Outlier Detection with IQR (Inter-quartile Range)

OBJECTIVE

Interquartile range (IQR) (https://en.wikipedia.org/wiki/Interquartile_range) is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, which is a popular method to detect outliers for preprocessing the data. The purpose of this lab is to implement IQR algorithm for outlier detection.

PREREQUISITES

Check the lecture slides Part II (Lecture 3.2) and be familiar with IQR algorithm below.

IQR Algorithm for Outlier Detection

1. Arrange the data in ascending order
2. Calculate Q1 (the first Quarter)
3. Calculate Q3 (the third Quartile)
4. Calculate IQR = (Q3 - Q1)
5. Calculate the lower Range T_{lower} = Q1 -(1.5 * IQR)
6. Calculate the upper Range T_{upper} = Q3 + (1.5 * IQR)
7. Detect outliers with the lower Range and the upper Range. If the data is not in the range [T_{lower},T_{upper}], then the data will be filtered out as outliers.

INSTRUCTIONS

- Implement IQR algorithm with Python
- Test IQR on the attribute "LotArea" of training data of **House Price Prediction** from Kaggle Data to detect and remove its outliers. The data can be downloaded from Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview).
- Compare the original data and preprocessed data by plotting (https://matplotlib.org/3.1.1/gallery/pyplots/boxplot_demo_pyplot.html#sphx-glr-gallery-pyplots-boxplot-demo-pyplot-py)

utbm université de technologie Belfort-Montbéliard
NVIDIA. DEEP LEARNING INSTITUTE
Georgia Tech
PRAIRIE VIEW A&M UNIVERSITY