

Index

- 10-fold cross-validation, 13–14, 25, 31, 150–151, 170, 182
- accuracy and error rate, 58, 91
- accuracy metric, 52, 94, 97, 103
 - balanced, 92, 381
- activation maps, 320–321
- active learning, 34–35, 251, 297, 375
- adjusted precision. *See* P@n metric
- adjusted Rand index, 253, 256, 258, 270, 374
- adversarial examples, 297, 299–300, 305
- agreement metrics, 95, 97–98
- agreement statistics, 85, 94, 97
 - Cohen’s kappa measure of agreement, 238
- Akaike information criterion (AIC), 268–269
- algorithmic bias, 7, 290, 293, 298–299, 353
- anomaly detection, 4, 41, 212–216, 218–220, 241
- anonymization techniques, 331, 333
- ANOVA, 185, 207, 309
- area under the curve (AUC), 63–64, 88–89, 117, 124–125, 216, 220–221, 247–248, 252, 255, 373–374, 378–383
 - multiclass, 124–125
- asymmetric costs, 85, 112, 374
- average error rate, 76, 145, 155
- average precision (AP), 216–218
- average prediction, 46, 134–136
- balanced accuracy, 65–66, 90–93, 98, 100, 111, 349
- balanced bootstrap sampling, 148
- balanced *F*-measure, 30, 62
- Bayesian statistics, 156, 159, 163–164, 202–203, 207
 - Bayesian analysis, 156–157, 162–163, 165–166, 201–203, 205, 207
 - Bayesian correlated *t*-test, 165–166, 206
 - Bayesian hierarchical correlated *t*-test, 167
 - Bayesian information criterion (BIC), 269
- bias and variance, 6, 33, 43–44, 46, 48–49, 134, 136
 - analysis, 43, 133–134
 - behavior, 44–45, 47–48, 134, 136, 150–151, 350
 - framework, 130
 - tradeoff, 6
- binomial distribution, 16–19, 21, 78, 132, 178
- BLEU, 238, 284
- blind testing, 311
- blocked cross-validation, 242
- Bonferroni correction, 166, 179–180, 184, 186, 207
- bootstrap confidence intervals, 167, 194
- bootstrapping, 80, 128–129, 144–151, 194, 351
 - bootstrap sampling, 135, 146–148, 194
- brevity penalty (BP), 284
- Brier’s score, 110, 114–116
- calibration, 6, 85–86, 110–115, 117, 122, 124, 307, 380
- central limit theorem, 8, 16, 19–21, 170
- chance agreement, 94–97
- class distributions, 91, 93, 97–99, 110–112, 126, 310, 346–347
- classifier uncertainty, 84, 87, 90, 110, 113
- class imbalance problem, 92–93, 97, 99, 101, 110–112, 212, 214–215, 217–218, 220, 296, 381
 - extreme, 104–107
 - and misclassification costs, 110–111
- class noise, 131, 219, 297, 380
- clustering, 6–7, 33–34, 41, 51, 69–71, 79, 243, 252–253, 255, 257–262, 270
 - and hierarchical clustering, 7, 252–253, 255, 257, 259, 261
 - metrics, 6, 71
- coefficient of determination, 196, 230–231
- comparing multiple classifiers, 53, 164, 168, 179, 188, 202, 207
 - on multiple domains, 164, 168, 184, 202, 207
- computational complexity, 73, 139, 141, 152, 260, 349–351, 383
- confidence intervals, 6, 8, 21–25, 28, 123, 132–133, 156–157, 161–162, 165, 167, 191–197, 199, 349
 - estimated, 22
 - theoretical, 72, 131

- confusion matrix, 51, 54–57, 64–65, 84, 86–87, 90, 92, 95, 99–101, 104, 106, 108–113, 228
- corrected resampled t -test, 171–172, 174
- correlated t -test, 164, 202, 205–207, 352
 - hierarchical, 205–207
- cost curves, 86–87, 126
- cost-sensitive learning, 112, 375
- credibility and trustworthiness of machine learning research, 336
