Prof.Sangsan Lee, TA:Shinkook Cha
School of Mechanical and Control Engineering
Handong Global University, Pohang

February 6, 2024

ASSIGNMENT 1 — Data Preprocessing and Visualization

1 Download the Dataset

- Download the `Wine` dataset provided from the UC Irvine Machine Learning Repository by using the **ucimlrepo** package: `https://archive.ics.uci.edu/dataset/109/wine`
- Download `diamonds.csv` from `https://www.kaggle.com/datasets/ulrikthygepedersen/diamonds?resource=download`
- **For Part 1, use the `Wine` dataset**
- **For Part 2, use the `diamonds` dataset**

2 Part 1

2.1 Skim through the Dataset

2.1.1 Table Visualization

- Print the `Wine` dataset's metadata and its variable information
- Print the first five rows of the dataset
- Print a table in which the columns are the variables and the rows are:
  1. count
  2. mean
  3. standard deviation
  4. minimum value
  5. 25% percentile
  6. median
  7. 75% percentile
  8. maximum value

2.1.2 Histogram Visualization

- Plot a histogram for each variables of the `Wine` dataset.

2.2 Train-Test Split

2.2.1 User-defined Function

- Define a train-test split function named `'Split_train_test'` from page 85 of our textbook.
- Divide the `Wine` dataset into train and test sets. Set the test ratio as 0.2
- Print the length of the train set and the test set.
- Describe two limitations of this function.

2.2.2 Scikit Learn's Function

- Using Scikit Learn's `train_test_split` function, divide the `Wine` dataset into train and test sets. Set the test ratio as 0.2
- Print the length of the train set and the test set.

2.3 Correlation

2.3.1 Correlation Visualization

- Visualize the correlation between each variables of the `Wine` dataset using scatter plot.

2.3.2 Correlation Calculation

- Calculate the correlation of `'Flavanoids'` with other variables and sort them in descending order.
- Name the variable with the highest correlation with `'Flavanoids'` apart from itself.

2.4 Missing Values

2.4.1 Check for missing Values

- Print the number of missing values in each variable in the `Wine` dataset.

2.4.2 Data Imputation

- Create missing values in the `Wine` dataset (Note: Missing rate is 0.1).
- Print the number of missing values in each variable.
- Impute the mean value of each variable into the missing values.

3 Part 2

3.1 Encoding

3.1.1 Ordinal Encoding

- For all categorical variables from the `diamonds` dataset, encode the data and replace it to the variables in the original dataframe.

3.1.2 One-Hot Encoding

- For all categorical variables from the `diamonds` dataset, encode the data, convert it into numpy array format and print the first five sample's encoded `cut` variable.

3.2 Scaling

3.2.1 MinMaxScaling

- For all continuous variables in the `diamonds` dataset, min-max scale the data and replace it to the variables in the original dataframe.
- Print the first five rows of the new dataframe.

3.2.2 StandardScaling

- For all continuous variables in the `diamonds` dataset, standard scale (i.e., normalize) the data and replace it to the variables in the original dataframe.
- Print the first five rows of the new dataframe.

Submitted by Prof.Sangsan Lee, TA:Shinkook Cha on February 6, 2024.