Data Labeling in Machine Learning with Python : Explore Modern Ways to Prepare Labeled Data for Training and Fine-Tuning ML and Generative AI Models /

Suda, Vijaya Kumar,
author

Monografía

Take your data preparation, machine learning, and GenAI skills to the next level by learning a range of Python algorithms and tools for data labeling Key Features Generate labels for regression in scenarios with limited training data Apply generative AI and large language models (LLMs) to explore and label text data Leverage Python libraries for image, video, and audio data analysis and data labeling Purchase of the print or Kindle book includes a free PDF eBook Book Description Data labeling is the invisible hand that guides the power of artificial intelligence and machine learning. In today's data-driven world, mastering data labeling is not just an advantage, it's a necessity. Data Labeling in Machine Learning with Python empowers you to unearth value from raw data, create intelligent systems, and influence the course of technological evolution. With this book, you'll discover the art of employing summary statistics, weak supervision, programmatic rules, and heuristics to assign labels to unlabeled training data programmatically. As you progress, you'll be able to enhance your datasets by mastering the intricacies of semi-supervised learning and data augmentation. Venturing further into the data landscape, you'll immerse yourself in the annotation of image, video, and audio data, harnessing the power of Python libraries such as seaborn, matplotlib, cv2, librosa, openai, and langchain. With hands-on guidance and practical examples, you'll gain proficiency in annotating diverse data types effectively. By the end of this book, you'll have the practical expertise to programmatically label diverse data types and enhance datasets, unlocking the full potential of your data. What you will learn Excel in exploratory data analysis (EDA) for tabular, text, audio, video, and image data Understand how to use Python libraries to apply rules to label raw data Discover data augmentation techniques for adding classification labels Leverage K-means clustering to classify unsupervised data Explore how hybrid supervised learning is applied to add labels for classification Master text data classification with generative AI Detect objects and classify images with OpenCV and YOLO Uncover a range of techniques and resources for data annotation Who this book is for This book is for machine learning engineers, data scientists, and data engineers who want to learn data labeling methods and algorithms for model training. Data enthusiasts and Python developers will be able to use this book to learn data exploration and annotation using Python libraries. Basic Python knowledge is beneficial but not necessary to get started

Título: Data Labeling in Machine Learning with Python Explore Modern Ways to Prepare Labeled Data for Training and Fine-Tuning ML and Generative AI Models Vijaya Kumar Suda

Edición: 1st ed

Editorial: Birmingham, UK Packt Publishing [2024] 2024

Descripción física: 1 online resource (398 pages)

Nota general: Includes index

Contenido: Cover -- Title Page -- Copyright -- Acknowledgments -- Contributors -- Table of Contents -- Preface -- Part 1: Labeling Tabular Data -- Chapter 1: Exploring Data for Machine Learning -- Technical requirements -- EDA and data labeling -- Understanding the ML project life cycle -- Defining the business problem -- Data discovery and data collection -- Data exploration -- Data labeling -- Model training -- Model evaluation -- Model deployment -- Introducing Pandas DataFrames -- Summary statistics and data aggregates -- Summary statistics -- Data aggregates of the feature for each target class -- Creating visualizations using Seaborn for univariate and bivariate analysis -- Univariate analysis -- Bivariate analysis -- Profiling data using the ydata-profiling library -- Variables section -- Interactions section -- Correlations -- Missing values -- Sample data -- Unlocking insights from data with OpenAI and LangChain -- Summary -- Chapter 2: Labeling Data for Classification -- Technical requirements -- Predicting labels with LLMs for tabular data -- Data labeling using Snorkel -- What is Snorkel? -- Why is Snorkel popular? -- Loading unlabeled data -- Creating the labeling functions -- Labeling rules -- Constants -- Labeling functions -- Creating a label model -- Predicting labels -- Labeling data using the Compose library -- Labeling data using semi-supervised learning -- What is semi-supervised learning? -- What is pseudo-labeling? -- Labeling data using K-means clustering -- What is unsupervised learning? -- K-means clustering -- Inertia -- Dunn's index -- Summary -- Chapter 3: Labeling Data for Regression -- Technical requirements -- Using summary statistics to generate housing price labels -- Finding the closest labeled observation to match the label -- Using semi-supervised learning to label regression data -- Pseudo-labeling Using data augmentation to label regression data -- Using k-means clustering to label regression data -- Summary -- Part 2: Labeling Image Data -- Chapter 4: Exploring Image Data -- Technical requirements -- Visualizing image data using Matplotlib in Python -- Loading the data -- Checking the dimensions -- Visualizing the data -- Checking for outliers -- Performing data preprocessing -- Checking for class imbalance -- Identifying patterns and relationships -- Evaluating the impact of preprocessing -- Practice example of visualizing data -- Practice example for adding annotations to an image -- Practice example of image segmentation -- Practice example for feature extraction -- Analyzing image size and aspect ratio -- Impact of aspect ratios on model performance -- Image resizing -- Image normalization -- Performing transformations on images - image augmentation -- Summary -- Chapter 5: Labeling Image Data Using Rules -- Technical requirements -- Labeling rules based on image visualization -- Image labeling using rules with Snorkel -- Weak supervision -- Rules based on the manual visualization of an image's object color -- Real-world applications -- A practical example of plant disease detection -- Labeling images using rules based on properties -- Bounding boxes -- Example 1 - image classification - a bicycle with and without a person -- Example 2 - image classification - dog and cat images -- Labeling images using transfer learning -- Example - digit classification using a pre-trained classifier -- Example - person image detection using the YOLO V3 pre-trained classifier -- Example - bicycle image detection using the YOLO V3 pre-trained classifier -- Labeling images using transformations -- Summary -- Chapter 6: Labeling Image Data Using Data Augmentation -- Technical requirements -- Training support vector machines with augmented image data Kernel trick -- Data augmentation -- Image data augmentation -- Implementing an SVM with data augmentation in Python -- Introducing the CIFAR-10 dataset -- Loading the CIFAR-10 dataset in Python -- Preprocessing the data for SVM training -- Implementing an SVM with the default hyperparameters -- Evaluating SVM on the original dataset -- Implementing an SVM with an augmented dataset -- Training the SVM on augmented data -- Evaluating the SVM's performance on the augmented dataset -- Image classification using the SVM with data augmentation on the MNIST dataset -- Convolutional neural networks using augmented image data -- How CNNs work -- Practical example of a CNN using data augmentation -- CNN using image data augmentation with the CIFAR-10 dataset -- Summary -- Part 3: Labeling Text, Audio, and Video Data

