ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DAY – 9

Date: Jul 03, 2025

Mean, Median, and Mode are statistical measures used to describe the central tendency of a dataset. In machine learning, these measures are used to understand the distribution of data and identify outliers. Here, we will explore the concepts of Mean, Median, and Mode and their implementation in Python.

Mean

The "mean" is the average value of a dataset. It is calculated by adding up all the values in the dataset and dividing by the number of observations. The mean is a useful measure of central tendency because it is sensitive to outliers, meaning that extreme values can significantly affect the value of the mean.

In Python, we can calculate the mean using the NumPy library, which provides a function called mean().

Median

The "median" is the middle value in a dataset. It is calculated by arranging the values in the dataset in order and finding the value that lies in the middle. If there are an even number of values in the dataset, the median is the average of the two middle values.

The median is a useful measure of central tendency because it is not affected by outliers, meaning that extreme values do not significantly affect the value of the median.

In Python, we can calculate the median using the NumPy library, which provides a function called median().

Mode

The "mode" is the most common value in a dataset. It is calculated by finding the value that occurs most frequently in the dataset. If there are multiple values that occur with the same frequency, the dataset is said to be bimodal, trimodal, or multimodal.

The mode is a useful measure of central tendency because it can identify the most common value in a dataset. However, it is not a good measure of central tendency for datasets with a wide range of values or datasets with no repeating values.

In Python, we can calculate the mode using the SciPy library, which provides a function called mode().

Standard Deviation

Standard deviation is a measure of the amount of variation or dispersion of a set of data values around their mean. In machine learning, it is an important statistical concept that is used to describe the spread or distribution of a dataset.

Standard deviation is calculated as the square root of the variance, which is the average of the squared differences from the mean. The formula for calculating standard deviation is as follows –

$\sigma = \sqrt{[\Sigma(x-\mu)2/N]}$

```
step@step-HP-ProDesk-400-G5-SFF: ~
import statistics

data=[23,45,45,45,56,56,78,89,95,95]
mean=statistics.mean(data)
median=statistics.median(data)
mode=statistics.mode(data)
sd=statistics.stdev(data)

print("Data : ",data)
print("Mean : ",mean)
print("Median : ",median)
print("Mode : ",mode)
print("Standard Deviation : ",sd)
~
~
~
~
```

OUTPUT:

```
step@step-HP-ProDesk-400-G5-SFF: ~
step@step-HP-ProDesk-400-G5-SFF:~$ vim maths.py
step@step-HP-ProDesk-400-G5-SFF:~$ python3 maths.py
Data :  [23, 45, 45, 45, 56, 56, 78, 89, 95, 95]
Mean :  62.7
Median :  56.0
Mode :  45
Standard Deviation :  24.984661961558203
step@step-HP-ProDesk-400-G5-SFF:~$
```

Percentiles

Percentiles are a statistical concept used in machine learning to describe the distribution of a dataset. A percentile is a measure that indicates the value below which a given percentage of observations in a group of observations falls.

For example, the 25th percentile (also known as the first quartile) is the value below which 25% of the observations in the dataset fall, while the 75th percentile (also known as the third quartile) is the value below which 75% of the observations in the dataset fall.

Percentiles can be used to summarize the distribution of a dataset and identify outliers. In machine learning, percentiles are often used in data preprocessing and exploratory data analysis to gain insights into the data.

Python provides several libraries for calculating percentiles, including NumPy and Pandas.

```
>>> import numpy as np
... data = np.array([1, 2, 3, 4, 5])
... p25 = np.percentile(data, 25)
... p75 = np.percentile(data, 75)
... print('25th percentile:', p25)
... print('75th percentile:', p75)
...
25th percentile: 2.0
75th percentile: 4.0
>>>
>>>
```

Matplotlib

Matplotlib is a low level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter. Matplotlib is open source and we can use it freely. Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility.

Scatter Plot

With Pyplot, you can use the scatter() function to draw a scatter plot.

The scatter() function plots one dot for each observation. It needs two arrays of the same length, one for the values of the x-axis, and one for values on the y-axis:

```
step@step-HP-ProDesk-400-G5-SFF: ~
import matplotlib.pyplot as plt

# Sample data
x = [1, 3, 2, 6, 5]
y = [2, 4, 1, 3, 7]

# Create scatter plot
plt.scatter(x, y, color='red', marker='o')

# Add labels and title
plt.xlabel("X-axis")
plt.ylabel("Y-axis")
plt.title("Scatter Plot")

# Show grid and plot
plt.grid(True)
plt.show()
```

Output: