**Statistics For Machine Learning**

**Machine Learning Statistics:** In the field of machine learning (ML), statistics plays a pivotal role in extracting meaningful insights from data to make informed decisions. Statistics provides the foundation upon which various ML algorithms are built, enabling the analysis, interpretation, and prediction of complex patterns within datasets.

This article delves into the significance of statistics in machine learning and explores its applications across different domains.

**What is Statistics?**

[Statistics](https://www.geeksforgeeks.org/statistics/) is the science of collecting, organizing, analyzing, interpreting, and presenting data. It encompasses a wide range of techniques for summarizing data, making inferences, and drawing conclusions.

Statistical methods help quantify uncertainty and variability in data, allowing researchers and analysts to make data-driven decisions with confidence.

**What is Machine Learning?**

[Machine learning](https://www.geeksforgeeks.org/machine-learning/) is a branch of artificial intelligence (AI) that focuses on developing algorithms and models capable of learning from data without being explicitly programmed.

ML algorithms learn patterns and relationships from data, which they use to make predictions or decisions. Machine learning encompasses various techniques, including supervised learning, unsupervised learning, and reinforcement learning.

**Applications of Statistics in Machine Learning**

Statistics is a key component of machine learning, with broad applicability in various fields.

* Feature engineering relies heavily on statistics to convert geometric features into meaningful predictors for[machine learning algorithms.](https://www.geeksforgeeks.org/machine-learning-algorithms/)
* In image processing tasks like object recognition and segmentation, statistics accurately reflect the shape and structure of objects in images.
* Anomaly detection and quality control benefit from statistics by identifying deviations from norms, aiding in the detection of defects in manufacturing processes.
* Environmental observation and geospatial mapping leverage statistical analysis to monitor land cover patterns and ecological trends effectively.

Overall, statistics plays a crucial role in machine learning, driving insights and advancements across diverse industries and applications.

**Types of Statistics**

There are commonly two types of statistics, which are discussed below:

* **Descriptive Statistics:** “[De­scriptive Statistics](https://www.geeksforgeeks.org/descriptive-statistic/)” helps us simplify and organize big chunks of data. This makes large amounts of data easier to understand.
* **Inferential Statistics:** “[Inferential Statistics](https://www.geeksforgeeks.org/difference-between-descriptive-and-inferential-statistics/)” is a little different. It uses smaller data to draw conclusions about a larger group. It helps us predict and draw conclusions about a population.

**Descriptive Statistics**

Descriptive statistics summarize and describe the features of a dataset, providing a foundation for further statistical analysis.

| **Mean** | **Median** | **Mode** |
| --- | --- | --- |
| [Mean](https://www.geeksforgeeks.org/mean/) is calculated by summing all values present in the sample divided by total number of values present in the sample. | [Median](https://www.geeksforgeeks.org/median/) is the middle of a sample when arranged from lowest to highest or highest to lowest. in order to find the median, the data must be sorted.  For odd number of data points:  For even number of data points: | [Mode](https://www.geeksforgeeks.org/mode/)is the most frequently occurring value in the dataset. |

**Measures of Dispersion**

* **Range:** The difference between the maximum and minimum values.
* [**Variance**](https://www.geeksforgeeks.org/variance/)**:** The average squared deviation from the mean, representing data spread.
* [**Standard Deviation**](https://www.geeksforgeeks.org/variance-and-standard-deviation/)**:** The square root of variance, indicating data spread relative to the mean.
* [**Interquartile Range**](https://www.geeksforgeeks.org/interquartile-range/)**:** The range between the first and third quartiles, measuring data spread around the median.

**Measures of Shape**

* [**Skewness**](https://www.geeksforgeeks.org/skewness-measures-and-interpretation/)**:** Indicates data asymmetry.

A diagram of a function

Description automatically generated

*Types of Skewed data*

* [**Kurtosis**](https://www.geeksforgeeks.org/difference-between-skewness-and-kurtosis/)**:** Measures the peakedness of the data distribution.

A diagram of distribution of a distribution

Description automatically generated

*Types of Skewed data*

**Covariance and Correlation**

| **Covariance** | **Correlation** |
| --- | --- |
| [Covariance](https://www.geeksforgeeks.org/covariance-and-correlation-in-r-programming/) measures the degree to which two variables change together. | [Correlation](https://www.geeksforgeeks.org/how-to-find-correlation-coefficient-in-excel/) measures the strength and direction of the linear relationship between two variables. It is represented by correlation coefficient which ranges from -1 to 1. A positive correlation indicates a direct relationship, while a negative correlation implies an inverse relationship.  Pearson’s correlation coefficient is given by: |

**Visualization Techniques**

* **Histograms:** Show data distribution.
* **Box Plots:** Highlight data spread and potential outliers.
* **Scatter Plots:** Illustrate relationships between variables.

