

Devanshu Surana

PC-23, 1032210755

Panel C, Batch C1

DEC Lab Assignment 3

Aim: Preprocess data using python

Problem Statement: Data Pre-processing using Python
(Part II)

Objectives:

Data Integration

Data Redundancy and correlation Analysis

Tuple Duplication

Data Transformation

Normalization Min-max, z-score

Data smoothening - Binning Methods (on dataset such as csv/xls file)

Data Reduction

PCA method.

Theory:

1. ~~Pearson~~ Correlation: NumPy and SciPy implementation
Pearson correlation is a Statistical method used to measure the linear relationship betⁿ two continuous variables.

NumPy and SciPy, python libraries provide function like 'numpy' 'corrcoef' and 'scipy.stats' pearson' to calculate pearson correlation coefficients betⁿ data arrays.

These functions helps access know strongly and in what direction two variables are correlated with values ranging from -1 to 1 .

2. Pearson Correlation: Pandas Implementation

Pandas is a popular data manipulation library, offers the `pandas.DataFrame.corr` method to compute pearson correlation coefficients within a Dataframe.

We can use this method to quickly analyze and understand the relationships betⁿ multiple variable in a structured data set.

3. Visualization of Correlation:

To visualize correlation between variables, you can create correlation matrices and use various visualization techniques.

4. Heatmaps of correlation Matrices

Heatmaps are the common way to visualize correlation matrices where colors represent the strength and direction of correlation.

5. Feature Normalization and their techniques.

It is the process of scaling data to ensure all features have a similar influence on machine learning algo. common techniques includes:

Minimax Scaling

Standardization

Log Transformation

Conclusion :

Hence we have ^{studied} ~~submitted~~ Data preprocessing i.e Data integration, Transformation and Reduction.

FAQ's

1. Why do we need scaling?

→ Scaling is needed to ensure that all features in a dataset have a similar impact on machine learning models. It helps prevent features with larger from dominating those with smaller scales.

2. Benefits and Techniques of Binning in Python.

→ Binning discretizes continuous data into intervals or bins, simplifying complex data, reducing noise and aiding analysis. Techniques like equal-width and equal-frequency binning are used.

3. What is Data leakage. How to avoid any data leakage during the model testing process?

→ Data leakage happens when information from the test set unintentionally leaks into the training process, causing overly optimistic model performance. To prevent this, apply all data preprocessing and feature engineering steps consistently.

4. Which technique we should use Normalization or standardization?

→ It depends on the algorithm and the data distribution. Use standardization (z-score) for algorithms sensitive to varying scales or

normalization (Min-max Scaling) for models relying on values within specific ranges

5. What are the benefits of correlation Analysis?

→ It helps identify relationships between variables allowing prediction, understanding dependencies, selecting relevant features and detecting multicollinearity in datasets.

6. What is the significance of correlation analysis?

→ Correlation analysis is significant as it helps in understanding the strength and direction of relationships betⁿ var, aiding in decision-making, feature selection, and predicting the behaviour of related attributes in datasets.

7. What are the different kinds of correlation analysis. Discuss their strength and weakness.

→ a) Pearson Correlation Coefficient

Strengths:

1. Measures linear relationships betⁿ continuous variables.

2. Easy to interpret, with values ranging from its weakness:

1. Assumes that the variables are normally distributed and have a linear relationship.

b) Spearman Rank Correlation Coefficient ~~st.~~

Strengths:

1. Non-parametric, meaning it doesn't rely on specific data distribution assumption.

Measures monotonic (non-linear) relationship between variable. ∞

Weakness:

1. Less sensitive to subtle linear relationship
2. Ignores specific data values, only focusing on their ranks.

c) Kendall's Tau (τ)

Strengths:

1. Also non-parametric and suitable for non-linear relationship.
2. Measures association between variables ordinal data.

Weakness:

1. Less commonly used in applications compared to Pearson's and Spearman correlation.
2. Computationally more intensive for large dataset.

8. What are the factors that affect a correlation Analysis?

→ Sample size

- Data Distribution

and the choice of correlation coefficient can

~~affect~~ the results of a correlation analysis.

9. Write a short note on:

a. The correlation coefficient:

→ It quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1 where -1 is negative linear relationship $+1$ is positive linear relationship and 0 is

no relationship.

b. The p-value :

→ The p-value in correlation analysis measures the significance of the calculated correlation coefficient. A low p-value indicates a statically significant correlation meaning that the observed relationship is unlikely to be due to chance. Conversely a high p-value suggest a weaker or non-significant correlation.

