Index

Activation function, 29–32, 37, 60
Active learning, 173–174
Ada. Boost. M1 algorithm, 166–167
Area under curve (AUC), 13–14, 118–119
Artificial intelligence (AI), 1, 2
Artificial neural network (ANN), 26
Augmented reality (AR), 53
Automatic interaction detection method (AID method), 139–143
Automatic relevance detection (ARD), 49
Average linkage, 157–158

Backpropagation
  cost functions and training of neural networks using, 38–40
  equations, 62
Bagging, 158–159, 161–165, 169
  regularization through, 78
Basis expansion, 58–59
Basis function(s), 58–59
  regression, 28, 31
Batch
  gradient descent, 8–9
  size, 8–9
Bayesian approach, 26
Bayesian neural networks, 49
Bias, 2, 69
  bias-variance tradeoff, 68–70
Binary choice targeting model, 72
Binary classification, 3
Boosting, 158–159, 165–169
Bootstrap(ping), 158–161
  aggregation, 159
Business-to-business setting (B2B setting), 118–119
Business-to-customer setting (B2C setting), 118–119

Caravan insurance, 176
Chain-rule of calculus, 39, 63
Chaotic time series, 119
Chi Squared automatic interaction detection method (CHAID method), 139–143
Chi-squared statistic, 140
Choice rules, 121–122
Choice-based conjoint analysis (CBC analysis), 122
Churn
  modeling, 118–119
  prediction, 54–57
Classification
  models, 2–3
  NN for, 37–38
  performance assessment for classification tasks, 9–19
  trees, 150–155
Classification and regression trees (CART), 143–155, 175–176
Classifier, 93
Clustering models, 156
Coefficients, 2
Collaborative-based recommendation system, 116–117
Complete linkage, 157–158
Composite functions, 36
Computational learning theory, 85–86
Confusion matrix, 12–13
Conjoint analysis, 124–125
  methodology, 116
Connection weights, feature importance based on, 45–47

Consumer choice modeling, 121
Content-based recommendation system, 116–117
Convolutional neural networks (CNN), 2
Cosine similarity kernel, 130
Cost complexity
  criterion, 181
  pruning, 149–150
Cost function, 9
  and training of machine learning models, 3–4
Cross-entropy cost, 4, 6, 19–20, 38, 41, 151–152
Cross-validation, 70–72, 80
Cumulative response curve, 15–17
"Customer-focused" approach of marketing, 116

Decision trees, 155
  applications in marketing and sales, 171–176
  bootstrapping, bagging, and boosting, 158–169
  case studies, 176–179
  decision tree-based methods, 139–140
  early evolution of, 139–143
  random forest, 169–171
  and segmentation, 155–158
Default classification rule, 12–13
Dendrograms, 156, 158
Dependent variable, 2
Depth of network, 36
Descent, 9
Direct Marketing Educational Foundation (DMEF), 117–118, 173–174
Directional derivatives, 8
Distance to city center (DCC), 58
Dot product, 59, 91, 128

"Earnings before tax-to-equity ratio", 173–174
Empirical distribution, 5
Ensemble methods, regularization through, 78
Ensemble random forest approach, 175–176
Euclidean distance, 156–157
Euclidean norm, 91–92
Evolutionary local selection algorithm (ELSA), 52
Example-dependent costs, 175–176
Expectation, 69
Expected test error, 71–72
Explanatory variable, 2

Feature importance
  based on connection weights, 45–47
  based on partial derivatives, 49
  measurement, 42–49
Feature selection, 75
Feature space, 94, 129, 143
First Order Conditions (FOCs), 132
Fivefold cross-validation, 71
Forward stagewise additive modeling process, 165–166

