# Index

<table>
<thead>
<tr>
<th>Page Numbers</th>
<th>Terms and Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>179–182</td>
<td>Bootstrap, 179–182</td>
</tr>
<tr>
<td>2-fold matched pair t-test, 225</td>
<td>5 × 2 CV F-Test, 261–263</td>
</tr>
<tr>
<td>5 × 2 CV t-test, 261</td>
<td>10 × 10 CV, 284–289</td>
</tr>
<tr>
<td>86–88, 315</td>
<td>Accuracy, 86–88, 315</td>
</tr>
<tr>
<td>25, 311</td>
<td>Active learning, 25, 311</td>
</tr>
<tr>
<td>89–94, 110</td>
<td>Agreement statistic, 89–94, 110</td>
</tr>
<tr>
<td>10, 12, 14, 322, 36, 380</td>
<td>Aggregation, 10, 12, 14, 322, 36, 380</td>
</tr>
<tr>
<td>12, 105</td>
<td>Artificial data approach, 294, 301–304</td>
</tr>
<tr>
<td>12, 105</td>
<td>Asymmetric misclassification cost, 12, 105</td>
</tr>
<tr>
<td>143</td>
<td>Area Under the Curve (AUC), 143</td>
</tr>
<tr>
<td>102, 104, 317</td>
<td>Bias, 39, 167–168, 170, 178, 181, 225</td>
</tr>
<tr>
<td>133–134</td>
<td>Bias-variance analysis, 34, 40, 162, 167–168, 329</td>
</tr>
<tr>
<td>170, 395</td>
<td>Bias-variance decomposition, 35–39, 167–168, 170, 395</td>
</tr>
<tr>
<td>77, 79–81</td>
<td>Bonferroni adjustment, 11</td>
</tr>
<tr>
<td>143</td>
<td>Bonferroni Test, 254–255</td>
</tr>
<tr>
<td>214, 255</td>
<td>Bonferroni–Dunn Test, 214, 255</td>
</tr>
<tr>
<td>318, 321</td>
<td>Candidate evaluation function, 318, 321</td>
</tr>
<tr>
<td>30</td>
<td>Capacity, 30</td>
</tr>
<tr>
<td>89–93</td>
<td>Chance agreement, 89–93</td>
</tr>
<tr>
<td>265–266</td>
<td>Chi-squared table, 265–266</td>
</tr>
<tr>
<td>102, 104, 317</td>
<td>Class imbalance problem, 102, 104, 317</td>
</tr>
<tr>
<td>105, 118, 120</td>
<td>Class ratio, 105, 118, 120</td>
</tr>
<tr>
<td>118</td>
<td>Class skew, 118</td>
</tr>
<tr>
<td>72, 221–222, 265, 373</td>
<td>Cohen’s d, 72, 221–222, 265, 373</td>
</tr>
<tr>
<td>109, 304</td>
<td>Cohen’s κ (kappa), 109, 304</td>
</tr>
<tr>
<td>15, 21</td>
<td>Community experiments effect, 15, 21</td>
</tr>
<tr>
<td>59</td>
<td>Confidence coefficient, 59</td>
</tr>
<tr>
<td>59, 61, 70, 166, 324</td>
<td>Confidence parameter, 59, 61, 70, 166, 324</td>
</tr>
<tr>
<td>65</td>
<td>Confirmatory data analysis, 65</td>
</tr>
<tr>
<td>77–79</td>
<td>Confusion matrix, 77–79</td>
</tr>
<tr>
<td>196, 259–260, 278–279</td>
<td>Corrected random subsampling, 196, 259–260, 278–279</td>
</tr>
<tr>
<td>104–106</td>
<td>Cost, 104–106</td>
</tr>
<tr>
<td>133–134</td>
<td>Cost curve space, 133–134</td>
</tr>
<tr>
<td>60</td>
<td>Credible interval, 60</td>
</tr>
<tr>
<td>Cross-entropy, 146</td>
<td></td>
</tr>
<tr>
<td>Cross-validated t-test, 188</td>
<td></td>
</tr>
<tr>
<td>Cross-Validation (CV), 172</td>
<td></td>
</tr>
<tr>
<td>Cumulative distribution function (CDF), 45, 61</td>
<td></td>
</tr>
<tr>
<td>Data mining, 2, 6, 19–20, 206, 246, 294–297, 307, 350, 370</td>
<td></td>
</tr>
<tr>
<td>Dataset, 15</td>
<td></td>
</tr>
<tr>
<td>De-Facto approach/culture, 7–8</td>
