Project Three: Feature Extraction with Unsupervised Algorithms

STA 551 Foundations of Data Science

Unsupervised learning algorithms for feature extraction are widely used to automatically discover meaningful patterns, reduce dimensionality, or transform raw data into a more informative representation without relying on labeled data (i.e., the response variable).

The goal of this project is to implement some commonly used, simple unsupervised learning algorithms to extract new features implicitly from the existing feature variables.

To evaluate the benefits of model-based feature extraction, we will incorporate these extracted features into a binary classification model and assess their performance using appropriate metrics.

- **Data Set Requirements**: The dataset should include
  - A few numerical variables that are highly correlated (to perform PCA).
  - A binary categorical variable (to build a binary classification model, such as logistic regression, perceptron, decision tree, or bagging).
- **Suggested data sites**:
  - My teaching data repository (https://pengdsci.github.io/datasets/),
  - UCI Machine Learning Repository (https://archive.ics.uci.edu/), and
  - Kaggle (https://www.kaggle.com/datasets?fileType=csv).
- **Methodology**:
  - **Regular EDA and Feature Engineering**
    - Perform exploratory data analysis (EDA) and feature engineering as usual to prepare an analytical dataset.
  - **Unsupervised ML Algorithms for Feature Extraction**
    - Principal Component Analysis (PCA)
    - Clustering
    - Local Outlier Factor (LOF)
  - **Building Binary Classification Models**
    - Train models without using features extracted via unsupervised ML.
    - Train models with regular engineered features.
    - Compare the two approaches and report the advantages and disadvantages of unsupervised feature extraction.
- **Report Format**
  - The report should follow the same structure and components as the previous two projects.