Probability&Statistics for ML Notes

Ahmed Yasser

September 2025

Chapter 1

1.1.1 The Interplay Between Probability, Statistics, and Machine Learning

Remark (Difference between probability and statistics)

The difference between **probability** and **statistics** is that between **modeling the likelihood of future events** and **analyzing the frequency of past events**. Probability theory is useful for modeling **expected outcomes**, whereas statistics is useful for analyzing **sample outcomes**. The sample outcomes from a probabilistic model are only approximations of the probabilistically expected outcomes.

1.2 Representing Data

Remark

The simplest form of data used in statistics and machine learning is **tabular data**, which is also referred to as **multidimensional data**. This data typically contains a set of **observations**, which are represented by **rows** in the data table. Each observation contains a set of **fields**, which are represented by **columns** in the data table, and they describe the difefrent properties of the specific observation (row) at hand.

Note that:

- An observation is also reffered to as a **data point**, **database record**, **instance**, **example**, **transaction**, **entity**, **tuple**, **object**, **sample**, or **feature vector**.
- Fields are also referred to as **attributes**, **dimensions**, **variables**, or **features**.

Definition (Multidimensional Data)

A multidimensional data set is an $m \times n$ data matrix $D = \left[x_{ij}\right]_{m\times n}$, which may also be represented by a table containing m rows and n columns. The m rows of the data matrix are denoted by the m row vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m$, such that each vector $\boldsymbol{x}_i$ contains a set of n features or variables denoted by $[x_{i1} \dots x_{in}]$. Each vector is also referred to as an **observation**.

Remark

A data set with a single dimension ($n = 1$) is referred to as **univariate data**, and multidimensional data set with $n > 1$ is referred to as **multivariate data**.

Definition (Univariate Analysis and Multivariate Analysis)

When the analysis is performed on a single variable, it is referred to as **univariate analysis**. On the other hand, when the analysis is performed on multiple variables together, it is referred to as **multivariate analysis**.

Remark (Attributes Data Types)

Attributes can be of different types, such as:

- **Numerical**: Attributes with values that have a **natural ordering** and **quantifiable distances** between values. They can be further classified into:
  - **Discrete**: Attributes with values that are **countable** and **finite**.
  - **Continuous**: Attributes with values that are **uncountable** and **infinite**.

- **Categorical**: Attributes that take values from a **finite set of categories** without inherent numeric meaning. They can be further classified into:
  - **Nominal**: Categories with **no ordering** or ranking.
  - **Ordinal**: Categories with a **meaningful ordering** but **no quantifiable distances**.

Remark

Categorical data can be transformed to **binary numerical** data through **one-hot encoding**, while numerical data can be transformed to **binary numerical** data through **discretization**.

1.3 Summarizing and Visualizing Data

Definition (Summary Statistics)

Summary statistics are used to describe the main characteristics of a data set. Two common forms of summaries of the data distribution include:

- **Measures of Central Tendency**: Identify representative points corresponding to central regions of the data. Common measures include **arithmetic mean**, **median**, and **mode**.
- **Measures of Dispersion**: Model the degree of spread of the distribution from the center of the data. Common measures include **variance**, **standard deviation**, **range**, and **interquartile range**.

Note that the measures of central tendency and dispersion are **univariate summary statistics** because they are based on only one dimension. Several forms of summary statistics, such as **covariance** and **correlation** provide an idea of how different attributes are related to one another.

Remark

While summary statistics offer a concise, numerical description of data distribution, they are often insufficient to capture the underlying patterns and relationships. Therefore, data visualization is necessary for a comprehensive understanding.

We use different types of visualizations depending on the **number of variables** being analyzed:

- **For a single variable**, we use **univariate visualizations**. A **histogram**, for instance, is a common example that partitions the variable into bins and plots a frequency measure (like raw or relative frequency) for each bin.
- To capture the relationship between **two or more variables**, we use **multivariate visualizations**. A **scatterplot** is a key example, where the attributes are represented along the axes and each observation is plotted as a marker.