- credible intervals, 22
- critical values, 27, 76, 78–79, 180, 182–183, 186, 188, 192, 361, 365, 372
- cross-entropy metric and Brier's score, 114
- cross-validation, 12, 71, 74, 122, 140, 143, 145, 151, 165, 171–172, 175, 241–243, 270
 - cross-validated t -test, 172, 174
 - CV test, 174–175, 177
 - F -test, 172, 175
 - t -test, 164, 172, 175
 - regular, 73, 242
 - stratified k -fold cross-validation, 138, 140
- crowdsourcing, 296–297
- cumulative distribution function (CDF), 9, 23
- data
 - anonymization, 331
 - augmentation, 296, 310–312
 - bias detection, 308–309, 311
 - collection, 12, 161, 306
 - distribution, 134, 137, 141, 171, 268, 272–273
 - high-dimensional, 263, 271, 274
 - privacy, 308, 330–331, 333
- data envelopment analysis (DEA), 383
- data stream mining, 7, 211–213, 242–243, 245
- Davies–Bouldin Index (DBI), 69–71, 253–255
- deep deterministic policy gradient (DDPG), 249
- deep learning models, 235–236, 272, 337
- dimensionality reduction, 6–7, 33, 41–42, 252, 262–263, 265–266, 272
- discriminant power, 98, 106–107
- Dunn index, 254
- effect size, 6, 31–32, 157, 161–162, 165–168, 191–201
 - measures, 196, 199
 - and p -value, 199–200
 - recommended, 165, 167–168
- efficiency method, 383–384
- empirical risk, 12, 14–16, 20, 23–28, 37, 39–41, 45, 130–133, 139, 152–154
 - estimate, 20, 130–132, 147
 - minimization (ERM), 39, 50, 225
- error estimates, 6, 53, 72–73, 133–134, 136–139, 141, 149–150, 169, 349, 351
 - biased, 53, 73, 347
 - empirical, 112, 134, 176
- estimation statistics, 156–157, 159, 162, 165, 191, 193, 195, 197, 199, 207
- ethical issues, xvii, 304, 353
- evaluation
 - deep semi-supervised learning, 250
 - dimensionality reduction techniques, 262–263
 - divisive clustering, 260
 - GANs, 278–279, 281–283
 - hierarchical clustering, 257, 260
 - image segmentation, 234
 - latent variable models, 267
 - lifelong learning systems, 245
 - ordinal classification, 222
 - probabilistic PCA (PPCA), 273
 - PU Learning, 219
 - qualitative, 375–376
 - text generation, 237, 283
 - time series, 216, 243–244
 - variational autoencoders, 274
- exact match ratio, 225
- expected value, 9–11, 16–17, 25
- explainability, 7, 289, 291, 299, 308, 312–313, 315–325, 376
- extrinsic measures, 252–253, 255
- fairness, 5, 7, 289, 291, 294, 298, 303–304, 308–309, 327–329, 341
 - metrics, 327, 329–330
- false negative (FN), 29, 55–57, 59–61, 93, 95, 111, 120–121, 126, 226, 256, 328
- false negative rate (FNR), 59–60, 126–127
- false positive rate (FPR), 58–59, 63–64, 88–89, 98, 117–120, 122–123, 126, 216, 328
- false positives (FP), 55–57, 59–61, 95–96, 111, 121, 220, 226, 256, 328
- family wise error rate (FWER), 179
- fidelity, 277
- fit, 48–49, 67, 86, 230–232, 269, 274
 - goodness of, 67, 178, 267–269
- F -measure, 30, 62, 75, 87–89, 98, 104, 220–221, 252, 255, 373, 378
- Fréchet inception distance (FID), 278–279
- frequentist approach, 156, 162, 164–168
- Friedman test, 168, 184–188, 199
- generalization, 43, 48, 50, 115, 124, 131–132, 136, 180, 238, 268, 331
 - error, 16, 33, 38–39, 137

- error bounds, 50, 152
- generated
 - images, 277–281
