Data Preparation Workshop

IEEE Computational Intelligence Society

Tuesday 10/1 1:30 - 3:30

Topics Discovered

Pipeline (With Coding):

1. What is Data & Data Preprocessing
2. Read the Data
3. EDA (Visualizations, Correlations, etc)
4. Skewness (How to fix it)
5. Outliers
6. Cleaning
    a. Missing Values (Mean, Median, Mode, Most Frequent)
    b. Categorical Data (One Hot Encoding, Ordinal/Label Encoding)
    c. Scaling (Min Max Scaler, Standard Scaler)
7. Feature Engineering
    a. Feature Extraction
    b. Feature Selection
    c. PCA
8. Splitting the Data (Training, Validation, and Testing) sets

Topics Discovered

Tips & Tricks (Without Coding)

1. Unbalanced Data (Categorical Target)
2. Split the data before preprocessing (if not, Data Leakage)
3. Start Simple
4. Make a Feature for null values
5. Filling the null values using ffil, bfil
6. Ordinal Encoding Sorting Issue (Alphepitically)

What is Data?

Structured Data:
- Excel Sheets
- CSV File
- Relational Databases

Unstructured Data:
- Images
- Videos
- Text
- Audio

Semi-Structured Data:
- Json Files
- XML Files

CSV FILE
(Comma-Separated Value)

Does the data always come perfect and ready to use? ***NO***

Read the Data

Using Python

EDA
(Exploratory Data Analysis)

Distribution of the Data

Skewness

Correlation

Visualization

Skewness

- Right Skewness
- Left Skewness
- Balanced (Normal Distribution)

Outliers

- What are Outliers
- What We do with them:
    - Delete Them
    - Replace Them

- Outliers not always Bad.

Cleaning & Preprocessing

- Missing Values
- Categorical Data
- Scaling

Missing Values

- Drop Row
- Drop Column

- Fill the missing values

  - Numeric
    - Mean
    - Mode
    - Median

  - Categorical
    - Most Freq.

Encode the Categorical Data

One Hot Encoding
- What is it
- When we use it
- Results are Binary
- Cons
    - Curse of Dimensionality

Ordinal Encoding
- What is it
- When we use it

Label Encoding
- Same as Ordinal Enc.
- But for the Label Feat.

Scaling

- Why Scaling? To avoid Bias

- Scaling Methods:
    - Min Max Scaler
    - Standard Scaling

Min Max Scaler

- Equation
- When to use it

Standard Scaler

- Equation
- When to use it

Feature Engineering

- What is Feature Engineering
- Types of Feature Engineering
    - Feature Extraction
    - Feature Selection
    - PCA

Feature Extraction

- What is it
- Examples
    - Total Distance
    - Seconds to Minutes

Feature Selection

- What is it
- Methods Based on
    - Logic
    - Feature Importancies
- PCA

Splitting

We split the data usually into 3 sets:

- Training Set (70%)
- Validation Set (15%)
- Testing Set (15%)

Tips & Tricks

1. Unbalanced Data (Categorical Target) → Over / Under Sampling
2. Split the data before preprocessing (if not, Data Leakage)
3. Start Simple
4. Make a Feature for null values
5. Filling the null values using ffil, bfil (forward/backward filling)
6. Ordinal Encoding Sorting Issue (Alphepitically)

Thank You

Any Questions?