Gainsight and Survey Monkey company, 54
Gaussian distribution, 62
Gaussian errors, 38
Generalizability, 73
Generalization error, 9–10, 68
Gini coefficient, 17–19
Gini index, 151–152
Goodness-of-fit measure, 149–150
Gradient, 61
  boosting, 168
  with cross-entropy, 63
  descent, 9, 61
  gradient-based learning, 6–9
Gram matrix, 130
Greedy algorithm, 147–149, 155

Hard choices, 116–117
Hidden nodes, 31–32, 59–60

Hierarchical Bayesian method (HB method), 116
Hierarchical clustering, 156
Hit rate, 11–12
Hyperparameters, 66, 167
Hyperplanes, 88
  margin between classes, 99–100
  maximal margin classification, 101–106
  optimal separating hyperplane, 99–106
  separating, 88–89

Independent variable, 2
Inner product, 59, 91, 128
Intel's RealSense Vision Processor, 53
Internet Movie Database (IMDB), 116–117
"Inverted U" shape, 33–34
Irreducible error, 69

K-fold cross validation, 71
Karush–Kuhn–Tucker conditions, 132
Kernel(s), 94–98
  kernel-based nonlinear classifier, 114
  in machine learning, 90–99
  matrix, 130
  as measures of similarity, 91–94
  trick, 98–99
k^{th} degree polynomial kernel, 130

L_1 regularization, 74–75
  as constrained optimization problems, 75–76
  weight decay in, 81
L_2 regularization, 73–74
  as constrained optimization problems, 75–76
  weight decay in, 80–81
Lagrange multipliers, 131–132
Lasso, 73
Latitude of acceptance rule (LOA rule), 52, 121–122
Law of parsimony, 72
Lead qualification and scoring models, 52
Learning rate, 66
  with cross-entropy function, 63
  parameter, 7–8
Learning slowdown, 63
Leave-out-one cross-validation (LOOCV), 71
Lift chart, 15–17
Linear activation function for continuous regression outputs, 40, 62–63
Linear regression model, 2–3
Log odds, 86, 127
  ratio, 3
Log-likelihood, 19
Logistic regression, 3, 86
Logit leaf model (LLM), 175–176

Machine learning, 1–2
  implementation, 1
  industry applications, 1
  kernels in, 90–99
Margin, 104, 130
  width, 107
Maximal margin classification, 101–106
Maximum likelihood estimation (MLE), 4–6, 38
Maximum likelihood estimator, 19, 60–61
Mean squared error (MSE), 10, 58
Mini-batch gradient descent, 8–9
Misclassification costs, 175–176
Model distribution, 5
Monocentric land value model, 26–27
Multi-class classification, 37
Multicentric land value model, 27
Multilayer NNs, 36–37, 53
Multilayer perceptron (MLP), 175–176
Multinomial logit (MNL), 50–51

Natural language processing (NLP), 2, 53
"Net profit margin", 173–174
Neural interpretation diagrams (NID), 43–44
Neural networks (NN), 2, 25–26, 53
  applications to sales and marketing, 49–54
  case studies, 54–58
  for classification, 37–38
  cost functions and training of neural networks using backpropagation, 38–40
  early evolution, 25–26
  feature importance measurement and visualization, 42–49
  model, 26–38
  output nodes, 40–42
  for regression, 28–37
Next Product To Buy (NPTB), 54
Non-compensatory choice rules, 52, 121
Non-convex region, 146
Non-parametric methods, 139–140
Nonlinear maps and kernels, 94–98
Norm, 91–92, 128

Online learning, 8–9
Optimal classifier, 114
Ordinary least squares regression (OLS regression), 101
Out-of-bag observations, 163
Output nodes, 40–42
Overfitting, 66–68

Parameter norm penalty methods, 73–74
Partial derivatives, feature importance based on, 49
Percent correctly classified (PCC), 11–12, 124, 176
Perceptrons, 89
Permutation importance measure, 164
Pessimistic active learning (PAL), 173–174
Polynomial kernel, 114
Predicted mean squared error (PMSE), 126
Predicted MSE (PMSE), 10
Prediction rule, 143
Profile method for sensitivity analysis, 44–45
Propensity scoring model, 12–13
Prototypes, 173–174

