Probability Theory & Statistics

Jerry Peng

1 Probability Theory

1.1 Classical Probability

Classical probability is a concept of probability that assumes all outcomes in a sample space are equally likely. This approach is often used in situations where we deal with simple random processes, such as tossing a fair coin or rolling a fair die.

To explain classical probability using mathematics, we consider two key concepts:

- **Sample Space (S):** This is the set of all possible outcomes of a probability experiment.
- **Event (E):** An event is any subset of the sample space. It can include one or more outcomes.

The formula for classical probability is given by:

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of equally likely outcomes}}$$

This means the probability of an event is the ratio of the number of ways that event can occur to the total number of equally likely outcomes in the sample space.

1.2 Bayes' Theorem

Bayesian probability represents a different approach to understanding probability, one that is deeply rooted in the interpretation of probability as a measure of belief or certainty. In contrast to frequentist probability, which interprets probability through the frequency of events, Bayesian probability is subjective and depends on the prior knowledge or beliefs about the event in question.

Mathematically, it's expressed as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In classification problems, the composition of the aforementioned definition is as follows:

- $P(A|B)$ is posterior probability: It represents the probability of event A occurring given the feature B. This is the updated probability of the event after taking new evidence into account. It's calculated using Bayes' Theorem and combines the prior probability and the likelihood of the new evidence.
- $P(B|A)$ is likelihood: It represents the probability of feature B occurring given the category A. This is the probability of observing the given data under different hypotheses. In Bayesian analysis, we consider how likely the observed data is under different scenarios.
- $P(A)$ is the prior probability, representing the probability of category A occurring. This is the initial judgment or belief about the probability of an event, before considering new evidence.
- $P(B)$ is the evidence or marginal probability, representing the probability of feature B occurring. It can be calculated as

  $$P(B) = P(B|A) \times P(A) + P(B|\neg A) \times P(\neg A)$$

1.3 Conditional Probability

Conditional probability is a fundamental concept in probability theory that deals with the probability of an event occurring given that another event has already occurred. This concept is crucial for understanding the relationship between two events, particularly in scenarios where the occurrence of one event affects the likelihood of the occurrence of another.

The conditional probability of an event A given that B has occurred is denoted as $P(A|B)$, and is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

2 Descriptive Statistics

2.1 Basic Terms

Mean The "average" or "expected value" of the sample

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_1 + x_2 + \cdots + x_N)$$

Median The Median of a distribution is the value 'in the middle'

Mode The mode is the value that is "most common" in the distribution

Variance Variance is a statistical measure that represents the degree of spread or dispersion in a set of data points. It is calculated by

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

A high variance indicates that the data points are spread out widely around the mean, showing a high level of variability in the dataset. A low variance indicates that the data points are closer to the mean, showing less variability.

Standard Deviation Standard deviation indicates how much the values in a dataset deviate, on average, from the mean of the dataset. It is calculated by

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

A high standard deviation indicates that the data points are spread out over a wider range of values, implying greater variability in the dataset. A low standard deviation suggests that the data points are clustered closely around the mean, implying less variability.

2.2 Distributions

A probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. It's a fundamental concept in probability theory and statistics, describing how probabilities are distributed over the values of a random variable.

There are two main types of probability distributions, depending on whether the random variable is discrete or continuous:

1. **Discrete Probability Distributions:** These apply to scenarios where the set of possible outcomes is discrete (e.g., a countable set like the number of heads in coin tosses).

   - **Binomial Distribution:** Used for a fixed number of trials, each with the same probability of success. For example, the number of heads in a series of coin flips.

- **Poisson Distribution:** Describes the probability of a given number of events happening in a fixed interval of time or space. It's often used for counting events like the number of emails received in an hour.
- **Geometric Distribution:** Models the number of trials until the first success in a series of Bernoulli trials.

2. **Continuous Probability Distributions:** These apply when the set of possible outcomes can take on values in a continuous range.

- **Normal (Gaussian) Distribution:** Describes a continuous variable whose distribution is symmetric and bell-shaped. It's characterized by its mean and standard deviation.
- **Exponential Distribution:** Used for modeling the time between events in a process that occurs continuously and independently at a constant average rate.
- **Uniform Distribution:** All outcomes in a range are equally likely.

3 Inferential Statistics

3.1 Hypothesis Test

A hypothesis test, also known as a significance test, is a statistical method used to make inferences or decisions about population parameters based on a sample. It involves comparing observed data to what we would expect under certain statistical assumptions to determine whether any observed effects are statistically significant or if they could have occurred by random chance.

- **Null Hypothesis (H_0):** This is a statement of no effect or no difference. It represents the status quo or a baseline assumption that you wish to test against.
- **Alternative Hypothesis (H_α or H_1):** This is what you want to prove or verify. It's a statement indicating the presence of an effect or difference.
- **P-value:** The P-value is a probability that measures the evidence against the null hypothesis. It tells you how likely you would observe your data (or something more extreme) if the null hypothesis were true. A small P-value indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- **Decisions:** Based on the P-value and a predetermined significance level α, you make a decision:
  - If P-value $\leq \alpha$, reject the null hypothesis in favor of the alternative.
  - If P-value $> \alpha$, fail to reject the null hypothesis.

It's crucial to understand that a hypothesis test can have errors:

- **Type I Error:** Rejecting H_0 when it's true (False Positive).
- **Type II Error:** Failing to reject H_0 when H_α is true (False Negative).

3.2 Confidence Interval

A confidence interval (CI) is a type of interval estimate, used in statistics, that is computed from the observed data. It provides a range of values that is likely to contain the value of an unknown population parameter. The confidence interval offers an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.

The confidence interval is calculated as

$$CI = \bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$$

where

- $\bar{x}$ is the sample mean.

- z is the z-score corresponding to the chosen confidence level (e.g., 1.96 for 95% confidence).

- σ is the standard deviation of the population (if known) or the sample standard deviation (if the population standard deviation is unknown).

- n is the sample size

A confidence interval provides a range of values for the estimate, acknowledging that there is a certain level of uncertainty in any estimate from sample data. The wider the interval, the more uncertainty there is in the estimate. The confidence level chosen (90%, 95%, 99%, etc.) reflects how sure you want to be about the interval containing the true parameter; higher confidence levels result in wider intervals.

3.3 Correlation Functions

Correlation functions are statistical measures that describe the extent to which two variables are linearly related to each other. In statistics and data analysis, understanding the relationship between variables is essential, and correlation functions provide a quantifiable way to measure this relationship.

3.3.1 Pearson Correlation Coefficient

The Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y. It has a value between -1 and 1, where:

- 1 is total positive linear correlation

- 0 is no linear correlation

- -1 is total negative linear correlation

The formula for the Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where X_i and Y_i are individual sample points, while $\bar{X}$ and $\bar{Y}$ are the mean values of the variables X and Y, respectively.

3.3.2 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient is a non-parametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function. It's particularly useful when the data do not meet the assumptions of the Pearson correlation (e.g., non-linear relationships, non-normally distributed variables). The formula for Spearman's rank correlation is similar to Pearson's but applied to rank values:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of corresponding variables, and n is the number of observations..

References