

Contents

1	Introduction	1
1.1	Machine Learning	1
1.1.1	Learning Under Randomness	2
1.1.2	Learning Under Unconstrained Randomness	3
1.2	A Shortcoming of Statistical Learning Theory	3
1.2.1	The Hold-Out Estimate of Confidence	4
1.2.2	The Contribution of This Book	4
1.3	The Online Framework	5
1.3.1	Online Learning	5
1.3.2	Online/Offline Compromises	6
1.3.3	One-Off and Offline Learning	6
1.3.4	Induction, Transduction, and the Online Framework	7
1.4	Conformal Prediction	9
1.4.1	Nested Prediction Sets	9
1.4.2	Validity	10
1.4.3	Efficiency	11
1.4.4	Conditionality	12
1.4.5	Flexibility of Conformal Predictors	12
1.5	Probabilistic Prediction Under Unconstrained Randomness	13
1.5.1	Universally Consistent Probabilistic Predictor	13
1.5.2	Probabilistic Prediction Using a Finite Dataset	14
1.5.3	Venn Prediction	14
1.5.4	Conformal Predictive Distributions	14
1.6	Beyond Randomness	15
1.6.1	Testing Randomness	15
1.6.2	Online Compression Models	15
1.7	Context	16

Part I Set Prediction

2	Conformal Prediction: General Case and Regression	19
2.1	Confidence Predictors	19
2.1.1	Assumptions	19
2.1.2	Simple Predictors and Confidence Predictors	20
2.1.3	Validity	22
2.1.4	Randomized Confidence Predictors	24
2.1.5	Confidence Predictors Over a Finite Horizon	25
2.1.6	One-Off and Offline Confidence Predictors	26
2.2	Conformal Predictors	27
2.2.1	Bags	27
2.2.2	Nonconformity and Conformity	28
2.2.3	p-Values	29
2.2.4	Definition of Conformal Predictors	30
2.2.5	Validity	31
2.2.6	Smoothed Conformal Predictors	31
2.2.7	Finite-Horizon Conformal Prediction	32
2.2.8	One-Off and Offline Conformal Predictors	33
2.2.9	General Schemes for Defining Nonconformity	34
2.2.10	Deleted Conformity Measures	36
2.3	Conformalized Ridge Regression	37
2.3.1	Least Squares and Ridge Regression	37
2.3.2	Basic CRR	38
2.3.3	Two Modifications	42
2.3.4	Dual Form Ridge Regression	44
2.4	Conformalized Nearest Neighbours Regression	46
2.5	Efficiency of Conformalized Ridge Regression	48
2.5.1	Hard and Soft Models	49
2.5.2	Bayesian Ridge Regression	49
2.5.3	Efficiency of CRR	50
2.6	Are There Other Ways to Achieve Validity?	53
2.7	Conformal Transducers	54
2.7.1	Definitions and Properties of Validity	54
2.7.2	Normalized Confidence Predictors and Confidence Transducers	56
2.8	Proofs	58
2.8.1	Proof of Theorem 2.2	58
2.8.2	Proof of Theorem 2.7	58
2.8.3	Proof of Theorem 2.10	63
2.9	Context	63
2.9.1	Exchangeability vs Randomness	63
2.9.2	Conformal Prediction	64
2.9.3	Two Equivalent Definitions of Nonconformity Measures	66