 - samples, 273, 277, 279–280, 282–283
 - text, 235, 238–239, 284–285, 354
- generator, 43, 272–273, 279–280
- geometric mean (*G*-Mean), 80, 98, 104–106, 127, 373, 381
- Grad-CAM visualization technique, 321–322
- Hamming loss (HL), 226
- Hamming score (HS), 226–228
- Harabasz index, 253–255
- HDI, 165, 203–205
 - HDI graphs, 204
- heatmaps, 271, 321–322
- holdout, 6, 128, 141–143, 242–243
 - estimate, 132–133
 - method, 72, 74, 128–131, 134, 136, 142, 177, 242–244
- human-centric
 - evaluation, 238
 - machine learning, 7, 334
- human-centric evaluation, 238
- hypothesis, 6, 8, 21, 25–28, 75, 143, 149, 160–161, 175, 179–180, 183
- hypothesis test, 25, 27–29, 31, 172, 361
 - multiple, 184–186
 - statistical, 22–28, 169
- image
 - classification, 234, 251, 336–337
 - generation, 272, 279
 - segmentation, 5, 7, 41, 79, 211, 213, 233–235
- industrial strength evaluation, 5, 79, 290–306, 353
- information, mutual, 253, 256, 261–262
- interpolation, 276
- interpretability and explainability, 85, 260, 294, 299, 308, 312–313, 318, 323–324, 335
 - interpretable results, 323–327
 - interpretation, 22, 25, 123, 194, 197, 258–259, 320, 323–324, 326–327, 338, 346
 - interpretations and measures of fairness, 298
- isomap, 42, 263
- isometrics, 121
- Jaccard index, 234–235, 256–258
- Kendall's *W* test value, 168
- k*-fold cross-validation, 6, 72–73, 122, 128, 137–143, 148, 151–152, 170–171, 175–176, 270, 351
 - classical, 141
- label-based metrics, 225, 228–229
- label generation process, 12, 94–95
- language models, 235, 283–284
- latent variable modeling, 6–7, 33, 41–42, 252, 266–270, 272
 - latent variables, 42–43, 267, 270–273, 275
- leave-one-out, 73–74, 128–129, 137–139, 141–142, 151
- lifelong learning, 7, 80, 211–213, 244–247
 - metrics, 247
 - systems, 244–247
- lift charts, 125
- likelihood, 24, 109, 203, 269, 271
 - likelihood-based measures, 267
 - likelihood ratios, 98, 106–109, 111, 269
 - likelihood-ratio test, 269–270
- linear models, 42, 323–324
- local interpretable model-agnostic explanations (LIME), 299, 315–316, 318, 320
- log-likelihood, 268–270, 272, 274–275
- log loss, 87, 110, 114–116
- longest common subsequence (LCS), 285
- long tail distributions, 110
- loss, zero-one, 37, 47–48, 83, 130, 135
- loss function, 6, 15, 33, 36–38, 44–47, 49, 134, 153, 244, 274–275, 282
 - arbitrary, 44–46
- machine learning deployment, xvi, 302–303, 353–355
- macro-average, 65–66
- McNemar's test, 168–169, 171, 177–180
- mean absolute error (MAE), 66–68, 222–223, 231, 233, 241
- mean absolute percentage error (MAPE), 241
- mean opinion score (MOS), 282–283
- mean squared error (MSE), 67–68, 75, 115, 222–223, 230–231, 233, 270, 273
- micro, 6, 225, 228–229
- misclassification costs, 91, 93, 97, 110–112, 374
 - incorporating, 112
 - unequal, 91
- mitigating data bias, 309
- MLOps, 305–306
- model
 - deployment, 305–306
 - interpretable, 268, 312–313, 315–316
 - selection, 38–39, 48–50, 63, 130, 135–136, 142–143, 152, 268, 374–375
 - selection criterion, 39, 50
 - selection process, 50, 134, 136
 - visualization, 312
- multi-class classification, 51, 65–66, 85, 92, 115, 122, 124, 140, 211, 346