<td></td>
</tr>
<tr>
<td>DET curve, 136, 156, 160</td>
<td></td>
</tr>
<tr>
<td>Deterministic algorithm, 75</td>
<td></td>
</tr>
<tr>
<td>Domain specific metric, 145</td>
<td></td>
</tr>
<tr>
<td>Dunnett Test, 253–255, 351, 363, 382, 384</td>
<td></td>
</tr>
<tr>
<td>Effect size, 71–72, 211, 213, 221–222, 258, 264–265, 373</td>
<td></td>
</tr>
<tr>
<td>Efficiency method, 318–320</td>
<td></td>
</tr>
<tr>
<td>Empirical risk, 27–28</td>
<td></td>
</tr>
<tr>
<td>Empirical risk minimization, 29–30, 54, 159</td>
<td></td>
</tr>
<tr>
<td>Error estimation, 35, 74, 84, 161</td>
<td></td>
</tr>
<tr>
<td>Error rate, 86–87, 89, 94, 96, 104, 109, 133–134, 159, 162, 167, 175, 182, 187</td>
<td></td>
</tr>
<tr>
<td>Evaluation framework template, 336–338</td>
<td></td>
</tr>
<tr>
<td>Evaluation measure (performance measure), 75–77</td>
<td></td>
</tr>
<tr>
<td>Expected cost, 120, 122, 128, 135</td>
<td></td>
</tr>
<tr>
<td>Expected risk, 27–28, 163, 176, 202</td>
<td></td>
</tr>
<tr>
<td>Expected value of a random variable, 45–46</td>
<td></td>
</tr>
<tr>
<td>Exploratory data analysis, 66, 211</td>
<td></td>
</tr>
<tr>
<td>False negative, 69, 79, 89, 91, 95, 116–117, 135</td>
<td></td>
</tr>
<tr>
<td>False negative rate, 95</td>
<td></td>
</tr>
<tr>
<td>False positive, 69</td>
<td></td>
</tr>
<tr>
<td>False positive rate, 84, 94–95, 113–116, 121–128, 135, 149, 151–152, 370–371</td>
<td></td>
</tr>
<tr>
<td>F-measure, 104, 108, 157, 372</td>
<td></td>
</tr>
<tr>
<td>F-Ratio table, 351, 357–360</td>
<td></td>
</tr>
<tr>
<td>Friedman table, 351, 361</td>
<td></td>
</tr>
<tr>
<td>Gaussian distribution, 53–54, 60</td>
<td></td>
</tr>
<tr>
<td>Generalization, 16, 32, 34, 39, 41, 69, 87, 90, 93, 110, 131, 144, 164–165, 170, 176, 240, 297, 303, 323, 344</td>
<td></td>
</tr>
<tr>
<td>Generalization error, 16, 28–29, 32, 41, 54, 170, 323–324</td>
<td></td>
</tr>
<tr>
<td>Generic algorithm, 293–294, 301, 336–337, 340</td>
<td></td>
</tr>
<tr>
<td>Geometric mean, 96, 100–101, 103</td>
<td></td>
</tr>
<tr>
<td>Gini coefficient, 122, 139, 159</td>
<td></td>
</tr>
<tr>
<td>Gold standard, 145</td>
<td></td>
</tr>
<tr>
<td>Graphical performance measure, 112</td>
<td></td>
</tr>
<tr>
<td>H Measure, 81, 159, 310</td>
<td></td>
</tr>
<tr>
<td>Holdout method, 70, 161–162, 167, 169, 176, 202, 226</td>
<td></td>
</tr>
<tr>
<td>Holdout risk bound, 325–326, 345</td>
<td></td>
</tr>
<tr>
<td>Honestly significant difference (HSD), 252</td>
<td></td>
</tr>
<tr>
<td>Hypothesis testing, 42, 59–62, 64–68, 162, 184, 188, 211, 216–217, 230, 239, 290, 324</td>
<td></td>
</tr>
<tr>
<td>Indicator function, 32, 85–86, 141, 230, 316, 322–323, 379–381</td>
<td></td>
</tr>
<tr>
<td>Information theoretic measure, 137–138, 143, 159</td>
<td></td>
</tr>
<tr>
<td>Information Reward, 140, 143</td>
<td></td>
</tr>
<tr>
<td>Information Score, 84, 140–142, 157, 317, 376, 384–385, 388, 391</td>
<td></td>
</tr>
<tr>