**Probability Theory**

Probability theory forms the backbone of [statistical inference](https://www.geeksforgeeks.org/statistical-inference/#:~:text=Statistical%20inference%20is%20the%20process,on%20data%20from%20a%20sample.), aiding in quantifying uncertainty and making predictions based on data.

**Basic Concepts**

* [**Random Variables**](https://www.geeksforgeeks.org/random-variable/)**:** Variables with random outcomes.
* [**Probability Distributions**](https://www.geeksforgeeks.org/probability-distribution/)**:** Describe the likelihood of different outcomes.

**Common Probability Distributions**

* [**Binomial Distribution**](https://www.geeksforgeeks.org/what-is-binomial-probability-distribution-with-example/)**:** Represents the number of successes in a fixed number of trials.
* [**Poisson Distribution**](https://www.geeksforgeeks.org/poisson-distribution/)**:** Describes the number of events occurring within a fixed interval.
* [**Normal Distribution**](https://www.geeksforgeeks.org/normal-distribution/)**:** Characterizes continuous data symmetrically distributed around the mean.

[**Law of Large Numbers**:](https://www.geeksforgeeks.org/law-of-large-numbers/)

States that as the sample size increases, the sample mean approaches the population mean.

[**Central Limit Theorem**:](https://www.geeksforgeeks.org/central-limit-theorem-in-machine-learning/)

Indicates that the distribution of sample means approximates a normal distribution as the sample size grows, regardless of the population’s distribution.

**Inferential Statistics**

Inferential statistics involve making predictions or inferences about a population based on a sample of data.

**Population and Sample**

* [**Population**](https://www.geeksforgeeks.org/population-and-sample-statistic/)**:** The entire group being studied.
* **Sample:** A subset of the population used for analysis.

**Estimation**

* [**Point Estimation**](https://www.geeksforgeeks.org/point-estimation/)**:** Provides a single value estimate of a population parameter.
* [**Interval Estimation**](https://www.geeksforgeeks.org/confidence-interval/)**:** Offers a range of values (confidence interval) within which the parameter likely lies.
* [**Confidence Intervals**](https://www.geeksforgeeks.org/confidence-interval/)**:** Indicate the reliability of an estimate.

[**Hypothesis Testing**](https://www.geeksforgeeks.org/understanding-hypothesis-testing/)

* **Null and Alternative Hypotheses:** The null hypothesis assumes no effect or relationship, while the alternative suggests otherwise.
* **Type I and Type II Errors:** Type I error is rejecting a true null hypothesis, while Type II is failing to reject a false null hypothesis.
* **p-Values:** Measure the probability of obtaining the observed results under the null hypothesis.
* **t-Tests and z-Tests:** Compare means to assess statistical significance.

[**ANOVA (Analysis of Variance)**:](https://www.geeksforgeeks.org/anova-formula/)

Compares means across multiple groups to determine if they differ significantly.

[**Chi-Square Tests**:](https://www.geeksforgeeks.org/chi-square-test/)

Assess the association between categorical variables.

**Correlation and Regression:**

Understanding relationships between variables is critical in machine learning.

**Correlation**

* [**Pearson Correlation Coefficient**](https://www.geeksforgeeks.org/pearson-correlation-coefficient/)**:** Measures linear relationship strength between two variables.
* [**Spearman Rank Correlation**](https://www.geeksforgeeks.org/spearmans-rank-correlation/)**:** Assesses the strength and direction of the monotonic relationship between variables.

**Regression Analysis**

* [**Simple Linear Regression**](https://www.geeksforgeeks.org/ml-linear-regression/)**:** Models the relationship between two variables.
* [**Multiple Linear Regression**](https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/)**:** Extends to multiple predictors.
* [**Assumptions of Linear Regression**](https://www.geeksforgeeks.org/assumptions-of-linear-regression/)**:** Linearity, independence, homoscedasticity, normality.
* [**Interpretation of Regression Coefficients**](https://www.geeksforgeeks.org/regression-coefficients/)**:** Explains predictor influence on the response variable.
* [**Model Evaluation Metrics**](https://www.geeksforgeeks.org/metrics-for-machine-learning-model/)**:** R-squared, Adjusted R-squared, RMSE.

**Bayesian Statistics**

Bayesian statistics incorporate prior knowledge with current evidence to update beliefs.

[Bayes’ Theorem](https://www.geeksforgeeks.org/bayes-theorem/) is a fundamental concept in probability theory that relates conditional probabilities. It is named after the Reverend Thomas Bayes, who first introduced the theorem. Bayes’ Theorem is a mathematical formula that provides a way to update probabilities based on new evidence. The formula is as follows:

*P*(*A*∣*B*)=*P*(*B*)*P*(*B*∣*A*)⋅*P*(*A*)​, where

* *P*(*A*∣*B*): The probability of event A given that event B has occurred (posterior probability).
* *P*(*B*∣*A*): The probability of event B given that event A has occurred (likelihood).
* *P*(*A*): The probability of event A occurring (prior probability).
* *P*(*B*): The probability of event B occurring.