheets

## O

- observations. *See* data points
- open triangles, 43
- out-degree, 30
- outlier detection, 24
  - network comparison, 66
  - stealth outliers, 104
  - subgroup mining, 159
- overfitting
  - curse of dimensionality, 14, 16–17
  - overview of, 13

## P

- PageRank algorithm, 33, 198
- PageRank centrality, 34–35
  - community mining, 60
  - link prediction in social media, 59

- measuring in social networks, 38–39, 41, 42
- paths, defined, 28
- PCA (principal component analysis), 3, 119
- perplexity, 124–125
- Perron-Frobenius theorem, 34
- persistence diagrams
  - measurement tool validation, 159–161
  - persistent homology, 89, 91–93
  - poetry analysis project, 187–188
  - shape comparison, 110
  - subgroup mining, 157
- persistent homology, 88, 155–159
  - measurement tool validation, 160–162
  - multicore approaches, 195
  - network comparison, 90–94
  - outlier detection, 159
  - overview of, 88–89
  - stock market change point detection, 129
  - subgroup mining, 156–159
- PET (positron emission tomography), 64, 90–93
- philtropy package, 108
- `plot_persist()`, 157
- poetry analysis project, 180–188
  - analysis in R, 184–188
  - forms of poetry, 181–182
  - natural language processing pipeline, 180–181
  - normalizing vectors, 184
  - tagging parts of speech, 183–184
  - tokenizing text data, 183
  - topology-based NLP tools, 188
- point clouds, 82, 88–89, 146, 156–157, 162–163
- Poisson distribution, 56–57
- Poisson regression, 57, 58
- polynomials, 147–148
- positron emission tomography, 64, 90–93
- predictors. *See* independent variables
- pretrained transformer models, 189
- principal component analysis, 3, 119

- probability density functions, 106–108, 109
- probability distributions
  - entropy, 107–110
  - t-distributed stochastic neighbor embedding, 124
  - Wasserstein distance, 105–107
- propagation analysis. *See* spread analysis
- Pythagorean distance, 100. *See also* Euclidean distance
- Python
  - BERT model, 189
  - distributed computing, 194
  - help resources for, xxiii
  - natural language processing, 180–181
  - poetry analysis project, 183–184
  - quantum computing, 196–199

**Q**

- quantum algorithms, 197–200, 204
- quantum annealing, 197
- quantum approximation optimization algorithms (QAOA), 198
- quantum classifiers, 203–204
- quantum computing approaches, 193–205
  - image classifiers, 200–204
  - quantum algorithm development, 199–200
  - quantum network algorithms, 197–199
  - qubit-based model, 196–197
  - qumodes-based model, 197
- quantum maximum flow and minimum cut algorithms, 197–198
- QuantumOps package, 198–200, 203–204
- qubits
  - circuit-centric quantum classifiers, 203
  - quantum network algorithms, 198
  - qubit-based model, 196–197
  - qumodes-based model, 197
- Quora dataset, 136, 156, 160, 163, 166, 174

**R**

- R (programming language)
  - downloading, xxi–xxii
  - help resources for, xxii
  - installing, xxii
  - installing packages, xxii
  - vignettes, xxii–xxiii
- radius of networks, 35, 49–50
- random forests, 2
  - decision boundaries, 11
  - regression, 11–12
- random walk algorithms
  - diversity of vertices, 42
  - eigenvector centrality, 34, 38
  - link prediction, 59
  - map networks, 24
  - PageRank centrality, 35, 38–39
  - quantum algorithms, 198
  - walktrap algorithm, 61–64
- redundant predictors, 133, 135
- regex tokenizer, 183
- regression
  - defined, 2
  - dummy variables, 5–6
  - elastic net regression, 101
  - homotopy-based regression, 169–176
  - journal ranking, 66
  - linear regression, 2–3, 12
    - dgLARS algorithm, 136, 138
    - making predictions with social media network metrics, 57
    - multicollinearity, 7
    - vs. homotopy-based regression, 173–174, 176
  - logistic regression, 2
    - curse of dimensionality, 16
    - decision boundaries, 10, 11
    - dgLARS algorithm, 140
    - link functions, 136
    - multicollinearity, 7
    - overfitting, 13
    - vs. homotopy-based regression, 174–176
- overview of, 11–12
- Poisson regression, 57, 58

