2024

Skewness,variance,deviation

IBTIHAL HOMADI
20-16-0099 CS4

In data mining, these terms are all related to describing the spread and distribution of data within a dataset:

Skewness:

- **Meaning:** Skewness refers to the asymmetry of a distribution relative to a normal (Gaussian) distribution. It indicates whether the data is "tilted" to one side or the other.
- **Values:**
    - Positive skewness: The "tail" of the distribution extends more towards the right, indicating most values are concentrated on the left side.
    - Negative skewness: The "tail" extends more towards the left, indicating most values are concentrated on the right side.
    - Zero skewness: The distribution is symmetrical, resembling a normal bell curve.
- **Importance:** Knowing skewness is crucial for choosing appropriate statistical methods, as some assume normalcy. It can also indicate underlying causes in real-world data (e.g., income distribution might be skewed positively).

Variance:

- **Meaning:** Variance measures how spread out the data is from its mean (average). A high variance indicates that data points are far from the mean, while a low variance suggests they are clustered closely around the mean.
- **Calculation:** It's the average squared deviation from the mean.
- **Interpretation:** Variance helps understand data variability and can be used for tasks like comparing different groups or features.

Standard Deviation:

- **Meaning:** Standard deviation is the square root of variance. It's also a measure of spread but expressed in the same units as the original data, making it easier to interpret.
- **Interpretation:** Like variance, it indicates how much data deviates from the mean. A high standard deviation means the data is more spread out, while a low standard deviation suggests it's more concentrated.

Relationship:

- Variance and standard deviation are directly related by the square root operation.
- Skewness describes the shape of the distribution, while variance and standard deviation measure the spread.

Together, these terms help you:

- Understand the distribution of data in your dataset.
- Choose appropriate statistical methods for analysis.
- Identify potential biases or patterns in your data.