<td>Iso-curves, 122</td>
<td></td>
</tr>
<tr>
<td>Isometrics, 119–124, 133, 314</td>
<td></td>
</tr>
<tr>
<td>Iso-precision lines, 120</td>
<td></td>
</tr>
<tr>
<td>Jackknife, 175, 186, 203</td>
<td></td>
</tr>
<tr>
<td>Kappa statistic, 93</td>
<td></td>
</tr>
<tr>
<td>Kendall coefficient, 144</td>
<td></td>
</tr>
<tr>
<td>Kononenko and Bratko’s Information Score, 84, 140–142</td>
<td></td>
</tr>
<tr>
<td>Kullback-Leibler divergence, 139</td>
<td></td>
</tr>
<tr>
<td>Learning bias, 31, 33, 177, 323–324, 329, 339</td>
<td></td>
</tr>
<tr>
<td>Leave One Out, 162, 171, 173, 175–176, 194, 202, 327, 341</td>
<td></td>
</tr>
<tr>
<td>Lift chart, 132</td>
<td></td>
</tr>
<tr>
<td>Lift curve, 153–154</td>
<td></td>
</tr>
<tr>
<td>Likelihood ratio, 97–99</td>
<td></td>
</tr>
<tr>
<td>Loss function, 27, 32, 34–38, 40, 48, 52, 75, 106, 137, 144, 168, 204</td>
<td></td>
</tr>
<tr>
<td>Mann-Whitney U test, 129</td>
<td></td>
</tr>
<tr>
<td>McNemar’s Test, 68, 217, 226–231, 236, 264–265, 289, 373</td>
<td></td>
</tr>
<tr>
<td>McNemar’s contingency matrix, 227–229</td>
<td></td>
</tr>
<tr>
<td>Machine learning, 23–42, 279–280</td>
<td></td>
</tr>
<tr>
<td>Matched samples design, 216</td>
<td></td>
</tr>
<tr>
<td>Matched pairs design, 216, 320–321</td>
<td></td>
</tr>
<tr>
<td>Measure-based method, 318</td>
<td></td>
</tr>
<tr>
<td>Meta-learning, 294</td>
<td></td>
</tr>
<tr>
<td>Monotonic performance measure, 65, 77, 216</td>
<td></td>
</tr>
<tr>
<td>Multiclass, 25, 38, 97, 110, 131, 144, 230, 315</td>
<td></td>
</tr>
<tr>
<td>Multi-class classification, 85, 97, 110, 174–175</td>
<td></td>
</tr>
<tr>
<td>Multiclass focus, 85–86, 101</td>
<td></td>
</tr>
<tr>
<td>Multiclass ROC curve, 131</td>
<td></td>
</tr>
</tbody>
</table>
Multiplicty effect, 15, 299–300, 345
Negative predictive value (NPV), 99, 371
Nemenyi Test, 256–257, 275, 382, 388, 390–391
Nested k-fold cross validation, 178
“No Free Lunch” theorems, 292
noise, 36
Non-parametric test, 110, 131, 144, 164–165, 170, 176, 197, 240, 303, 323, 344
Normalized expected cost (NEC), 135
Null hypothesis, 65, 216
Null hypothesis statistical testing (NHST), 207, 289
Omnibus statistical test, 251
One-tailed test/One-sided test, 70, 220
One-Way ANOVA, 214, 245–246
One-Way repeated measure ANOVA, 240–245
Online learning, 25, 311
Ontology of performance measures, 81–82
Ontology of error-estimation methods, 163
Operating point, 112, 114, 116, 119, 122–123
Overview of the statistical tests, 214
Oversampling, 30, 32, 39, 167, 229, 299, 302, 325
Panacea approach to evaluation, 4, 7, 349
Parameter estimation, 34
Parameter selection, 4, 15, 76, 178
Parametric hypothesis testing, 68
Passive learning, 25
Perfect calibration, 140, 143
Performance metric, 4, 21, 83, 85–89, 94, 308–309, 311
Permutation test, 186, 199–203, 329, 373
Poisson distribution, 55–56
Positive predictive values (PPV), 13
Power, 69–70, 145, 203, 226, 230, 256, 260, 275