- reinforcement learning, 4
- Ricci curvature, 45. *See also*
  - Forman–Ricci curvature
- Riemannian manifolds, 18
- Rigetti, 197
- RStudio, xxii
- S**
- SBERT (sentence-based Bidirectional Encoder Representations from Transformers), 189
- scale-free graphs, 52–53
  - network comparison, 65–66
  - persistent homology, 90–93
- scatterplots, 11, 100, 119, 171
- selected predictors, 135
- Shannon entropy, 42
- shape comparison, 110–116
  - Fréchet distance, 111–112
  - Gromov-Hausdorff distance, 113–116
  - Hausdorff distance, 113
- simplicial complexes, 81–85
  - Čech complexes, 82
  - cochains, 141
  - Euler characteristic, 87
  - filtering, 82
  - flag complexes, 82–83
  - Mapper algorithm, 162
  - maximal cliques, 82–84
  - overview of, 81–82, 84
  - persistent homology, 88–89, 156–157
  - topological dimension, 82–84
  - Vietoris-Rips complexes, 82
- simulated annealing, 62
- single-linkage hierarchical clustering, 89, 156, 163
- SIR model. *See* susceptible-infected-resistant model
- SMOTE (Synthetic Minority Oversampling Technique), 8
- social networks
  - algebraic connectivity, 50
  - bot account detection, 18, 24, 64, 66
  - centrality, 30–42
  - clustering vertices, 59–64
  - diary entry prediction, 56–58
  - directed and undirected networks, 17–18
  - discrete exterior derivatives, 141–142
  - Forman–Ricci curvature, 46, 47
  - graph filtration, 76–80
  - influencers, 24, 30–31
  - link prediction, 58–59
  - network depth, 33
  - simplicial complexes, 82–83, 84
  - spread of misinformation, 9, 66–67
  - subgroup mining, 156–157
  - transitivity, 43, 44
  - triadic closure, 43
  - vertex vs. global metrics, 47
  - Watts-Strogatz graphs, 52
- spectral gap of networks, 50
- spectral measures of networks, 49–51
- spectral radius, 35, 49–50, 79
- spinglass algorithms, 62
- spinglass clustering, 62–64
- spread analysis, 66–74
  - disease spread tracking
    - between towns, 67–69
    - between windsurfers, 69–72
  - disrupting communication and disease spread, 72–74
- spreading processes on networks, 49
- spreadsheets
  - adjacency matrices, 27
  - classification, 10
  - dummy variables, 5–7
  - Euclidean vector space, 8
  - geometry of, 8–10
  - geospatial data, 8–9
  - structured data, 4
- stats package, 99
- stock market change point detection, 127–129
- strength (weighted degree) of vertices, 30, 82
- structured data, 4–17
  - defined, 4
  - dummy variables, 5–7
  - numerical spreadsheets, 8–10