2.9.4	The Two Meanings of Conformity in Conformal Prediction	66
2.9.5	Examples of Nonconformity Measures	67
2.9.6	Kernel Methods	67
2.9.7	Burnaev–Wasserman Programme	68
2.9.8	Completeness Results	68
3	Conformal Prediction: Classification and General Case	71
3.1	Criteria of Efficiency for Conformal Prediction	71
3.1.1	Basic Criteria	72
3.1.2	Other Prior Criteria	72
3.1.3	Observed Criteria	74
3.1.4	Idealised Setting	75
3.1.5	Conditionally Proper Criteria of Efficiency	77
3.1.6	Criteria of Efficiency that Are not Conditionally Proper ..	78
3.1.7	Discussion	82
3.2	More Ways of Computing Nonconformity Scores	84
3.2.1	Nonconformity Scores from Nearest Neighbours	84
3.2.2	Nonconformity Scores from Support Vector Machines ...	86
3.2.3	Reducing Classification Problems to the Binary Case	86
3.3	Weak Teachers	87
3.3.1	Imperfectly Taught Predictors	88
3.3.2	Weak Validity	89
3.3.3	Strong Validity	90
3.3.4	Iterated Logarithm Validity	91
3.3.5	Efficiency	91
3.4	Proofs	92
3.4.1	Proofs for Sect. 3.1	92
3.4.2	Proofs for Sect. 3.3	99
3.5	Context	105
3.5.1	Criteria of Efficiency	105
3.5.2	Examples of Nonconformity Measures	105
3.5.3	Universal Predictors	106
3.5.4	Weak Teachers	106
4	Modifications of Conformal Predictors	107
4.1	The Topics of This Chapter	107
4.2	Inductive Conformal Predictors	108
4.2.1	Inductive Conformal Predictors in the Online Mode	108
4.2.2	Inductive Conformal Predictors in the Offline and Semi-Online Modes	109
4.2.3	The General Scheme for Defining Nonconformity	112
4.2.4	Normalization and Hyper-Parameter Selection	113
4.3	Further Ways of Computing Nonconformity Scores	114
4.3.1	Nonconformity Measures Considered Earlier	115
4.3.2	De-Bayesing	116

4.3.3	Neural Networks and Other Multiclass Scoring Classifiers	117
4.3.4	Decision Trees and Random Forests	118
4.3.5	Binary Scoring Classifiers	118
4.3.6	Logistic Regression	119
4.3.7	Regression and Bootstrap	120
4.3.8	Training Inductive Conformal Predictors	120
4.4	Cross-Conformal Prediction	122
4.4.1	Definition of Cross-Conformal Predictors	122
4.4.2	Computational Efficiency	123
4.4.3	Validity and Lack Thereof for Cross-Conformal Predictors	123
4.5	Transductive Conformal Predictors	125
4.5.1	Definition	125
4.5.2	Validity	126
4.6	Conditional Conformal Predictors	126
4.6.1	One-Off Conditional Conformal Predictors	127
4.6.2	Mondrian Conformal Predictors and Transducers	128
4.6.3	Using Mondrian Conformal Transducers for Prediction ..	130
4.6.4	Generality of Mondrian Taxonomies	131
4.6.5	Conformal Prediction	132
4.6.6	Inductive Conformal Prediction	133
4.6.7	Label-Conditional Conformal Prediction	134
4.6.8	Object-Conditional Conformal Prediction	134
4.7	Training-Conditional Validity	135
4.7.1	Conditional Validity	135
4.7.2	Training-Conditional Validity of Inductive Conformal Predictors	137
4.8	Context	140
4.8.1	Computationally Efficient Hedged Prediction	140
4.8.2	Specific Learning Algorithms and Nonconformity Measures	140
4.8.3	Training Conformal Predictors	140
4.8.4	Cross-Conformal Predictors and Alternative Approaches ..	140
4.8.5	Transductive Conformal Predictors	141
4.8.6	Conditional Conformal Predictors	141

Part II Probabilistic Prediction

5	Impossibility Results	145
5.1	Introduction	145
5.2	Diverse Datasets	146
5.3	Impossibility of Estimation of Probabilities	147
5.3.1	Binary Case	147
5.3.2	Multiclass Case	148

