Statistics & Probability for Machine Learning

Statistics and Probability form the mathematical foundation for many machine learning algorithms. They help describe, understand, and make predictions from data.

---

* Statistics in Machine Learning

Statistics is the study of data collecting, summarizing, and interpreting data.

-> Descriptive Statistics:

- Mean: Average value
- Median: Middle value
- Mode: Most frequent value
- Standard Deviation (SD): Measure of how spread out values are
- Variance: Square of SD

-> Inferential Statistics:

Used to make predictions or inferences about a population based on a sample.

Examples in ML:

- Mean squared error (MSE)
- Confidence intervals in A/B testing
- Statistical hypothesis testing (p-value, t-test)

-> Distribution:

- Normal Distribution: Bell-curve, used in Gaussian Naive Bayes, regression

- Skewness & Kurtosis: Used for understanding feature behavior

---

* Probability in Machine Learning

Probability measures how likely an event is to occur.

-> Key Concepts:

- Random Variables

- Probability Distribution: e.g., Bernoulli, Binomial, Normal

- Conditional Probability: P(A|B)

- Bayes Theorem: Core of Naive Bayes classifier

-> Use Cases in ML:

- Naive Bayes classification

- Predictive modeling

- Bayesian inference

- Hidden Markov Models (HMMs)

Bayes Theorem:

$P(A|B) = [P(B|A) * P(A)] / P(B)$

---

* Applications in ML Algorithms

- Naive Bayes: Based on probability and Bayes theorem

- Logistic Regression: Uses Bernoulli probability distribution

- Clustering: Often assumes data distributions (e.g., Gaussian Mixture Models)

- Random Forest: Uses statistical measures like Gini Impurity and Entropy

---

* Summary

Understanding statistics and probability allows you to:

- Choose the right model

- Evaluate results with confidence

- Interpret outputs of algorithms