Precision, 13, 70–71, 84, 99–100, 102, 104, 109
Precision-Recall curve (PR Curve), 122–133
Prior class probability, 13
Probabilistic algorithm, 75
Probability density function (PDF), 45, 54
Probability distribution, 44–46, 55, 58, 67, 139, 331
Probability space, 44
Qualitative metric, 311–312, 319
Quantitative metric, 309–311
R, 42
Random variable, 42, 44–48, 52, 54, 58, 60
Randomization, 15, 162, 183–185, 188, 202, 345, 348
Ranking classifier, 75, 118, 129, 248
Recall, 137–138
Receiver Operating Characteristic (ROC), 111, 112–113
regression, 27–28
Regularization, 15, 162, 183–185, 188, 202, 345, 348
Repeated-measures design, 216
Replicability, 185, 203, 262–263, 291, 294
Repository approach, 294
Resampled matched pair t-test, 225
Resampling, 166–167, 175–176
Resampling framework, 171–172, 180
Resampling statistics, 14–15
Resubstitution error, 161, 164, 181, 202
risk, 27–28
$\times \times k CV$, 261–263
ROC Analysis, 112–113
ROC Convex Hull (ROCCH), 122
ROC Curve generation, 124–126
ROC Space, 113–120, 123, 128, 130, 153–134, 136, 314
Root Mean Squared Error (RMSE), 137–143, 157
Sample mean, 45–46, 48, 51–53, 57–58, 218, 240
Sample standard deviation, 47, 60–61, 176, 218
Sampling distribution, 57–59, 63, 216, 218, 247, 332, 348
S Coefficient, 90–91, 110
Scoring classifier, 117–118
Scott’s $\pi$ (pi) coefficient, 91
Semi-supervised learning, 24–25
Sensitivitiy, 95–99
Sign Test, 228–229, 231–233, 236–239, 267, 289, 355
Silver standard, 89
Simple and intuitive measure (SIM), 319–320
Single class focus, 87, 94, 102
Skew, 104–106
Skew ratio, 110, 112, 119–120, 123, 130, 310
Specialization, 34, 39
Specificity, 13, 82, 95
Standard deviation, 47, 52
Standard error, 60–61, 251
Statistical distribution, 46
Statistical hypothesis testing, 161–62, 65–66, 68, 162, 230, 290
Statistical significance testing, 6, 14–15, 21, 57, 74, 162, 185, 188, 203, 206
## Index

| Statistical table, 22, 351 | Type II error, 42, 69–70, 347 |
| Statistical test, 206–215 | UCI Machine Learning Repository, 7 |
| Stratified k-fold Cross-Validation, 189, 191, 193 | variance, 11, 37, 40, 46 |
| Structural risk minimization, 29, 30 | visualization approach, 321–323 |
| Summary statistic, 128–130, 310 | visualization-based combination metric, 321–323 |
| Supervised learning, 1–3, 21, 24–25, 310 | |
| Task specific algorithm, 293–294 | |
| Training set bound, 126–128, 176, 324 | Wilcoxon’s Signed Rank Test, 233–235, 236 |
| True negative, 19, 96, 102, 104, 113 | Wilcoxon’s Rank Sum Test, 129 |
| True negative rate, 95–96, 113 | Wilcoxon Table, 235, 238, 267–268, 356 |
| True positive, 79, 84, 94–95, 116, 128, 132 | Within-group variation, 240, 242 |
| True positive rate, 94–95, 128, 132 | |
| True risk, 27–29, 61, 68, 162, 164–165, 204, 324 | Zero-one loss, 27, 32, 36, 39, 163, 169 |
| t-table, 261 | Z-table, 351 |
| t-test, 224 | |
| Tukey Test, 251–253, 257, 275, 362 | |
| Two-tailed test/Two-sided test, 67–68 | |