Types of Data

Data can be categorized based on its nature and measurement levels:

1) Based on Measurement Scales:

- **Nominal Data**:- Categorical data with no inherent order (e.g colors, gender)
- **Ordinal Data**:- Categorical data with a meaningful order but no fixed intervals (e.g rankings)
- **Interval Data**:- It is a type of numerical data where the difference between values is meaningful and consistent, but there is no true zero point.
  Example:- Temperature in Celsius or Farenheit, calender years, and IQ scores.
- **Ratio Data**:- It is the numerical data that has meaningful intervals and a true zero point, allowing for calculation of ratios.
  Examples:- Weight, height, age, income, distance

2) Based on Structure

- Univariate Data:- Focuses on analysing one variable at a time. It is the data with a single variable (e.g height of students)
  Example:- Analyzing the distribution of students heights using histograms or boxplots

- Multivariate Data :- It is the data with multiple variables (e.g, weight, height and age of students) It involves analyzing relationships between two or more variables.

Example :- Studying the correlation between height, weight and age of students.

Key Techniques

Univariate Analysis: Measures of central tendency (mean, median) and dispersion (variance, standard deviation)

Multivariate Analysis :- Scatterplots, covariance, correlation, regression and dimensionality reduction techniques like PCA

Covariance in Machine Learning

Covariance plays a fundamental role in understanding relationships between variables and optimizing machine learning models.

Covariance quantifies the directional relationship between two variables. It indicates whether changes in one variable are associated with changes in another

Covariance measures how much two random variables change together.

For two variables X and Y, covariance is calculated as

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

where

$x_i, y_i \rightarrow$ individual data points of X and Y respectively

$\bar{x}, \bar{y} \rightarrow$ means of X and Y

$n \rightarrow$ no. of data points

1) Positive Covariance:
If $Cov(X,Y) > 0$, it means that as X increases, Y tends to increase, and vice versa.
Example:- Hours studied and test scores may have positive covariance.

2) Negative Covariance:-
If $Cov(X,Y) < 0$, it means that as X increases, Y tends to decrease and vice versa
Example:- Speed of a car and travel time may have negative covariance.

3) Zero Covariance:-
If $Cov(X,Y) = 0$, it means there is no linear relationship between X and Y.

Importance in Machine Learning

1) Feature Selection:- Covariance helps to identify relationships between features. Features with high covariance may carry redundant information, which can negatively affect models.
Solution:- Use dimensionality reduction techniques like PCA which rely on the covariance matrix

Principal Component Analysis (PCA) :- PCA transforms data into a new co-ordinate system using the eigen vectors of the covariance matrix. These eigen vectors represent orthogonal directions (principal components) with the largest variance captured in the first components.

3) Model Interpretation :- Understanding covariance helps in interpreting multivariate relationships and understanding feature dependencies in models.

4) Portfolio Optimization in Finance :- Machine learning models in finance use covariance to diversify assets and minimize risk.

Correlation

Correlation is a statistical measure that expresses the strength and direction of the linear relation -ship between two variables. It quantifies how changes in one variable are associated with changes in another.

Types of Correlation

1) Positive Correlation :- If one variable increases, the other also increases.
Example :- Height and weight
Correlation coefficient (r) > 0

2) Negative Correlation:- If one variable increases, the other decreases.
Example:- Speed of a car and travel time.
Correlation coefficient (r) < 0

3) No Correlation:- If no relationship exists between the two variables. Example:- Shoe size and IQ.
Correlation coefficient (r) ≈ 0

$$\text{Correlation coefficient}(r) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

where $Cov(X,Y)$ is the covariance between X and Y.
σ_X, σ_Y are the standard deviations of X and Y.

Properties of r:-

- $-1 \leq r \leq 1$
- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

Importance of Correlation in Machine Learning

1) Feature selection:-
- Correlation helps identify the relationship between features and the target variable.
- Highly correlated features with target variable are more likely to contribute to model performance.

Example:- In predicting house prices features like "area" and "number of rooms" may have high positive correlation with the target price.

Removing Redundant Features :- Features with high correlation to each other (multicollinearity) may not add new information to the model. By removing such redundant features, the model can be simplified without sacrificing accuracy.

3) Evaluating Relationships :- It helps understand the strength of relationships between independent and the dependent variables.
Example :- Correlation can confirm if increasing advertising speed improve sales.

4) Feature Engineering :- Correlation insights guide the creation of new feature. Example :- Combining two weakly correlated variables may lead to better predictors.

5) Improving Model Interpretability :- Correlation analysis helps explain how features affect predictions, improving transparency in machine learning ~~models~~ procers.

6) Dimensionality Reduction :- Techniques like Principal Component Analysis (PCA) rely on correlation to reduce features while retaining the most important information.

Covariance vs. Correlation

1) Covariance measures the direction of a linear relationship but not its strength.

- Correlation measures the strength and direction of the linear relationship between two variables.

2) Covariance is scale-dependent (units depend on the variables being measured) ~~while correlation is dimensionless (scale-independent).~~
Correlation is dimensionless (scale-independent).
Values are standarized between -1 and 1.

3) Covariance has no fixed range, can take any value between $-\infty$ to $+\infty$

In Correlation, the correlation coefficient (r) varies between -1 to $+1$

$$-1 \leq r \leq 1.$$

Orthogonality in Machine Learning

Orthogonality is a concept from linear algebra that refers to vectors being perpendicular to each other in a multidimensional space.
In machine learning, orthogonality has important applications in designing models, optimizing performance and interpreting data.
Two vectors are orthogonal if their dot product is zero, i.e

$$u \cdot v = 0$$

This means there is no overlap or interaction between two vectors. Orthogonality implies that vectors are independent of each other.

Importance of Orthogonality in Machine Learning

1) **Feature Independence**:- In machine learning, orthogonality ensures that features (variables) are independent of each other. This reduces redundancy and helps models better understand the data.
Example:- Orthogonal features allow models to avoid multicollinearity, improving interpretability and stability.

2) **Orthogonal Basis in Dimensionality Reduction**:- Techniques like Principal Component Analysis (PCA) create orthogonal components (principal axes) that represent uncorrelated features. This simplifies data representation while preserving maximum variance, improving computational efficiency.

3) **Gradient Descent and Optimization**:- In optimization, orthogonal directions are preferred because they allow for independent updates to model parameters. Orthogonality ensures that adjustments to one parameter do not interfere with others, leading to faster convergence during training.

Regularization:- Orthogonal regularization encourages learned features or weights in a neural network to be orthogonal. This prevents overfitting by ensuring diverse and non-redundant feature extraction.

5) Error Decomposition in Models:- Orthogonality simplifies the decomposition of errors into independent components. For instance, in regression, residuals (errors) are orthogonal to the fitted values. This allows for clearer diagnostics and better understanding of the model's performance.

6) Model Interpretability:- When weights or features are orthogonal, it becomes easier to interpret the contributions of individual components in the predictions.

Orthogonality implies zero correlation between features or vectors. However zero correlation does not always mean orthogonality unless the relationship is linear.

Example, in PCA, orthogonal components are uncorrelated are linearly independent

Mathematical Representation

Let X be a matrix of features in n-dimensional space. PCA ensures:

- Orthogonality of Principal Components:
  If u_1 and u_2 are principal components,

$$u_1 \cdot u_2 = 0$$

This ensures there is no redundancy in the reduced feature space.