Introduction to Machine Learning

Data Collection

Andres Mendez-Vazquez

August 12, 2025

Outline

Outline

Introduction

Data collection is foundational to modern technologies

- However, it poses significant technical challenges.

Introduction

Data collection is foundational to modern technologies

- However, it poses significant technical challenges.

Data Quality Issues

- Incomplete Data: Missing values, incomplete records.
- Inaccurate Data: Measurement errors, incorrect labeling.
- Outdated Data: Data may no longer represent the current state.
- Noisy Data: Contains irrelevant or random variation.

Furthermore

Bias in Data Collection

- Selection Bias: Not all groups are equally represented (e.g., only surveying urban areas).
- Sampling Bias: Non-random sampling leads to distorted results.
- Observer/Interviewer Bias: The person collecting data influences the results.
- Reporting Bias: Only certain outcomes are recorded or shared.

Furthermore

Bias in Data Collection

- Selection Bias: Not all groups are equally represented (e.g., only surveying urban areas).
- Sampling Bias: Non-random sampling leads to distorted results.
- Observer/Interviewer Bias: The person collecting data influences the results.
- Reporting Bias: Only certain outcomes are recorded or shared.

Ethical and Privacy Concerns

- Informed Consent: Data subjects may not be aware their data is collected.
- Sensitive Information: Personally identifiable or health-related data.
- Surveillance Concerns: Data collected without consent (e.g., tracking).
- Data Ownership: Who owns the collected data? Who can access it?

Not less important

Logistical & Technical Challenges

- Cost: Data collection can be expensive and time-consuming.
- Access: Some data is behind paywalls or not publicly available.
- Scale: Large-scale data may require special infrastructure (e.g., IoT or satellite data).
- Real-time vs Batch: Hard to collect in real-time or with low latency.

Not less important

Logistical & Technical Challenges

- Cost: Data collection can be expensive and time-consuming.
- Access: Some data is behind paywalls or not publicly available.
- Scale: Large-scale data may require special infrastructure (e.g., IoT or satellite data).
- Real-time vs Batch: Hard to collect in real-time or with low latency.

Data Integration Problems

- Inconsistent Formats: Combining data from different sources can be messy.
- Duplicate Entries: Same data may appear multiple times.
- Different Standards: Units, naming conventions, and structures may vary.

Outline

What is Computer Vision?

What is Computer Vision?

- Enables machines to interpret and understand visual information.
- Mimics human vision using digital images and videos.
- Intersection of AI, ML, and image processing.

What is Computer Vision?

What is Computer Vision?

- Enables machines to interpret and understand visual information.
- Mimics human vision using digital images and videos.
- Intersection of AI, ML, and image processing.

What is Image Processing?

- Enhancing or analyzing digital images.
- Low-level processing: noise removal, contrast enhancement.
- Key tool in computer vision systems.

Outline

Image Preprocessing

- Grayscale conversion
- Gaussian blur for noise reduction
- Thresholding
- Morphological operations

Edge Detection

- Highlights object boundaries.
- Algorithms: Sobel, Canny, Laplacian.
- Used in object recognition and segmentation.

Image Filtering

- Low-pass filters (blurring)
- High-pass filters (sharpening)
- Kernel-based convolution techniques

Image Segmentation

- Partitioning an image into multiple regions.
- Techniques: Watershed, K-means, U-Net (deep learning)
- Essential for tasks like medical imaging and scene understanding.