5.4	Proof of Theorem 5.2	149
5.4.1	Probability Estimators and Statistical Tests	149
5.4.2	Complete Statistical Tests	150
5.4.3	Restatement of the Theorem in Terms of Statistical Tests	150
5.4.4	The Proof of the Theorem	152
5.5	Context	152
5.5.1	More Advanced Results	152
5.5.2	Density Estimation, Regression Estimation, and Regression with Deterministic Objects	153
5.5.3	Universal Probabilistic Predictors	154
5.5.4	Algorithmic Randomness Perspective	154
6	Probabilistic Classification: Venn Predictors	157
6.1	Introduction	157
6.2	Venn Predictors	158
6.2.1	Validity of One-Off Venn Predictors	159
6.2.2	Are There Other Ways to Achieve Perfect Calibration? ..	160
6.2.3	Venn Prediction with Binary Labels and No Objects	162
6.3	A Universal Venn Predictor	163
6.4	Venn–Abers Predictors	163
6.4.1	Full Venn–Abers Predictors	163
6.4.2	Inductive Venn–Abers Predictors	165
6.4.3	Probabilistic Predictors Derived from Venn Predictors ...	166
6.4.4	Cross Venn–Abers Predictors	168
6.4.5	Merging Multiprobability Predictions into a Probabilistic Prediction	169
6.5	Proofs	170
6.5.1	Proof of Theorem 6.4	170
6.5.2	PAVA and the Proof of Lemma 6.6	172
6.5.3	Proof of Proposition 6.7	173
6.6	Context	178
6.6.1	Risk and Uncertainty	178
6.6.2	John Venn, Frequentist Probability, and the Problem of the Reference Class	178
6.6.3	Online Venn Predictors Are Calibrated	179
6.6.4	Isotonic Regression	179
7	Probabilistic Regression: Conformal Predictive Systems	181
7.1	Introduction	181
7.2	Conformal Predictive Systems	182
7.2.1	Basic Definitions	182
7.2.2	Properties of Validity	185
7.2.3	Simplest Example: Monotonic Conformity Measures	186
7.2.4	Criterion of Being a CPS	187
7.3	Least Squares Prediction Machine	188

7.3.1	Three Kinds of LSPM	188
7.3.2	The Studentized LSPM in an Explicit Form	190
7.3.3	The Offline Version of the Studentized LSPM	191
7.3.4	The Ordinary LSPM	192
7.3.5	Asymptotic Efficiency of the LSPM	193
7.3.6	Illustrations	196
7.4	Kernel Ridge Regression Prediction Machine	198
7.4.1	Explicit Forms of the KRRPM	198
7.4.2	Limitation of the KRRPM	201
7.5	Nearest Neighbours Prediction Machine	202
7.6	Universal Conformal Predictive Systems	204
7.6.1	Definitions	205
7.6.2	Universal Conformal Predictive Systems	205
7.6.3	Universal Deterministic Predictive Systems	206
7.7	Applications to Decision Making	206
7.7.1	A Standard Problem of Decision Making	207
7.7.2	Examples	208
7.7.3	Asymptotically Efficient Decision Making	209
7.7.4	Dangers of Overfitting	209
7.8	Computationally Efficient Versions	210
7.8.1	Inductive Conformal Predictive Systems	210
7.8.2	Cross-Conformal Predictive Distributions	213
7.8.3	Practical Aspects	214
7.8.4	Beyond Randomness	215
7.9	Proofs and Calculations	216
7.9.1	Proofs for Sect. 7.2	216
7.9.2	Proofs for Sect. 7.3	217
7.9.3	Proof of Theorem 7.16	220
7.9.4	Proofs for Sect. 7.8	221
7.10	Context	222
7.10.1	Conformal Predictive Distributions	222
7.10.2	Conformal Predictive Distributions with Kernels	222
7.10.3	Venn Prediction for Probabilistic Regression	223
7.10.4	Universal Consistency and Universality	223
7.10.5	Various Notions of Convergence in Law	223
7.10.6	Decision Theory	223
7.10.7	Postprocessing of Predictive Distributions	224

Part III Testing Randomness

8	Testing Exchangeability	227
8.1	Testing Exchangeability	227
8.1.1	Exchangeability Supermartingales	228
8.1.2	Conformal Test Martingales	229
8.2	Testing for Concept and Label Shift in Anticausal Classification	233