- structured data (*continued*)
    - supervised learning, 10–17
      - classification, 10–11
      - curse of dimensionality, 14–17
      - overfitting, 13
      - regression, 11–12
  - subgroup mining
    - Mapper algorithm, 161–166
    - persistent homology, 156–159
  - supervised learning, 10–17
    - algorithm viewed as function, 2–3
    - classification, 10–11
    - combining with unsupervised learning, 3
    - curse of dimensionality, 14–17
    - overfitting, 13
    - overview of, 2–3
    - prediction
      - diary entry prediction in social media, 56–58
      - link prediction in social media, 58–59
    - regression, 11–12
    - training and testing data, 3
  - support vector machines, 2, 172
  - susceptible-infected-resistant model
    - defined, 68
    - disease spread tracking
      - between towns, 67–69
      - windsurfers, 69–72
    - disrupting communication and disease spread, 72–74
  - Synthetic Minority Oversampling Technique (SMOTE), 8
- T**
- tabular data. *See* structured data
  - tangent lines, 96, 132–133
  - tangent planes, 96, 132, 133
  - tangent spaces, 96, 124, 131–135, 149
    - Euclidean space and, 113
    - geodesics and, 96
    - linear algebra and, 133
    - nonselected predictors, 135
    - redundant predictors, 135
    - selected predictors, 135
  - targets. *See* dependent variables
  - TDA. *See* topological data analysis
  - TDAmapper package, 163–166
  - TDAstats package, 156–157, 161
  - t-distributed stochastic neighbor embedding (t-SNE), 124–125, 185–186, 189–190
  - test error
    - curse of dimensionality, 14
    - overfitting, 13
  - testing data
    - defined, 3
    - overfitting, 13
  - text data
    - overview of, 20–21
    - poetry analysis project, 179–191
    - vector embeddings, 20–21
  - topological data analysis, 159–166
    - graph filtration, 76
    - Mapper algorithm, 161–166
    - measurement tool validation, 159–161
    - multicore approaches, 194–195
    - persistent homology, 155–159
    - subgroup mining, 156–159
  - topological dimension, 82–85
  - tori, 86, 168
  - training data, defined, 3
  - training error, 13
  - transitivity
    - community mining, 60
    - global network metrics, 49
    - vertex metrics, 43–45
  - Treebank tokenizer, 183
  - triadic closure, 43–44, 79
  - triangle inequality condition, 101
  - TSdist package, 111
  - t-SNE (t-distributed stochastic neighbor embedding), 124–125, 185–186, 189, 190
  - Twitter, 17–18, 26, 30–31
- U**
- UCI credit default dataset, 138–139
  - underfitting, 13–14

- undirected networks, 19
  - converting to directed, 39–40
  - defined, 26
  - Facebook, 17, 26
  - interconnectivity of networks, 48
  - networks in R, 27
- unstructured data, 17–21
  - image data, 18–20
  - network data, 17–18
  - text data, 20–21
- unsupervised learning, 59
  - clustering vertices, 59–64
    - evaluating quality outcome of clusters, 61–62
    - exploring networks with random walks, 61
    - overview of, 59–60
    - running clustering algorithms, 62–64
    - spinglass clustering, 62
  - combining with supervised learning, 3
  - overview of, 3
- unweighted networks, 27–28, 30–32, 34–35

## V

- vector embeddings, 20–21
- vertex metrics, 30–47
  - centrality, 30–42
    - authority centrality, 35
    - betweenness of vertices, 32–33
    - closeness of vertices, 31
    - degree of vertices, 30–31
    - eigenvector centrality, 33–34
    - hub centrality, 35
    - Katz centrality, 35
    - measuring centrality in
      - example social network, 36–42
      - PageRank centrality, 34–35
    - community mining, 59–60
    - defined, 47
    - diversity of vertices, 42
    - eccentricity of vertices, 45
    - efficiency of vertices, 44–45
    - Forman–Ricci curvature, 45–47

- prediction
  - diary entry prediction in social media, 56–58
  - link prediction in social media, 58–59
- triadic closure, 43–44

- vertices
  - betweenness of, 32–33, 64
  - closeness of, 31
  - community mining, 59–64
  - degree of, 30–31
  - depiction of, 25
  - directed and undirected networks, 26
  - distance between, 28
  - diversity of, 42
  - eccentricity of, 45, 49
  - efficiency of, 44–45
  - neighboring, 28
  - overview of, 25
  - strength of, 30
- Vietoris–Rips complexes, 82

## W

- walktrap algorithm, 61–64
- Wasserstein distance, 93, 105–107
- Watts–Strogatz graphs, 52–53, 61
  - network comparison, 65–66
  - persistent homology, 90–93
- wave functions, 197
- weighted degree (strength) of vertices, 30, 82
- weighted networks
  - adjacency matrices, 27
  - degree of vertices, 30
  - disease spread tracking, 67–69
  - diversity of vertices, 42
  - eigenvector centrality, 34
  - graph filtration, 76–81, 84–85
  - PageRank centrality, 35
  - paths and network distance, 28–29
  - simplicial complexes, 84–85
  - strength of vertices, 30

## X

- Xanadu, 197

## Y

- YouTube, 4