

Contents

Preface	xiii
I Opening	1
1 Introduction	3
1.1 Discovering the Laws of Nature	3
1.1.1 A Horse Named Clever Hans	4
1.1.2 Two Models of the Solar System	6
1.1.3 Lights and Shadows in AI Models	6
1.1.4 The Cow-on-the-Beach Problem	10
1.2 Book Outline	11
II Background	17
2 Minimizing Training Error	19
2.1 Introduction	19
2.2 Prediction and the Problem of Induction	21
2.2.1 The Problem of Induction	22
2.3 Empirical Risk Minimization	29
2.3.1 Linear and Logistic Regression	30
2.3.2 Vapnik-Chervonenkis Learning Theory	33
2.3.3 Universal Consistency, and the Lack Thereof	36
2.3.4 Distribution-Dependent Bounds	37
2.4 Zero Training Error in Modern Practice	38
2.4.1 Controlling Effective Capacity	39
2.4.2 Characterizing the Learned Minimizer	40
2.5 Simplify-and-Memorize	45
2.5.1 The Cow-on-the-Beach Problem	46
2.6 Failing to Generalize Out-of-Distribution	48

3	Theories of Causation	55
3.1	Introduction	55
3.2	Classic Theories of Causation	60
3.2.1	Ancient and Rationalist Philosophers	60
3.2.2	David Hume	61
3.2.3	Immanuel Kant	63
3.3	Regularity Theories	67
3.4	Interventionist Theories	70
3.4.1	Example: Hooke's Law	71
3.4.2	What Is an Intervention?	73
3.4.3	Range of Interventions	74
3.4.4	Autonomy of Interventions	75
3.4.5	Criticisms	77
3.5	Counterfactual Theories	79
3.5.1	Criticisms	80
3.6	Causation and Explanation	85
3.7	Causation and Time	86
3.8	Causation and Cognition	88
4	Practice of Causation	93
4.1	Introduction	93
4.2	Variable Selection	94
4.3	Structural Equation Models	97
4.3.1	The Causal Graph	99
4.3.2	Observational Distribution	100
4.3.3	Interventional Distributions	103
4.3.4	Counterfactual Distributions	104
4.3.5	Instrumental Variables	104
4.3.6	Do-Calculus	107
4.4	Causal Discovery	108
4.5	Potential Outcomes Framework	111
4.5.1	Naïve Difference Estimator	113
4.5.2	Eliminating Selection Bias by Randomization	113
4.5.3	Reducing Selection Bias by Valid Adjustment	115
4.5.4	Exploring a Linear Case	118
4.5.5	Limitations	118
4.6	Conclusion	122

5	The Invariance Principle	127
5.1	Introduction	127
5.2	What Is an Environment?	129
5.2.1	Interventions as Environments	132
5.2.2	Environments as Domains, Attributes, and Groups	134
5.3	Invariant Risk Minimization	135
5.3.1	A Linear Example	137
5.4	IRMv1, a Tractable Approximation	139
5.4.1	Posing IRM as a Regularized ERM	139
5.4.2	Controlling the Capacity of the Classifier	142
5.4.3	Controlling the Capacity of the Representation	142
5.4.4	Empirical Performance of IRMv1	143
5.4.5	Theoretical Analyses of IRMv1	145
5.5	Trade-offs and Pitfalls	148
5.5.1	Null Invariance	150
5.5.2	Delta Invariance	151
5.6	The Invariance Principle in Other Fields of Knowledge	152
5.6.1	Invariance in Mathematics	153
5.6.2	Invariance in Physics	154
5.6.3	Invariance and Metaphysics	157
5.6.4	Invariance and Relevance	158
5.6.5	Invariance and Truth	163
6	Domain Generalization Algorithms	165
6.1	Introduction	165
6.2	Invariant Causal Prediction	166
6.2.1	Methods Related to ICP and IRM	167
6.2.2	Causal Concepts Related to ICP and IRM	169
6.3	Distributionally Robust Optimization	171
6.4	Data Subsampling and Reweighting	172
6.5	Matching Representations	175
6.6	Invariance Through Time	176
6.7	Following Common Gradients	177
6.8	Model Evaluation and Selection	179
7	Discovering Environments	181
7.1	Introduction	181
7.2	Environment Discovery in Humans	183
7.3	Methods for Environment Discovery	185
7.4	Discovering Environments with XRM	187