8.3	Multistage Nonrandomness Detection	239
8.3.1	CUSUM-Type Change Detection	241
8.3.2	Shiryaev–Roberts Change Detection	241
8.4	When Do We Retrain?	243
8.4.1	The Ville Procedure for the Wine Quality Dataset	244
8.4.2	CUSUM and Shiryaev–Roberts for Wine Quality	248
8.4.3	Fixed and Variable Schedules	250
8.5	Proofs	255
8.5.1	Proof of Corollary 8.9	255
8.5.2	Proofs for Sect. 8.3	256
8.5.3	A Proof for Sect. 8.4	258
8.6	Context	259
8.6.1	Conformal Test Martingales	259
8.6.2	The Other Conformal Martingales	260
8.6.3	Optimality of Sequential Testing Procedures	261
8.6.4	Standard Procedures for Change Detection and Their Optimality	261
8.6.5	Deciding When to Retrain	262
8.6.6	Avoiding Numerical Problems	262
8.6.7	Low-Dimensional Dynamic Models	263
8.6.8	Islands of Randomness	263
9	Efficiency of Conformal Testing	265
9.1	Conformal Testing with a Known Critical Region	265
9.1.1	Randomness Probability vs Exchangeability Probability	266
9.1.2	Conformal Probability	267
9.2	Conformal Testing with a Known Alternative	269
9.2.1	Bayes–Kelly Algorithm	270
9.2.2	Bayes–Kelly Algorithm with a Composite Alternative ...	274
9.2.3	Changepoint Alternatives	275
9.2.4	More Natural Conformal Test Martingales for Changepoint Alternatives	282
9.2.5	Symmetric Markov Alternatives	284
9.2.6	Asymmetric Markov Alternatives	289
9.2.7	Mixing Over All Markov Alternatives	291
9.3	Testing the Validity of Putative Test Martingales	294
9.4	Proofs and Discussions	298
9.4.1	Proof of Proposition 9.2	298
9.4.2	Proof of Proposition 9.5	299
9.4.3	An Alternative Interpretation of the Bayes–Kelly Algorithm and the Proof of Proposition 9.9	300
9.5	Context	301
9.5.1	Randomness and Exchangeability: Algorithmic Perspective	301
9.5.2	Testing Exchangeability and Randomness	303

10	Non-conformal Shortcut	305
10.1	General Picture	305
10.2	Case of Regression	307
10.2.1	Probability Integral Transformation in Place of Conformal Prediction	307
10.2.2	The Simple Jumper Protection Procedure	308
10.2.3	A Simulation Study	310
10.3	Case of Classification	314
10.3.1	Testing Predictions by Betting	314
10.3.2	Protecting Prediction Algorithms	318
10.3.3	Experimental Results	319
10.3.4	Classification with the Brier Loss Function	321
10.3.5	Theoretical Guarantees	323
10.4	Complements and Proofs	324
10.4.1	Aggregating Algorithm	324
10.4.2	Proof of Theorem 10.3	326
10.5	Context	327
10.5.1	Regression	327
10.5.2	Classification	328
10.5.3	Standard Loss Functions	329
10.5.4	Aggregating Algorithm	330

Part IV Online Compression Modelling

11	Generalized Conformal Prediction	333
11.1	Introduction	333
11.2	Online Compression Models and Their Modifications	334
11.2.1	Online Compression Models	334
11.2.2	Conformal Transducers and Validity of OCM	337
11.2.3	Finite-Horizon Result	338
11.2.4	Repetitive Structures	339
11.2.5	One-Off Structures (OOS)	340
11.3	Exchangeability Model and Its Modifications	342
11.3.1	Exchangeability OOS	342
11.3.2	Conditional-Exchangeability OOS	342
11.3.3	Exchangeability OCM	343
11.3.4	Generality and Specificity	343
11.3.5	Inductive-Exchangeability Models	344
11.3.6	Mondrian-Exchangeability Models	345
11.3.7	Exchangeability with a Known Dataset Shift	345
11.4	Gaussian Model and Its Extensions	346
11.4.1	Gaussian Model	346
11.4.2	Gauss Linear Model	348
11.4.3	Multivariate Gaussian Model	349
11.4.4	Some Comparisons	351