Quadratic cost, 63
  function, 76, 83

Radial basis function kernel (RBF kernel), 99, 130, 175–176
Radial basis kernel, 126
Radial kernel, 123
Random forest, 2, 139–140, 169–171
  applications in marketing and sales, 171–176
Randomization approach for weight and input variable significance, 48–49
Receiver operating characteristics curve (ROC curve), 13–14
Recency, frequency, monetary value analysis (RFM analysis), 173–174
Rectified Linear Units (ReLU), 33
Recurrent neural networks (RNN), 2
Recursive binary partitioning, 145
Regression
  cost complexity pruning, 149–150
  greedy algorithm, 147–149
  models, 2–3
  NN for, 28–37
  performance assessment for, 9–19
  trees, 147–150
Regularization, 66, 72–78
  through bagging and ensemble methods, 78
  through early stopping, 77

through input noise, 76
through sparse representations, 77–78
Rent value
location *vs.*, 125
prediction, 57–58
Response variables, 1–3
Ridge regression, 73
"Root" node, 149–150

Sales and marketing
applications of NN to, 49–54
SVM applications in, 114–120
Sampling variability, 69
Satisficing rule, 52
Segmentation, 155–158
Segmentation, targeting and positioning (STP), 50, 155–156
Self-organizing feature maps (SOFM), 115
Separability, 109–110
Separating hyperplanes, 88–89, 127
Sequential binary programming (SBP), 173–174
Shannon's entropy, 151–152
Sigmoid activation function, 33
for binary outputs, 40–41, 63
Sigmoid function, 33, 36
Sigmoid kernel, 130
Similarity, kernels as measures of, 91–94
Slack variable, 107–109
Soft margins, 107
Softmax activation function for multiclass outputs, 42, 64
Softmax function, 37
Sparse representations, regularization through, 77–78
Sparsity, 75, 81–82
Stochastic gradient boosting algorithm, 169
Stochastic gradient descent (SGD), 8–9
Stopping rule, 148–149
Streaming data, 8–9
Sum of squares (SS), 147, 150, 181
cost, 4, 19
error cost, 38
Supervised learning models, 1
Supervised segmentation, 155–156
Support vector classifier, 106–114
Support vector clustering (SVC), 115
Support vector machine (SVM), 2, 85–86, 175–176
applications in marketing and sales, 114–120
case studies, 120–127
early evolution, 85–86
hyperplanes, 88–89
kernels in machine learning, 90–99
nonlinear classification using, 86–88
optimal separating hyperplane, 99–106
support vector classifier and, 106–114
Support vectors, 102
SVMauc technique, 118–119

Taiwan Ratings Corporation, 118–119
Target variables, 1–3
Test data, 9–10, 71
Test error, 9–10, 66, 68
Text analysis, 119–120
Text classification, 120
THAID, 139–143
Top decile lift, 17
Training error, 9–10, 66
Training of machine learning models, 1–9
cost functions and training of machine learning models, 3–4
gradient-based learning, 6–9
MLE, 5–6
regression and classification models, 2–3
Tree size, 149–150
Tree-based model, 175–176

"Trial-and-repeat" purchase models, 2
True data distribution, 5, 71–72

Underfitting, 69
Units, 25–26
Universal approximation theorem, 33–34
Unsupervised segmentation models, 156

Vapnik–Chervonenkis theory, 85–86
Variance, 69
Virtual reality (VR), 53
Visualization, 42–49

Weight decay, 72–78
  in L_1 regularization, 81
  in L_2 regularization, 80–81
  parameter, 74, 80, 150
Weight(s), 2
  vectors, 31
  weight-based input importance method, 45
  weighted additive rule, 52
Wine quality, 178
Wolfe Dual program, 106, 133

XOR problem, 113