7.4.1	Twin Setup, Holding-out of Data	187
7.4.2	Twin Training, Flipping Labels	190
7.4.3	Twin Model Selection, Counting Label Flips	191
7.4.4	Environment Discovery, via Cross-Mistake Formula	191
7.4.5	Relationship to the Invariance Principle	192
7.5	Performance of XRM	192
7.5.1	When Does XRM Work, and When Does It Fail?	195
7.6	Memorization, Twin Training, and Label Flipping	197
7.7	One Example, One Environment	202

8 Invariance in LLMs 205

8.1	Introduction	205
8.2	The Language Modeling Task	206
8.2.1	Next-token Prediction	206
8.2.2	Large Language Model Training Pipeline	208
8.2.3	The Scaling Hypothesis	208
8.3	The Transformer Architecture	209
8.3.1	Tokenization	210
8.3.2	Token Embeddings	211
8.3.3	Position Embeddings	211
8.3.4	Attention	213
8.3.5	Multilayer Perceptrons	215
8.3.6	Normalization	215
8.3.7	Skip Connections	215
8.3.8	Unembedding	216
8.3.9	Putting Everything Together	216
8.4	In-Context Learning	216
8.4.1	In-context Learning Versus in-Weight Learning	218
8.4.2	Context Is Environment	221
8.5	In-Context Risk Minimization	223
8.5.1	Theoretical Guarantees	224
8.5.2	Empirical Performance	226
8.5.3	ICRM and the Invariance Principle	227
8.6	Chain of Thought	229
8.7	Future Prediction	232
8.8	LLMs: Just Fiction Machines?	233

IV Related Technologies 235

9 Learning Diverse Features 237

9.1	Introduction	237
-----	------------------------	-----

9.2	Diverse Features via a Single Model	238
9.2.1	Spectral Decoupling	238
9.2.2	Dropout	239
9.2.3	Self-Supervised Learning	240
9.3	Diverse Features via Ensembles	242
9.3.1	Combining Model Predictions	244
9.3.2	Combining Model Features	245
9.3.3	Combining Model Weights	245
10	Learning from Combinations of Examples	247
10.1	Introduction	247
10.2	The mixup Data-Augmentation Technique	249
10.2.1	Why Does mixup Work?	252
10.2.2	Points of Care	257
10.3	Mixing on Hidden Layers	257
10.4	Mixing Not at Random	260
10.5	Mixing Unlabeled Data	262
10.6	Mixing by Extrapolation	263
10.7	Mixing at Test Time	264
11	Uncertainty Estimation	267
11.1	Introduction	267
11.2	Learning with Uncertainty	271
11.3	Two Types of Uncertainty	272
11.3.1	Aleatoric Uncertainty	273
11.3.2	Epistemic Uncertainty	274
11.4	Combining Uncertainty	278
11.4.1	Reject-or-classify	278
11.4.2	Unified Uncertainty Calibration	279
11.4.3	Theoretical Analysis of U2C	281
11.5	Empirical Analysis of U2C	282
11.6	Uncertainty in Sequential Environments	283
11.7	Uncertainty and Invariance	285
12	Fairness and Alignment	287
12.1	Introduction	287
12.2	The Status Quo in Learning Fair Classifiers	289
12.2.1	Evaluating Performance Metrics	290
12.2.2	Evaluating Fairness Metrics	290
12.3	Unfairness in-the-Wild: A Catalog	292
12.4	From Fairness to Invariance and Back	294
12.5	The Framework of Stress Environments	296

12.5.1	Flexibility	296
12.5.2	Subjectivity	297
12.5.3	Robustness	298
12.5.4	Limitations	299
12.6	Relation to Prior Work	300

V Closing 301

13 Esoterica 303

13.1	Introduction	303
13.2	Buddhist Teachings	305
13.2.1	The Doctrine of Dependent Origination	306
13.2.2	The Doctrine of No-Self	307
13.3	The Headless Way	311
13.4	Living by the Invariance Principle	315

Bibliography 319

Index 361