# Machine Learning

## Week 1 • Class 2

### Core Terminologies

### Features, Labels, Models, Training vs Testing

## Class Objective

- Understand key machine learning terminologies
- Identify features and labels in a dataset
- Explain what a model is
- Distinguish between training and testing phases

## Why Terminology Matters

Machine learning relies heavily on precise terminology.

Understanding core terms helps you:

- Read ML papers and documentation
- Communicate effectively with teams
- Design correct ML pipelines
- Avoid common beginner mistakes

## What Is Data in Machine Learning?

In machine learning, **data** is the foundation.

Data typically consists of:

- Inputs (features)
- Outputs (labels)
- Examples (rows or records)

The quality of data directly impacts model performance.

# Features

## What Are Features?

**Features** are individual measurable properties or characteristics of data.

They represent the **inputs** given to a machine learning model.

## Examples of Features

| Problem | Features |
|---|---|
| House price prediction | Area, number of rooms, location |
| Email spam detection | Word frequency, sender, subject |
| Student performance | Attendance, assignment scores |

Features are usually represented as columns in a dataset.

## Feature Types (High-Level)

Common feature types include:

- Numerical (age, salary, marks)
- Categorical (gender, city, product type)
- Binary (yes/no, true/false)
- Text and image-based features

Feature selection is a critical ML step.

# Labels

## What Are Labels?

**Labels** are the outcomes or target values the model is trying to predict.

They represent the **correct answer** for supervised learning.

## Examples of Labels

| Problem | Label |
|---|---|
| House price prediction | House price |
| Email classification | Spam or Not Spam |
| Disease detection | Disease present or not |

Labels are usually the final column in a dataset.

## Features vs Labels

| Features | Labels |
|---|---|
| Input variables | Output variable |
| Given to model | Predicted by model |
| Independent | Dependent |
| Multiple per dataset | Usually one |

Understanding this distinction is essential.

## What Is a Model?

A **machine learning model** is a mathematical representation that:

- Learns patterns from data
- Maps features to labels
- Makes predictions on new data

The model contains learned parameters.

## Examples of Models

Common machine learning models include:

- Linear Regression
- Decision Trees
- k-Nearest Neighbors
- Support Vector Machines
- Neural Networks

Different problems require different models.

# Training a Model

## What Is Training?

**Training** is the process where a model:

- Learns from historical data
- Adjusts internal parameters
- Minimizes prediction error

Training uses labeled data.

## Training Data

Training data:

- Is the largest portion of the dataset
- Contains both features and labels
- Is used to teach the model patterns

Typically 70–80% of data is used for training.

# Testing a Model

## What Is Testing?

**Testing** evaluates how well the model performs on **unseen data**.

The model does not learn during testing.

## Testing Data

Testing data:

- Is kept separate from training data
- Contains known labels
- Measures model accuracy and reliability

Typically 20–30% of data is used for testing.

## Training vs Testing

| Training | Testing |
|---|---|
| Model learns | Model evaluates |
| Uses majority of data | Uses held-out data |
| Adjusts parameters | No parameter updates |
| Risk of overfitting | Checks generalization |

## Why Split Data?

Data is split to:

- Prevent memorization
- Measure real-world performance
- Detect overfitting
- Ensure fairness in evaluation

Without splitting, results are misleading.

## Overfitting (Concept Preview)

Overfitting occurs when:

- Model performs very well on training data
- Model performs poorly on testing data

This will be covered in detail in later classes.

## Class Summary

- Features are input variables
- Labels are target outputs
- Models learn patterns from data
- Training teaches the model
- Testing evaluates performance

## Next Class

## Types of Learning Problems – Classification