- multi-class focus, 85–86, 90–91, 93, 95, 97
- multi-criteria
 - evaluation, 383
 - metrics, 87, 381–382
- multi-label classification, 211–212, 223–225, 227, 229
- multiple resampling, 128, 144–145, 147, 149–150

- negative predictive value (NPV), 60–61, 98, 100–101
- Nemenyi test, 164, 188–190, 199
- NMI, 255, 261–262, 270
- nonmonotonic metrics, 85
- nonparametric approaches, 28–29, 179–180, 194
- nonparametric confidence intervals, 167, 191, 194
- nonparametric tests, 29, 166–168, 180, 184
- normal distribution, 11, 16–19, 22, 24–26, 28, 31, 75, 170, 173, 353, 361
- NPV. *See* negative predictive value
- null hypothesis, 24, 26–31, 75–79, 160–162, 165, 167–168, 173, 175, 178–180, 182–186, 188–191, 196, 198–200, 269
- null hypothesis significance testing (NHST), 156–157, 160–163, 165, 168–169, 171–189, 191, 196, 199, 202, 207
- observed
 - data, 39, 156, 266–272, 274
 - results, 154, 163, 165, 208, 351
- one-class learning, 214, 218
- one-tailed tests, 28, 30–31, 76–77
- ordinal classification, 6, 211–212, 221–223
- outlier detection, 6, 211, 213, 215, 217
- overfitting, 39, 41, 43–44, 48–49, 67, 69, 134, 230, 233, 268–269
 - avoiding, 268–269
- overlapping clusters, 253, 271
- overlapping training sets, 171–172
- P@n metric, 216–218
- paired *t*-tests, resampled, 172
- parametric approach, 28–29, 179, 195
- parametric confidence intervals, 165, 191, 195–196
- partial dependence plots (PDP), 314–315
- perplexity, 283–284
- pointwise mutual information (pmi), 75–77, 79, 173, 181, 383–385
- positive predictive value (PPV), 60–61, 98, 100–103
- positive sentiment prediction, 316, 318, 320
- positive unlabeled (PU), 79, 212, 218–219
- positive unlabeled classification, 6, 211
- post-hoc tests, 184, 187–188
- power analysis, 6, 157, 191–201
- precision, 30, 60–62, 89, 98, 102–105, 125–126, 215–218, 220–221, 227–229, 255, 279, 285
- precision and recall, 62, 80, 89, 102–104, 125–126, 215–216, 220, 279, 285
- precision-recall curve (PR-curve), 86–87, 125–126, 216, 218
- privacy
 - and data protection rights of individuals, 308
 - preserving, 330–331
 - preserving data publishing, 331
 - protecting model, 330
 - risks in machine learning, 304
 - and security in machine learning, 330
- probabilistic classifiers, 84, 86, 109–110, 113, 117, 221
- probability distributions, 8–10, 15, 17, 19–20, 27, 163, 235, 239, 280
- purity, 252, 255
- p*-values, 76–77, 159–160, 163, 165, 169, 186, 190, 199–200, 269
 - actual, 169
 - corrected, 181
- qualitative metrics, 373, 375, 383
- q*-value, 189, 249
- Rand index, 256
- random subsampling, 128, 144–145, 172–173
- random variables, 6, 8–14, 16–17, 19–23, 25
 - continuous, 8–10, 15–16
- receiver operating characteristic analysis. *See* ROC analysis
- region of practical equivalence (ROPE), 166–168, 203–205, 207
- regression analysis, 6, 51, 66–69, 211, 222, 229–232, 240–241, 309, 324
- reinforcement learning, 6, 211, 213, 248–250, 375
- reliability metrics, 87, 380–381
- repeatability, 291, 335–337, 339
- replicability in machine learning, 335, 337, 339
- reproducibility, 291, 305, 309, 335–340
- resampling, 45, 50, 122, 128–152, 351
