Data-Centric Machine Learning with Python : The Ultimate Guide to Engineering and Deploying High-Quality Models Based on Good Data /

Christensen, Jonas,
author

Monografía

Join the data-centric revolution and master the concepts, techniques, and algorithms shaping the future of AI and ML development, using Python Key Features Grasp the principles of data centricity and apply them to real-world scenarios Gain experience with quality data collection, labeling, and synthetic data creation using Python Develop essential skills for building reliable, responsible, and ethical machine learning solutions Purchase of the print or Kindle book includes a free PDF eBook Book Description In the rapidly advancing data-driven world where data quality is pivotal to the success of machine learning and artificial intelligence projects, this critically timed guide provides a rare, end-to-end overview of data-centric machine learning (DCML), along with hands-on applications of technical and non-technical approaches to generating deeper and more accurate datasets. This book will help you understand what data-centric ML/AI is and how it can help you to realize the potential of 'small data'. Delving into the building blocks of data-centric ML/AI, you'll explore the human aspects of data labeling, tackle ambiguity in labeling, and understand the role of synthetic data. From strategies to improve data collection to techniques for refining and augmenting datasets, you'll learn everything you need to elevate your data-centric practices. Through applied examples and insights for overcoming challenges, you'll get a roadmap for implementing data-centric ML/AI in diverse applications in Python. By the end of this book, you'll have developed a profound understanding of data-centric ML/AI and the proficiency to seamlessly integrate common data-centric approaches in the model development lifecycle to unlock the full potential of your machine learning projects by prioritizing data quality and reliability. What you will learn Understand the impact of input data quality compared to model selection and tuning Recognize the crucial role of subject-matter experts in effective model development Implement data cleaning, labeling, and augmentation best practices Explore common synthetic data generation techniques and their applications Apply synthetic data generation techniques using common Python packages Detect and mitigate bias in a dataset using best-practice techniques Understand the importance of reliability, responsibility, and ethical considerations in ML/AI Who this book is for This book is for data science professionals and machine learning enthusiasts looking to understand the concept of data-centricity, its benefits over a model-centric approach, and the practical application of a best-practice data-centric approach in their work. This book is also for other data professionals and senior leaders who want to explore the tools and techniques to improve data quality and create opportunities for small data ML/AI in their organizations

Título: Data-Centric Machine Learning with Python The Ultimate Guide to Engineering and Deploying High-Quality Models Based on Good Data Jonas Christensen, Nakul Bajaj and Manmohan Gosada ; foreword by Kirk D. Borne

Edición: 1st ed

Editorial: Birmingham, England Packt Publishing [2024] 2024

Descripción física: 1 online resource (378 pages)

Bibliografía: Includes bibliographical references and index

Contenido: Cover -- Title Page -- Copyright and Credits -- Foreword -- Contributors -- Table of Contents -- Preface -- Part 1: What Data-Centric Machine Learning Is and Why We Need It -- Chapter 1: Exploring Data-Centric Machine Learning -- Understanding data-centric ML -- The origins of data centricity -- The components of ML systems -- Data is the foundational ingredient -- Data-centric versus model-centric ML -- Data centricity is a team sport -- The importance of quality data in ML -- Identifying high-value legal cases with natural language processing -- Predicting cardiac arrests in emergency calls -- Summary -- References -- Chapter 2: From Model-Centric to Data-Centric - ML's Evolution -- Exploring why ML development ended up being mostly model-centric -- The 1940s to 1970s - the early days -- The 1980s to 1990s - the rise of personal computing and the internet -- The 2000s - the rise of tech giants -- 2010-now - big data drives AI innovation -- Model-centricity was the logical evolutionary outcome -- Unlocking the opportunity for small data ML -- Why we need data-centric AI more than ever -- The cascading effects of data quality -- Avoiding data cascades and technical debt -- Summary -- References -- Part 2: The Building Blocks of Data-Centric ML -- Chapter 3: Principles of Data-Centric ML -- Sometimes, all you need is the right data -- Principle 1 - data should be the center of ML development -- A checklist for data-centricity -- Principle 2 - leverage annotators and SMEs effectively -- Direct labeling with human annotators -- Verifying output quality with human annotators -- Codifying labeling rules with programmatic labeling -- Principle 3 - use ML to improve your data -- Principle 4 - follow ethical, responsible, and well-governed ML practices -- Summary -- References -- Chapter 4: Data Labeling Is a Collaborative Process Understanding the benefits of diverse human labeling -- Understanding common challenges arising from human labelers -- Designing a framework for high-quality labels -- Designing clear instructions -- Aligning motivations and using SMEs -- Collaborating iteratively -- Dealing with ambiguity and reflecting diversity -- Understanding approaches for dealing with ambiguity in labeling -- Measuring labeling consistency -- Summary -- References -- Part 3: Technical Approaches to Better Data -- Chapter 5: Techniques for Data Cleaning -- The six key dimensions of data quality -- Installing the required packages -- Introducing the dataset -- Ensuring the data is consistent -- Checking that the data is unique -- Ensuring that the data is complete and not missing -- Ensuring that the data is valid -- Ensuring that the data is accurate -- Ensuring that the data is fresh -- Summary -- Chapter 6: Techniques for Programmatic Labeling in Machine Learning -- Technical requirements -- Python version -- Library requirements -- Pattern matching -- Database lookup -- Boolean flags -- Weak supervision -- Semi-weak supervision -- Slicing functions -- Active learning -- Uncertainty sampling -- Query by Committee (QBC) -- Diversity sampling -- Transfer learning -- Feature extraction -- Fine-tuning pre-trained models -- Semi-supervised learning -- Summary -- Chapter 7: Using Synthetic Data in Data-Centric Machine Learning -- Understanding synthetic data -- The use case for synthetic data -- Synthetic data for computer vision and image and video processing -- Generating synthetic data using generative adversarial networks (GANs) -- Exploring image augmentation with a practical example -- Natural language processing -- Privacy preservation -- Generating synthetic data for privacy preservation -- Using synthetic data to improve model performance When should you use synthetic data? -- Summary -- References -- Chapter 8: Techniques for Identifying and Removing Bias -- The bias conundrum -- Types of bias -- Easy to identify bias -- Difficult to identify bias -- The data-centric imperative -- Sampling methods -- Other data-centric techniques -- Case study -- Loading the libraries -- AllKNN undersampling method -- Instance hardness undersampling method -- Oversampling methods -- Shapley values to detect bias, oversample, and undersample data -- Summary -- Chapter 9: Dealing with Edge Cases and Rare Events in Machine Learning -- Importance of detecting rare events and edge cases in machine learning -- Statistical methods -- Z-scores -- Interquartile Range (IQR) -- Box plots --

ISBN: 1-80461-241-3

Materia: Machine learning Python (Computer program language) Data mining

Autores: Bajaj, Nakul, author Gosada, Manmohan, author Borne, Kirk D., writer of foreword

Enlace a formato físico adicional: 1-80461-812-8

Baratz Innovación Documental

- Gran Vía, 59 28013 Madrid
- (+34) 91 456 03 60
- informa@baratz.es