Machine Learning: Understanding Data

Vikas Thammanna Gowda

01/23/2025

1 Types of Data

1.1 Nominal (Categorical) Data

Definition: Nominal data represents categories or labels without any intrinsic order or ranking. These categories are mutually exclusive and do not have a quantitative value.
Characteristics:

- No natural order or hierarchy.
- Cannot perform arithmetic operations on them.
- Often represented as text or numbers assigned as labels.

Examples:

- Gender: Male, Female, Non-binary
- Color: Red, Blue, Green
- Country of Origin: USA, Canada, India

1.2 Ordinal Data

Definition: Ordinal data represents categories with a meaningful order or ranking, but the intervals between the categories are not uniform or measurable.
Characteristics:

- Can be ordered or ranked.
- Differences between ranks are not consistent or meaningful.

Examples:

- Education Level: High School < Bachelor's < Master's < PhD
- Customer Satisfaction: Very Unsatisfied < Unsatisfied < Neutral < Satisfied < Very Satisfied
- Socioeconomic Status: Low < Middle < High

1.3 Numerical Data

Numerical data, also known as quantitative data, represents measurable quantities and can be divided into two subcategories:

a) Continuous Data

Definition: Continuous data can take any value within a range and can be divided into smaller fractions.
Examples:

- Height: 5.4 ft, 6.2 ft
- Weight: 68.5 kg, 72.3 kg
- Temperature: 98.6°F, 37.5°C

b) Discrete Data

Definition: Discrete data consists of countable values, typically integers.
Examples:

- Number of students in a class: 25, 30, 32
- Number of cars in a parking lot: 10, 15, 20
- Test scores: 85, 90, 95

1.4 Binary Data

Definition: Binary data is a type of categorical data with only two possible values or states, often represented as 0 and 1.
Characteristics:

- Typically used for "yes/no", "true/false", or "present/absent" types of data.
- Can be thought of as a special case of nominal data.

Examples:

- Gender (simplified): Male (0), Female (1)
- Light Switch: Off (0), On (1)
- Loan Status: Approved (1), Denied (0)

1.5 Summary Table

Data Type	**Definition**	**Examples**
Nominal	Categories without a meaningful order.	Gender, Colors, Country
Ordinal	Categories with a meaningful order.	Education Levels, Satisfaction Levels, Rankings
Numerical	Measurable quantities (discrete/continuous).	Height, Weight, Number of students, Test scores
Binary	Two possible states (0/1).	Yes/No, On/Off, True/False

1.6 Exercise

1. Which of the following is an example of nominal data?
   (a) Test scores (85, 90, 95)
   (b) Height (5.4 ft, 6.2 ft)
   (c) Colors (Red, Blue, Green)
   (d) Customer satisfaction levels (Very Unsatisfied, Neutral, Very Satisfied)
2. What is a characteristic of ordinal data?
   (a) It has measurable intervals between values.
   (b) It cannot be ordered or ranked.
   (c) It represents categories with a meaningful order.
   (d) It has only two possible states (e.g., 0 and 1).
3. Which of the following represents discrete data?
   (a) Temperature: 98.6°F
   (b) Number of students in a class: 25
   (c) Height: 6.2 ft
   (d) Weight: 68.5 kg
4. Binary data is often represented as:
   (a) Integers and floats
   (b) A range of continuous values
   (c) Two possible states, such as 0 and 1
   (d) Multiple categories without hierarchy
5. What differentiates nominal data from ordinal data?
   (a) Nominal data is quantitative, while ordinal data is qualitative.
   (b) Ordinal data has a meaningful order, while nominal data does not.
   (c) Nominal data can take fractional values, while ordinal data cannot.
   (d) Both are the same type of data.
6. Which of the following is continuous data?
   (a) Number of cars in a parking lot: 15
   (b) Temperature: 37.5°C
   (c) Test scores: 90
   (d) Customer satisfaction level: Satisfied

2 Key Concepts in Data Analysis

2.1 Unit of Observation

Definition: The unit of observation is the entity or object being measured or analyzed in a dataset. It refers to what each row in a dataset represents.
Key Points:

- Determines the level of detail or granularity of the data.
- Units of observation can be individuals, groups, events, or objects depending on the dataset.

Examples:

- **Individuals:** A dataset about students may have each student as the unit of observation.
- **Events:** A dataset on car accidents may have each accident as the unit of observation.
- **Groups:** A dataset on countries' GDP may have each country as the unit of observation.

2.2 Instance

Definition: An instance (or record, data point, observation) is a single occurrence or row in the dataset, representing one unit of observation.
Key Points:

- Each instance corresponds to one unit of observation.
- In a tabular dataset, each row represents one instance.

Examples:

- **Students Dataset:** Each row is an instance, e.g., a single student.
- **House Prices Dataset:** Each row is an instance representing one house.

2.3 Features

Definition: Features (or attributes, variables, columns) are the properties, characteristics, or attributes describing each instance. These are the measurable inputs or dimensions of the data.
Key Points:

- Features are represented as columns in a dataset.
- Features can be numerical (e.g., height, age) or categorical (e.g., gender, color).

Examples:

- **Students Dataset:**
  - **Features:** Name, Age, Grade, Gender.
  - **Instance:** A single student with specific values for these features (e.g., "John, 16, 11th Grade, Male").
- **Cars Dataset:**
  - **Features:** Make, Model, Year, Price.
  - **Instance:** A specific car (e.g., "Toyota, Corolla, 2020, $20,000").

2.4 Relationship Between the Terms

- The **unit of observation** describes what each row represents in the dataset.
- An **instance** is a single row in the dataset, corresponding to one unit of observation.
- **Features** are the attributes (columns) describing each instance.

2.5 Example Dataset

Name	Age	Grade	Gender
John	16	11th	Male
Sarah	15	10th	Female
Michael	17	12th	Male

- **Unit of Observation:** Students (each row represents a student).
- **Instance:** Each row in the table (e.g., John, Sarah, Michael).
- **Features:** Name, Age, Grade, Gender.

2.6 Exercise

1. The unit of observation in a dataset refers to:
   (a) The entity or object being analyzed in the dataset.
   (b) The attributes describing each instance.
   (c) The number of columns in the dataset.
   (d) The unique ID assigned to each instance.
2. In a dataset of house prices, the instance represents:
   (a) A feature such as the house size.
   (b) A row corresponding to a specific house.
   (c) The entire dataset.
   (d) The unit of observation.
3. Features in a dataset are:
   (a) Rows in a dataset.
   (b) The properties or attributes describing each instance.
   (c) Entities being measured.
   (d) The summary of all observations.
4. In a dataset about students, which of the following is a feature?
   (a) Name, Age, Grade
   (b) Each student
   (c) The dataset itself
   (d) A specific row in the dataset
5. What is the relationship between the terms unit of observation, instance, and features?
   (a) Instances are columns, features are rows, and the unit of observation is the dataset.
   (b) Features are columns, instances are rows, and the unit of observation is what rows represent.
   (c) The unit of observation is a column, and features are rows.
   (d) All three terms refer to the same concept.
6. In a dataset where each row represents a car, which is true?
   (a) The unit of observation is “cars”.
   (b) Features could include “Make” and “Model”.
   (c) Each row is an instance.
   (d) All of the above.