UmdeMi

Summary of Statistics in ML

Overview

In this file, I will summarize some important topics on probability and statistics.

And know Why Statistics is given so much importance for excelling into Machine learning and artificial intelligence?

Goals

1. Know Why Learn Statistics and what it is?
2. Know some important terminology of statistics and probability
3. Know why statistics is important for mastering AI/ML skills

Milestones

Basic Probability Theory and Statistics

- **Random Experiment:-**

  is a physical situation whose outcome cannot be predicted until it is observed

- Sample Space:-

is a set of all possible outcomes of a random experiment.

- Random variable:-

- is a variable whose possible values are numerical outcomes.

- **Random variables types:-**

1. D*iscrete Random Variable*
2. Continuous Random Variable

- Probability:-

- is the measure of the likelihood that an event will occur in a Random Experiment.

- Takes values between 0 and 1.

- Conditional Probability:-

- is a measure of the probability of an event given that another event has already occurred.

- is usually written as P(A|B)

- Independence:-

- Two events are independent of each other, if the probability that one event occurs in no way affects the probability of the other event occurring.

- **P(A, B) = P(A) * P(B) where P(A) =! 0 and P(B) =! 0**

- Conditionally independence:-

- Two events A and B are conditionally independent given a third event C precisely if the occurrence of A and the occurrence of B are independent events in their conditional probability distribution given C

- **P(A | C, B | C) = P(A | C) * P(B | C)**

- Expectation:-

expectation **E(X)** is what you would expect the outcome of an experiment to be average if you repeat the experiment a large number of time.

- Variance:-

- is a measure of how concentrated the distribution of a random variable X is around its mean.

- The average of the squared differences from the Mean.

- Probability Distribution:-

- Is a mathematical function that maps the all possible outcomes of a random experiment with its associated probability.

- **Probability Distribution types:-**

1. Discrete Probability Distribution
2. Continuous Probability Distribution

- Joint Probability Distribution:-

Joint distribution function of X and Y is:-

F xy(X, Y) = P(X = x, Y = y)

- Conditional Probability Distribution CPD:-

- If Z is random variable who is dependent on other variables X and Y, then the distribution of P(Z|X,Y) is called CPD of Z

- **Some of the important operations**

  - Conditioning/Reduction:-
  - Marginalisation

What is Statistics and why is it important in machine learning?

- Why Learn Statistics?

Statistics and Statistical methods are required to find answers to the questions that we have about data.

Statistics allows us to understand the data used to train a ML model and to interpret the results of testing different ML models

- What is Statistics?

Is a collection of methods for working with data and using data to answer questions.

It can be both a classical method from statistics and a modern algorithm used for feature selection or modeling.

Statistics is divided into two large groups of methods:

- **descriptive statistics** for summarizing data
- **inferential statistics** for drawing conclusions from samples of data.

- Descriptive Statistics

refer to methods for summarizing raw observations into information that we can understand and share.

Common methods as :-

- Mean
- Median
- Variance
- standard deviation
- Graphical methods as Charts

- Inferential Statistics

methods that aid in quantifying properties of the domain or population from a smaller set.

Common methods as :-

- Expected value
- The amount of spread
- Statistical hypothesis testing

Why Statistics Is Important For Mastering AI/ML Skills

- Why statistics?

Many ML techniques and algorithms are either fully borrowed from or heavily rely on the theory from statistics.

Both subjects have their own applications when it comes to AI and Ml, for instance, continuity and differentiability in maths are widely used in AI/ML algorithms.

It is important to learn the subject to interpret the results of logistic regression or you will end up being baffled by how bad your models perform due to non-normalised predictors.

- Finding The Right Course

The demand for the subjects mostly comes from Economics students and is widely seen as a subset of it as well.

Why is Statistics given so much importance for excelling into Machine learning / artificial intelligence?

- **Statistics** is basically math used in a way to give a different type of context of all the historical data.
- **Statistics** deriving correlations between features/columns tells you whether they have a relationship.
- Non-linear models like **neural networks** that might not be as important because of the more sophisticated math.

- Statistics are important because:-
  - It gives us an entire overview of the dataset.
  - It allows us to use the right type of algorithm according to data.
  - They are able to solve much more complex problems.
- It is important to learn the subject to interpret the results of logistic regression.
- **Programming** is basically two things combined together. **Maths** and **logic**
- In AI or ML the decisions are made based on a logical intelligence given to it.
- Statistics is the background mechanics that work behind machine learning and AI