

Table of Contents

Preface	xi
---------	----

1

Introduction to Data Imbalance in Machine Learning			1
Technical requirements	2	Challenges and considerations when dealing with imbalanced data	19
Introduction to imbalanced datasets	2		
Machine learning 101	4	When can we have an imbalance in datasets?	20
What happens during model training?	7	Why can imbalanced data be a challenge?	20
Types of dataset and splits	8	When to not worry about data imbalance	23
Cross-validation	9	Introduction to the imbalanced-learn library	24
Common evaluation metrics	10	General rules to follow	27
Confusion matrix	10	Summary	28
ROC	15	Questions	29
Precision-Recall curve	17	References	30
Relation between the ROC curve and PR curve	18		

2

Oversampling Methods			33
Technical requirements	34	SMOTE	39
What is oversampling?	34	How SMOTE works	40
Random oversampling	36	Problems with SMOTE	42
Problems with random oversampling	39	SMOTE variants	43
		Borderline-SMOTE	43

ADASYN	47	Guidance for using various oversampling techniques	55
Working of ADASYN	47	When to avoid oversampling	56
Categorical features and SMOTE variants (SMOTE-NC and SMOTEN)	49	Oversampling in multi-class classification	57
Model performance comparison of various oversampling methods	54	Summary	59
		Exercises	60
		References	60

3

Undersampling Methods 63

Technical requirements	63	Tomek links	77
Introducing undersampling	64	Neighborhood Cleaning Rule	78
When to avoid undersampling the majority class	65	Instance hardness threshold	79
Fixed versus cleaning undersampling	66	Strategies for removing easy observations	80
Undersampling approaches	69	Condensed Nearest Neighbors	80
Removing examples uniformly	70	One-sided selection	82
Random UnderSampling	70	Combining undersampling and oversampling	82
ClusterCentroids	72	Model performance comparison	83
Strategies for removing noisy observations	74	Summary	86
ENN, RENN, and AllKNN	74	Exercises	86
		References	87

4

Ensemble Methods 89

Technical requirements	90	Boosting techniques for imbalanced data	102
Bagging techniques for imbalanced data	91	AdaBoost	103
UnderBagging	96	RUSBoost, SMOTEBoost, and RAMOBoost	104
OverBagging	97	Ensemble of ensembles	106
SMOTEBagging	99	EasyEnsemble	107
Comparative performance of bagging methods	101		

Comparative performance of boosting methods	109	Summary	113
		Questions	113
Model performance comparison	110	References	113

5

Cost-Sensitive Learning 115

Technical requirements	116	Cost-Sensitive Learning for decision trees	126
The concept of Cost-Sensitive Learning	116	Cost-Sensitive Learning using scikit-learn and XGBoost models	128
Costs and cost functions	116	MetaCost – making any classification model cost-sensitive	133
Types of cost-sensitive learning	117	Threshold adjustment	137
Difference between CSL and resampling	118	Methods for threshold tuning	140
Problems with rebalancing techniques	118	Summary	146
Understanding costs in practice	119	Questions	147
Cost-Sensitive Learning for logistic regression	120	References	147

6

Data Imbalance in Deep Learning 149

Technical requirements	150	Text analysis using Natural Language Processing	162
A brief introduction to deep learning	150	Data imbalance in deep learning	163
Neural networks	151	The impact of data imbalance on deep learning models	165
Perceptron	152	Overview of deep learning techniques to handle data imbalance	168
Activation functions	153	Multi-label classification	169
Layers	153	Summary	172
Feedforward neural networks	154	Questions	172
Training neural networks	155	References	172
The effect of the learning rate on data imbalance	159		
Image processing using Convolutional Neural Networks	160		

7

Data-Level Deep Learning Methods 175

Technical requirements	176	Document-level augmentation	202
Preparing the data	176	Character and word-level augmentation	203
Creating the training loop	178		
Sampling techniques for deep learning models	180	Discussion of other data-level deep learning methods and their key ideas	205
Random oversampling	180	Two-phase learning	205
Dynamic sampling	182	Expansive Over-Sampling	205
Data augmentation techniques for vision	185	Using generative models for oversampling	206
		DeepSMOTE	207
		Neural style transfer	208
Data-level techniques for text classification	199	Summary	209
Dataset and baseline model	201	Questions	209
		References	210

8

Algorithm-Level Deep Learning Techniques 213

Technical requirements	213	Class-dependent temperature Loss	235
Motivation for algorithm-level techniques	214	Class-wise difficulty-balanced loss	237
Weighting techniques	215	Discussing other algorithm-based techniques	239
Using PyTorch's weight parameter	216	Regularization techniques	239
Handling textual data	220	Siamese networks	239
Deferred re-weighting – a minor variant of the class weighting technique	224	Deeper neural networks	240
		Threshold adjustment	240
Explicit loss function modification	227	Summary	241
Focal loss	227	Questions	241
Class-balanced loss	232	References	242

9

Hybrid Deep Learning Methods 245

Technical requirements	246	Online Hard Example Mining	262
Using graph machine learning for imbalanced data	246	Minority class incremental rectification	264
Understanding graphs	246	Utilizing the hard sample mining technique in minority class incremental rectification	265
Graph machine learning	247	Summary	268
Dealing with imbalanced data	247	Questions	268
Case study – the performance of XGBoost, MLP, and a GCN on an imbalanced dataset	250	References	268
Hard example mining	261		

10

Model Calibration 271

Technical requirements	271	The calibration of model scores to account for sampling	286
Introduction to model calibration	271	Platt's scaling	288
Why bother with model calibration	273	Isotonic regression	289
Models with and without well-calibrated probabilities	273	Choosing between Platt's scaling and Isotonic regression	291
Calibration curves or reliability plot	274	Temperature scaling	291
Brier score	276	Label smoothing	291
Expected Calibration Error	277	The impact of calibration on a model's performance	294
The influence of data balancing techniques on model calibration	279	Summary	295
Plotting calibration curves for a model trained on a real-world dataset	282	Questions	296
Model calibration techniques	285	References	297

Appendix

Machine Learning Pipeline in Production 299

Machine learning training pipeline	299	Inferencing (online or batch)	301
------------------------------------	-----	-------------------------------	-----

Assessments	303
--------------------	------------

Chapter 1 – Introduction to Data Imbalance in Machine Learning	303	Chapter 7 – Data-Level Deep Learning Methods	308
Chapter 2 – Oversampling Methods	306	Chapter 8 – Algorithm-Level Deep Learning Techniques	308
Chapter 3 – Undersampling Methods	307	Chapter 9 – Hybrid Deep Learning Methods	309
Chapter 4 – Ensemble Methods	307	Chapter 10 – Model Calibration	310
Chapter 5 – Cost-Sensitive Learning	307		
Chapter 6 – Data Imbalance in Deep Learning	307		

Index	313
--------------	------------

Other Books You May Enjoy	322
----------------------------------	------------
