Machine Learning

Introduction to Machine Learning with Statistics

Introduction

- **In this Tutorial we are going to Cover –**
  - The Nature of Probability and Statistics
  - Frequency Distributions and Graphs
  - Data Description
  - Probability and Counting Rules
  - Hypothesis Testing
  - Machine Learning Introduction
  - Real Life Application of Machine Learning
  - Preprocessing Structured Data
  - Feature Selection and Engineering
  - Model Selection and Evaluation
  - Correlation and Regression
  - Trees and Forests
  - Nearest Neighbors
  - Support Vector Machines
  - Naïve Bayes
  - Clustering
  - Data Mining

  Etc.

Machine Learning

Descriptive and Inferential Statistics

Descriptive & Inferential Statistics

- **Definition of Statistics:**
  - **Statistics** is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

- **Definition of a Variable:**
  - A variable is a characteristic or attribute that can assume different values.

  - **Data** are the values (measurements or observations) that the variables can assume. Variables whose values are determined by chance are called **random variables**.

  - A collection of data values forms a **data set**. Each value in the data set is called a data value or a **datum.**

Descriptive & Inferential Statistics

- Data can be used in different ways. The body of knowledge called statistics is sometimes divided into two main areas, depending on how data are used. The two areas are
  1. Descriptive statistics
  2. Inferential statistics

- **Descriptive statistics:**
  - This statistics consists of the collection, organization, summarization, and presentation of data.
  - It can present data in some meaningful form, such as charts, graphs, or tables.

Descriptive & Inferential Statistics

- **Inferential statistics:**
  - **Inferential statistics** consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.
  - Here, the statistician tries to make inferences from *samples* to *populations.* Inferential statistics uses **probability,** i.e., the chance of an event occurring.
  - If you play cards, dice, bingo, and lotteries, you win or lose according to the laws of probability.

Descriptive & Inferential Statistics

➢ **Population and Sample:**

- A **population** consists of all subjects (human or otherwise) that are being studied.

- Most of the time, due to the expense, time, size of population, medical concerns, etc., it is not possible to use the entire population for a statistical study; therefore, researchers use samples.

- A **sample** is a group of subjects selected from a population.

Machine Learning

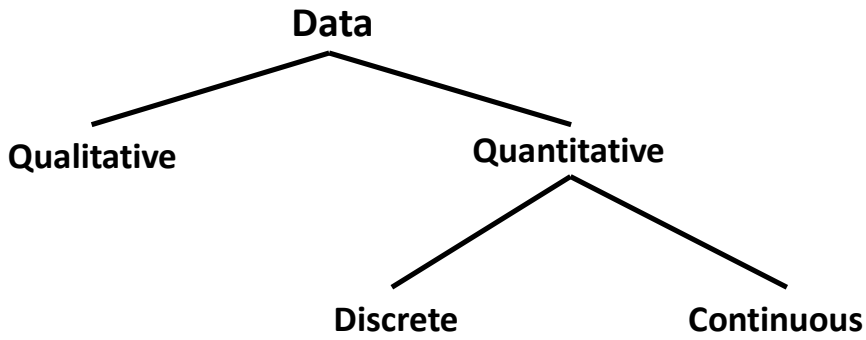
Variables & Types of Data

Variables & Types of Data

- Variables can be classified as qualitative or quantitative.
- **Qualitative variables** are variables that can be placed into distinct categories, according to some characteristic or attribute. For example, if subjects are classified according to gender (male or female), then the variable *gender* is qualitative.
- Other examples of qualitative variables are religious preference and geographic locations.

Variables & Types of Data

- **Quantitative variables** are numerical and can be ordered or ranked. For example, the variable *age* is numerical, and people can be ranked in order according to the value of their ages.
- Quantitative variables can be further classified into two groups: discrete and continuous.
- **Discrete variables** assume values that can be counted. E.g. number of children in a family.
- **Continuous variables** can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals. E.g. temperature.



Machine Learning

Measurement Scales

Measurement Scales

- **Measurement Scales:**
- They are of four types:

- The **nominal level of measurement** classifies data into mutually exclusive (non overlapping), exhausting categories in which no order or ranking can be imposed on the data. As example political party (Democratic, Republican, Independent, etc.), marital status (single, married, divorced, widowed, separated).

- The **ordinal level of measurement** classifies data into categories that can be ranked; however, precise differences between the ranks do not exist. As example exam grades (A, B, C, D, E, F).

Measurement Scales

- The **interval level of measurement** ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero.

- There is a meaningful difference of 1 point between an IQ of 109 and an IQ of 110.

- The **ratio level of measurement** possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population.

- For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1.

Machine Learning

Data Collection and Sampling Techniques

Data Collection & Sampling Techniques

- Data can be collected in a variety of ways. One of the most common methods is through the use of surveys. Surveys can be done by using a variety of methods -
  - Telephone Surveys
  - Mailed Questionnaire Surveys
  - Personal interview surveys
  - Surveying records or direct observation of situations.

Data Collection & Sampling Techniques

- Statisticians use four basic methods of sampling: **Random**, **Systematic**, **Stratified**, and **Cluster** sampling.
- **Random samples** are selected by using chance methods or random numbers.
- Researchers obtain **Systematic samples** by numbering each subject of the population and then selecting every *k*th subject.
- Researchers obtain **Stratified samples** by dividing the population into groups (called strata) according to some characteristic that is important to the study, then sampling from each group.

Data Collection & Sampling Techniques

- Researchers also use **Cluster samples.** Here the population is divided into groups called clusters by some means such as geographic area or schools in a large school district, etc. Then the researcher randomly selects some of these clusters and uses all members of the selected clusters as the subjects of the samples.

- **Convenience sample.** Here a researcher uses subjects that are convenient. For example, the researcher may interview subjects entering a local mall to determine the nature of their visit or perhaps what stores they will be visiting.

Machine Learning

Observational and Experimental Studies

Observational & Experimental Studies

- In an **observational study,** the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.
- In an **experimental study,** the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

Observational & Experimental Studies

- The **independent variable** in an experimental study is the one that is being manipulated by the researcher. The independent variable is also called the **explanatory variable.**

- The resultant variable is called the **dependent variable** or the **outcome variable.** The group that received the special instruction is called the **treatment group** while the other is called the **control group.** The treatment group receives a specific treatment (in this case, instructions for improvement) while the control group does not.

Machine Learning

Exercise on Sampling Techniques

Exercise on Sampling Techniques

- **Classify each sample as random, systematic, stratified, or cluster.**

1. In a large school district, all teachers from two buildings are interviewed to determine whether they believe the students have less homework to do now than in previous years.

2. Every seventh customer entering a shopping mall is asked to select her or his favourite store.

3. Nursing supervisors are selected using random numbers to determine annual salaries.

Exercise on Sampling Techniques

- **Classify each sample as random, systematic, stratified, or cluster. (Contd.)**

4. Every 100th hamburger manufacturer is checked to determine its fat content.

5. Mail carriers of a large city are divided into four groups according to gender (male or female) and according to whether they walk or ride on their routes. Then 10 are selected from each group and interviewed to determine whether they have been bitten by a dog in the last year.