Exploratory Data Analysis (EDA)

Theory and Math Behind Key Concepts

Outline

Missing Values

Outliers

Irrelevant / Redundant Data

Data Type Correction

Categorical Variables

Normalization and Standardization

Summary Table

Missing Values

Missing Values

Definition: Data entries where values are not recorded.

Types:

- MCAR – Missing Completely at Random
- MAR – Missing at Random
- MNAR – Missing Not at Random

Imputation Example: Mean

$$x_{\text{imputed}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Outliers

Outliers

Definition: Observations significantly different from the rest.

Detection Methods:

- **IQR Method:**

  $$\text{IQR} = Q_3 - Q_1, \quad \text{Lower} = Q_1 - 1.5 \times \text{IQR}, \quad \text{Upper} = Q_3 + 1.5 \times \text{IQR}$$

- **Z-Score Method:**

  $$z_i = \frac{x_i - \mu}{\sigma}$$

- If $|z_i| > 3$, it's an outlier.

Irrelevant / Redundant Data

Irrelevant or Redundant Data

Definition: Features that add little or no useful information.

Methods:

- Correlation Analysis
- Variance Thresholding
- Principal Component Analysis (PCA)

Pearson Correlation:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Data Type Correction

Data Type Correction

Goal: Ensure appropriate types for modeling (e.g., integers, floats, categories).

Examples:

- Convert strings to `datetime`
- Map numerical codes to categories
- Cast floats to integers (if discrete)

Categorical Variables

Categorical Variables

Types:

- Nominal (e.g., color)
- Ordinal (e.g., education level)

Encoding Methods:

- **One-Hot Encoding:**

$$\text{Red} \rightarrow [1, 0, 0], \quad \text{Blue} \rightarrow [0, 1, 0]$$

- **Label Encoding:**

$$\text{Low} = 0, \quad \text{Medium} = 1, \quad \text{High} = 2$$

Normalization and Standardization

Normalization

Definition: Rescale features to a fixed range, usually [0, 1].

Formula:

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Use case: Distance-based models like KNN, K-Means.

Standardization

Definition: Rescale features to mean 0 and standard deviation 1.

Formula:

$$x_i' = \frac{x_i - \mu}{\sigma}$$

Use case: Models assuming normal distribution (e.g., PCA, linear regression).

Summary Table

EDA Summary Table

Step	Method	Math
Missing Values	Mean Imputation	$x' = \frac{1}{n}\sum x_i$
Outliers	IQR, Z-score	$z = \frac{x-\mu}{\sigma}$
Redundant Data	Correlation, PCA	$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$
Data Types	Type casting	N/A
Categorical	One-Hot/Label Encoding	Encoding to vectors or integers
Normalization	Min-Max Scaling	$x' = \frac{x-\min(x)}{\max(x)-\min(x)}$
Standardization	Z-score Scaling	$x' = \frac{x-\mu}{\sigma}$

Thank you!