 - regimens, 74, 117, 121–122, 129–130
 - schemes, 73, 133–134, 144, 151, 165
- resampling methods, 50, 53–54, 128–130, 134–137, 140, 142–143, 145, 150–152, 154, 156, 172
 - multiple, 130, 143, 148
 - simple, 128, 130, 137, 142–143, 150–151
- responsible machine learning, 50, 79, 290, 308–341
- risk, xv, 14–16, 31, 33, 36–39, 136, 138–139, 142, 144–145, 147, 152–153, 307, 309, 330–331
 - estimates, 25, 131, 141
- ROC analysis, 62–64, 84, 86–87, 117, 119, 121–124, 215–216, 248, 343, 349–350, 373–374
 - isometrics, 126
 - multi-class problems, 124
 - ROC space, 63–64, 117–119, 121, 124, 127, 378
- ROC-AUC metric. *See* area under the curve (AUC)
- ROC curve, 62–64, 87–89, 117, 119–127, 216, 349, 374–375
 - multi-class ROC curves, 124–125
- root mean squared error (RMSE), 67–68, 230–231, 233, 241, 379–382
- ROUGE, 238, 285
- R-precision, 216–217
- R-square measure, 67
 - adjusted, 66, 68
- sample statistics, 20–21, 28–30, 129, 170
- sample variance, 10–11, 75, 78
- sampling, 21, 25, 72, 135–136, 144, 146, 149, 171, 272, 282
 - sampling distribution, 20–21, 25, 75

- scoring classifiers, 83–84, 86, 109, 119, 121, 373
- Semantic propositional image caption evaluation (SPICE), 239
- sensitivity, 59–61, 63, 97, 99–100, 102–107, 110, 164, 231, 259, 280, 323
- sensitivity and specificity, 59–60, 63, 80, 98–101, 104–107, 349
- sequence-to-sequence models, 237
- SHapley Additive exPlanations (SHAP), 299, 317–318, 320
- signed-rank test, 166, 168, 180–184, 186–187, 191–192, 194, 198–199, 205–206
- significance level, 31, 76, 78–79, 156, 173, 178, 180, 183, 188–189, 200–201, 365
- sign test, 6, 51, 53, 74, 78–79, 164, 166, 168, 180, 361, 365
- silhouette analysis, 261
- simple and intuitive measure (SIM), 383–385
- Spearman rank correlation coefficient, 265
- structural risk minimization, 6, 39–40, 50
- sum squared error (SSE), 66–68, 230–233
- tables of critical values, 78–79, 361, 365, 372
- testing sets
 - test folds, 12–15, 123, 143
 - test partition, 133, 143, 151
- testing standards, 290–292
- training and testing sets, 56, 72–74, 140, 150, 242
 - testing sets, 11–12, 15–17, 49–50, 54–57, 67–69, 71–76, 83–84, 110, 112–113, 121–122, 124, 131–133, 136, 138–143, 150–153, 170–173, 217–218, 225–226, 242–243, 284
 - training sets, 34, 37–39, 45–47, 56, 71–72, 130–132, 134–138, 140–145, 148–149, 152, 169–174, 177, 202–203, 219–220, 242
- transfer learning, 250
- transformer models, 237
- true negative (TN), 55, 57, 59–60, 63, 95, 103–104, 118, 173, 256
- true negative rate (TNR), 59, 63, 105, 111, 118
- true positive rate (TPR), 58–59, 61, 63–64, 88–89, 105, 111, 117–120, 122–123, 125, 127, 328
- true positives (TP), 55, 57, 95, 121, 125, 256
- true risk, 23, 28, 37–39, 130, 132–133, 138, 152–153
- t*-test, 24, 26, 51, 53, 74, 76–79, 164–165, 168–172, 181, 184, 194, 196–198, 201, 309, 351–353
- Tukey test, 361, 372
- two-tailed test, 27–28, 30–31, 76–77, 79, 188
- type II error, 8, 29–32, 161, 200–201, 352
- underfitting, 48–49, 134
- underlying distribution, 10–12, 20–22, 38
- Wilcoxon's signed-rank test, 181, 185, 198
- Wilcoxon statistics, 181–183, 198
- Youden's index, 87, 98, 106