11.5	Proofs	353
11.5.1	Proof of Theorem 11.2	353
11.5.2	Proof of Proposition 4.4	355
11.5.3	Proof of Theorem 8.8	355
11.5.4	Proof of Proposition 11.6	357
11.6	Context	358
11.6.1	Kolmogorov's Programme	359
11.6.2	Repetitive Structures	360
11.6.3	Exchangeability Model	361
11.6.4	Gaussian Models	362
11.6.5	Markov Model	362
12	Generalized Venn Prediction and Hypergraphical Models	363
12.1	Introduction	363
12.2	Venn Prediction and Venn-Conditional Conformal Prediction in One-Off Structures	364
12.2.1	Venn Prediction	364
12.2.2	Venn-Conditional Conformal Prediction	365
12.3	Hypergraphical Models	366
12.3.1	Hypergraphical Repetitive Structures	367
12.3.2	Generality of Finitary Repetitive Structures	368
12.3.3	Generality of Hypergraphical Models	368
12.3.4	Junction-Tree Models	369
12.3.5	Combinatorics of Junction-Tree Models	370
12.3.6	Shuffling Datasets	371
12.3.7	Decomposability in Junction-Tree Models	372
12.4	Venn Prediction in Hypergraphical Models	372
12.4.1	Prediction in Junction-Tree Models	373
12.4.2	Universality of the Fully Object-Conditional Venn Predictor	374
12.4.3	An Experiment with a Simple Causal Network	375
12.5	Conformal Prediction in Hypergraphical Models	378
12.5.1	Conditional Probability Conformity Measure	378
12.5.2	LED Datasets	379
12.5.3	Hypergraphical Assumptions for LED Datasets	380
12.5.4	Results for the LED Datasets	381
12.5.5	Label-Conditional Conformal Prediction for Hypergraphical OCMs	384
12.5.6	Further Results for the LED Datasets	385
12.5.7	Summary of the Experiments	388
12.6	Context	388
12.6.1	Conformal Prediction in Hypergraphical Models	388
12.6.2	Additive Models	389
13	Contrasts and Perspectives	391
13.1	Introduction	391

13.2	Inductive Learning	392
13.2.1	Jacob Bernoulli's Learning Problem	393
13.2.2	Statistical Learning Theory	396
13.2.3	The Quest for Data-Dependent Bounds	399
13.2.4	The Hold-Out Estimate	401
13.3	Transductive Learning	402
13.3.1	Student and Fisher	402
13.3.2	Conformal Predictive Distributions in the Gaussian Model	405
13.3.3	Tolerance Regions	405
13.3.4	Dempster-Hill Procedure	407
13.3.5	Transduction in Statistical Learning Theory	409
13.3.6	Validity and Efficiency in Probabilistic Prediction	412
13.4	Bayesian Learning	413
13.4.1	Bayesian Ridge Regression	413
13.4.2	An Illustration	415
13.5	Further Development of Conformal Prediction	417
13.5.1	Moving Beyond Exchangeability	418
13.5.2	Causality	418
13.5.3	Anomaly Detection	419
13.6	Proof of Proposition 13.3	420
13.7	Context	421
13.7.1	Inductive Prediction	421
13.7.2	Transductive Prediction	421
13.7.3	Online Transduction	422
13.7.4	Bayesian Prediction	422
13.7.5	Conformal Prediction: Further Developments	422
A	Probability Theory	423
A.1	Basics	423
A.1.1	Kolmogorov's Axioms	423
A.1.2	Convergence	425
A.2	Independence and Products	425
A.2.1	Products of Probability Spaces	426
A.2.2	Randomness Model	427
A.3	Expectations and Conditional Expectations	427
A.4	Markov Kernels and Regular Conditional Distributions	429
A.5	Exchangeability	431
A.5.1	De Finetti's Representation Theorem	431
A.5.2	Conditional Probabilities Given a Bag	432
A.6	Theory of Martingales	433
A.6.1	Basic Results	434
A.6.2	Limit Theorems	435
A.6.3	Hoeffding's Inequality	436
A.7	Bibliographical Remarks	437

A.7.1	Conditional Probabilities	438
A.7.2	Martingales	438
A.7.3	Hoeffding's Inequality.....	438
B	Datasets and Computations	439
B.1	USPS Dataset	439
B.2	Wine Quality Dataset	440
B.3	Bank Marketing Dataset	440
B.4	Normalization	440
B.5	Randomization and Reshuffling	441
B.6	Bibliographical Remarks.....	442
C	FAQ	443
C.1	Unusual Features of Conformal Prediction	443
C.2	Independence of Errors	444
C.3	Ways of Applying Conformal-Type Predictors	445
C.4	Conformal Prediction vs Standard Methods	446
	References	449
	Index	467