EN3150 Homework 01 Data Pre-processing

Sampath K. Perera

August 7, 2023

Task: Comparing of different data normalization methods.

1. Visit the following web page to get an idea about the available data sets in scikit-learn and how to load them. https://scikit-learn.org/stable/datasets.html
2. Load an available dataset of your choice. For example, you can load the Iris dataset (Since, we have used California housing dataset in the class, please do not use it)
3. Explore the dataset: Print the feature names, target variable (if applicable), and any relevant information about the dataset.
4. Select the features: Choose a subset of features (e.g., say two features) from the dataset for the normalization comparison. [Hint: See the mean, variance, 1^{st} quartile (25^{th} quantile) and the 3^{rd} quartile (75^{th} quantile) of the features and select features that may contain outliers.]
5. Apply different normalization methods. Use the following normalization methods from 'sklearn.preprocessing':
   - Min-Max Scaling (MinMaxScaler)
   - Standard Scaling (StandardScaler)
   - Robust Scaling (RobustScaler)
   - Power Transformer
6. Normalize the data:Apply each normalization method to the selected features.
7. Visualize the data before and after normalization.Create scatter plots or histograms of the original and normalized data to visualize the effects of each normalization method on the feature distributions.
8. Compare how each normalization method scales the data and its impact on outliers.
9. Interpret the Results. Discuss the effects of each normalization method on the data's distribution, scale, and outlier handling. Which methods are more robust for outliers? Why?

Listing 1: Loading an dataset

```
import pandas as pd
from sklearn.datasets import fetch_california_housing

# Load the California housing dataset
dataset = fetch_california_housing()
X_full, y_full = dataset.data, dataset.target
feature_names = dataset.feature_names

# Access the target variable (data labels)
target = dataset.target
target_names = dataset.target_names

# Print the name of the target variable
print("target_names")
print(target_names)
# Print the feature names
print("Feature Names:")
print(feature_names)
```

- Additional resources

    1. Scikit-learn preprocessing data documentation