-- Chapter 7: Labeling Text Data -- Technical requirements -- Real-world applications of text data labeling -- Tools and frameworks for text data labeling -- Exploratory data analysis of text -- Loading the data -- Understanding the data -- Cleaning and preprocessing the data -- Exploring the text's content -- Analyzing relationships between text and other variables -- Visualizing the results -- Exploratory data analysis of sample text data set -- Exploring Generative AI and OpenAI for labeling text data -- GPT models by OpenAI -- Zero-shot learning capabilities -- Text classification with OpenAI models -- Data labeling assistance -- OpenAI API overview -- Use case 1 - summarizing the text -- Use case 2 - topic generation for news articles -- Use case 3 - classification of customer queries using the user-defined categories and sub-categories -- Use case 4 - information retrieval using entity extraction -- Use case 5 - aspect-based sentiment analysis -- Hands-on labeling of text data using the Snorkel API Hands-on text labeling using Logistic Regression -- Hands-on label prediction using K-means clustering -- Generating labels for customer reviews (sentiment analysis) -- Summary -- Chapter 8: Exploring Video Data -- Technical requirements -- Loading video data using cv2 -- Extracting frames from video data for analysis -- Extracting features from video frames -- Color histogram -- Optical flow features -- Motion vectors -- Deep learning features -- Appearance and shape descriptors -- Visualizing video data using Matplotlib -- Frame visualization -- Temporal visualization -- Motion visualization -- Labeling video data using k-means clustering -- Overview of data labeling using k-means clustering -- Example of video data labeling using k-means clustering with a color histogram -- Advanced concepts in video data analysis -- Motion analysis in videos -- Object tracking in videos -- Facial recognition in videos -- Video compression techniques -- Real-time video processing -- Video data formats and quality in machine learning -- Common issues in handling video data for ML models -- Troubleshooting steps -- Summary -- Chapter 9: Labeling Video Data -- Technical requirements -- Capturing real-time video -- Key components and features -- A hands-on example to capture real-time video using a webcam -- Building a CNN model for labeling video data -- Using autoencoders for video data labeling -- A hands-on example to label video data using autoencoders -- Transfer learning -- Using the Watershed algorithm for video data labeling -- A hands-on example to label video data segmentation using the Watershed algorithm -- Computational complexity -- Performance metrics -- Real-world examples for video data labeling -- Advances in video data labeling and classification -- Summary -- Chapter 10: Exploring Audio Data -- Technical requirements Real-life applications for labeling audio data -- Audio data fundamentals -- Hands-on with analyzing audio data -- Example code for loading and analyzing sample audio file -- Best practices for audio format conversion -- Example code for audio data cleaning -- Extracting properties from audio data -- Tempo -- Chroma features -- Mel-frequency cepstral coefficients (MFCCs) -- Zero-crossing rate -- Spectral contrast -- Considerations for extracting properties -- Visualizing audio data with matplotlib and Librosa -- Waveform visualization -- Loudness visualization -- Spectrogram visualization -- Mel spectrogram visualization -- Considerations for visualizations -- Ethical implications of audio data -- Recent advances in audio data analysis -- Troubleshooting common issues during data analysis -- Troubleshooting common installation issues for audio libraries -- Summary -- Chapter 11: Labeling Audio Data -- Technical requirements -- Downloading FFmpeg -- Azure Machine Learning -- Real-time voice classification with Random Forest -- Transcribing audio using the OpenAI Whisper model -- Step 1 - importing the Whisper model -- Step 2 - loading the base Whisper model -- Step 3 - setting up FFmpeg -- Step 4 - transcribing the YouTube audio using the Whisper model -- Classifying a transcription using Hugging Face transformers -- Hands-on - labeling audio data using a CNN -- Exploring audio data augmentation -- Introducing Azure Cognitive Services - the speech service -- Creating an Azure Speech service -- Speech to text -- Speech translation -- Summary -- Chapter 12: Hands-On Exploring Data Labeling Tools -- Technical requirements -- Azure Machine Learning data labeling -- Label Studio -- pyOpenAnnotate -- Data labeling using Azure Machine Learning -- Benefits of data labeling with Azure Machine Learning -- Data labeling steps using Azure Machine Learning Image data labeling with Azure Machine Learning

ISBN: 1-80461-378-9

Materia: Python (Computer program language) Machine learning Computer programming

Enlace a formato físico adicional: 9781804610541

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