Machine Learning

Free Machine Learning Training – Session **5**

How to Access Datasets

In this session, we will discuss how to access datasets. As we have repeatedly emphasized, the field of machine learning heavily relies on datasets, which are used to train models and make accurate predictions.

What is a dataset?

Datasets are collections of data organized in a specific format.

They can include any type of data, such as a series of numbers arranged sequentially (also called an array) or a table of various data points.

Types of Data in a Dataset

01 > Numerical Data

Examples include house prices (e.g., 20,000,000) or temperature (e.g., 40).

02 > Categorical Data

Examples include yes/no (like the "Purchased" column in the table above), male/female, blood type, etc.

03 > Ordinal Data

These are similar to categorical data but can be ordered or ranked. For example, educational degrees (diploma, bachelor's, master's, Ph.D.) have a specific order

Types of Datasets

Image Datasets

ImageNet

CIFAR-10

MNIST

Text Datasets

Gutenberg Task Dataset

IMDb Film Reviews Dataset

Time Series Datasets

Stock market data

Weather data

Sensor readings

Tabular Datasets

These are organized in tables or spreadsheets, with rows representing samples and columns representing features. They are used for tasks like regression and classification.

Why Do We Need Datasets?

As mentioned earlier, preprocessed and ready-to-use datasets are crucial for machine learning projects.

They provide the foundation for training accurate and reliable models. However, working with large datasets can pose challenges in terms of management and processing.

To address these challenges, we need to apply certain techniques to the dataset

Data Preprocessing

Data preprocessing is a critical step in preparing datasets for machine learning. It involves transforming raw data into a format suitable for training models.

Common preprocessing techniques include:

- Cleaning data to remove errors.
- Standardizing data to scale it within a specific range.
- Scaling features to ensure they have similar ranges.
- Handling missing values through imputation or deletion.

Splitting Datasets

Training Dataset	Test Dataset
This is the part of the dataset used to train the machine learning model.	This is the part of the dataset used to evaluate the model's performance.

Popular Sources for Machine Learning Datasets

- Kaggle Datasets
- UCI Machine Learning Repository
- AWS Public Datasets
- Google Dataset Search
- Microsoft Research Open Data
- Awesome Public Datasets (GitHub)
- Government Datasets
- Computer Vision Datasets
- Scikit-learn Datasets

“

💡 Mastering Machine Learning requires dedication, practice, and continuous learning.

Stay with ComeToMachine ♥

Thank you!

Do you have any questions?

Alizadeh.c2m@gmail.com
ComeToMachine.ir


ComeToMachine