Indian Institute of Technology Dharwad

CS214: Artificial Intelligence Laboratory

Lab 3: Outlier Detection, Standardization, Normalization and Correlation of Data

Note

Two problems need to be worked out in the lab. **Problem Statement 1** is compulsory. **Problem Statement 2** is a bonus.

Problem Statement 1

You are given the ***"landslide_data_original.csv"*** file. This dataset contains the readings from various sensors installed at 10 locations around the Mandi district in Himachal Pradesh. These sensors give the details about the factors like temperature, humidity, pressure, rain etc. Write a python program to do the following.

1. **Outlier Detection**

   (a) Read the data into a dataframe using pandas. Obtain the boxplot of all the attributes (exclude the attributes *"dates"* and *"stationid"*). Observe the number of outliers in each attribute and their values. Outliers are the values that do not satisfy the condition: $(Q1 - 1.5 \times IQR) < x < (Q3 + 1.5 \times IQR)$ where, IQR is the Interquartile range $(Q3 - Q1)$, where $Q1$ and $Q3$ are the lower and upper quartiles.

   (b) Replace these outliers with the median of the attribute. After outlier correction, save it as ***"landslide_data_corrected.csv"*** file since it is required in the next task. Plot the boxplot again and observe the difference. Do you still get outliers? Why? (You may use $Q1 = df.quantile(0.25)$ and $Q3 = df.quantile(0.75)$ in pandas)

2. **Correlation**

   (a) Find the Pearson's and Spearman's correlation coefficients for each pair of attributes for the dataset after outlier correction.

   (b) Visualize these correlation coefficients as a correlation matrix. (Use the heatmap plot to visualize it). Observe the correlation coefficients and the degree of correlation (shown in Table 1), and print the relation for each attribute pair based on the degree of correlation.

   (c) Let's assume *"rain"* as the target attribute. Now, find and visualize (use bar plot) the Pearson's and Spearman's correlation coefficients of the target attribute with all attributes for the dataset, both before and after outlier correction. Further, show the scatter plots between the target attribute and all other attributes and analyze them with their respective correlation coefficient.

3. **Standardization and Normalization**

   (a) Observe the range of the values in each attribute (Use the data obtained after outlier correction). Find the minimum and maximum values in each attribute.

   (b) Perform the Min-Max normalization of this data to have the range of values between 0-1. Do not use any inbuilt function/library for Min-Max normalization. Also, observe the correlation on the normalized dataset and compare it with the previously calculated correlation.

Degree of Correlation	**Positive Correlation**	**Negative Correlation**
Perfect Correlation	+1.0	-1.0
Very Strong Correlation	$(0.5, 1.0)$	$(-1.0, -0.5)$
Strong Correlation	$(0.3, 0.5]$	$[-0.5, -0.3)$
Moderate Correlation	$(0.1, 0.3]$	$[-0.3, -0.1)$
Weak Correlation	$(0.0, 0.1]$	$[-0.1, 0.0)$
Zero/No Correlation (uncorrelated)	0.0	0.0

Table 1: Degree of Correlation

(c) Perform Min-Max normalization to have the range of values between 0-20. Do not use any inbuilt function/library for Min-Max normalization.

(d) Use the data obtained after outlier correction. Find the mean and standard deviation of the attributes. Standardize each attribute (exclude the attributes *"dates"* and *"stationid"*) using the relation $x_{new} = \frac{x-\mu}{\sigma}$ where μ is mean and σ is standard deviation. Compare the mean and standard deviations before and after the standardization. Do not use any inbuilt function/library for standardization. Also, observe the correlation on the standardized dataset and compare it with the previously calculated correlation..

(e) Repeat steps (b), (c), and (d) using scikit-learn instead of pandas. You can use the functions *StandardScaler* and *MinMaxScaler* in scikit-learn.

Problem Statement 2

A dataset related to red variants of the Portuguese "Vinho Verde" wine is given. This dataset contains the values of different physicochemical tests from each sample of red wine [1]. The original goal of the dataset is to model wine quality based on physicochemical tests. The attributes of the dataset based on physicochemical tests are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH value, sulphates, alcohol content, and the last attribute is quality. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

You are given the ***"winequality_red_original.csv"*** file. Write a Python program to do the following.

1. **Outlier Detection**

(a) Read the data into a dataframe using pandas. Obtain the boxplot of all the attributes (exclude the attribute *"quality"*). Observe the number of outliers in each attribute and their values. Outliers are the values that do not satisfy the condition: $(Q1 - 1.5 \times IQR) < x < (Q3 + 1.5 \times IQR)$ where, IQR is the Interquartile range $(Q3 - Q1)$, where $Q1$ and $Q3$ are the lower and upper quartiles.

(b) Replace these outliers with the median of the attribute. After outlier correction, save it as ***"winequality_red_corrected.csv"*** file since it is required in the next task. Plot the boxplot again and observe the difference. Do you still get outliers? Why? (You may use $Q1 = df.quantile(0.25)$ and $Q3 = df.quantile(0.75)$ in pandas).

2. **Correlation**

   (a) Find the Pearson's and Spearman's correlation coefficients for each pair of attributes for the dataset before outlier correction and after outlier correction.

   (b) Visualize these correlation coefficients as a correlation matrix. (Use the heatmap plot to visualize it). Observe the correlation coefficients and the degree of correlation (shown in Table 1), and print the relation for each attribute pair based on the degree of correlation.

   (c) Let's assume *"quality"* as the target attribute. Now, find and visualize (use bar plot) the Pearson's and Spearman's correlation coefficients of the target attribute with all attributes for the dataset, both before and after outlier correction. Further, show the scatter plots between the target attribute and all other attributes and analyze them with their respective correlation coefficient.

3. **Standardization and Normalization**

   (a) Observe the range of the values in each attribute (Use the data obtained after outlier correction). Find the minimum and maximum values in each attribute.

   (b) Perform the Min-Max normalization of this data to have the range of values between 0-1. Do not use any inbuilt function/library for Min-Max normalization. (Make sure that you do not alter the attribute: *"quality"*). Also, observe the correlation on the normalized dataset and compare it with the previously calculated correlation.

   (c) Perform Min-Max normalization to have the range of values between 0-20. Do not use any inbuilt function/library for Min-Max normalization.

   (d) Use the data obtained after outlier correction. Find the mean and standard deviation of the attributes. Standardize each attribute (exclude the attributes *"quality"*) using the relation $x_{new} = \frac{x-\mu}{\sigma}$ where μ is mean and σ is standard deviation. Compare the mean and standard deviations before and after the standardization. Do not use any inbuilt function/library for standardization. Also, observe the correlation on the standardized dataset and compare it with the previously calculated correlation.

   (e) Repeat steps (b), (c), and (d) using scikit-learn instead of pandas. You can use the functions *StandardScaler* and *MinMaxScaler* in scikit-learn.

Note

Please upload the completed Jupyter Notebook for each problem statement separately to Moodle for evaluation. Save the files using the format:

1. **your_rollnumber_problem_statement1.ipynb**
2. **your_rollnumber_problem_statement2.ipynb**.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.