Introduction to Data Science **Assignment 6**

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh** TA(s): Mohammad Saadati

Deadline: Tuesday, Khordad 1st, 11:59 PM

Introduction

In this assignment, we will delve into dimensionality reduction and unsupervised learning tasks. Firstly, we should preprocess the provided dataset to prepare it for analysis. Next, we will apply dimensionality reduction techniques to simplify the dataset's complexity. Then, we will use unsupervised learning algorithms to tackle the task. Finally, we evaluate and analyze the results for comparison.

Dataset

In 2014, some researchers published an article called "Impact of c1HbA Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records." They gathered data on diabetic patients from many hospitals and clinics in America. Some of this data, about 200,000 items with 50 features, has been shared with the public in a way that keeps people's identities private.

Task

1. Preprocessing

Data Preprocessing or Data Preparation is a data mining technique that transforms raw data into an understandable format for ML algorithms. Real-world data is usually noisy (contains errors, outliers, duplicates), incomplete (some values are missed), and could be stored in different places and different formats. The task of Data Preprocessing is to handle these issues.

The dataset contains both numerical and textual values, along with outliers and null values. These inconsistencies can disrupt clustering accuracy. Normalize the data extensively and provide detailed explanations for each normalization step in the report file.

2. Dimensionality Reduction

Some data mining algorithms, like K-Means, struggle with accurately clustering data when confronted with numerous features, leading to high dimensionality. This issue isn't exclusive to datasets with hundreds or thousands of features; even just ten features can pose accuracy challenges.

Feature or dimensionality reduction aims to address this by transforming the original feature set into a smaller set of derived features that retain most of the original information.

Principal Component Analysis (PCA) is a widely used technique for feature reduction. It condenses the original dataset into a set number of features known as principal components. The desired number of principal components must be specified.

In this section, utilize PCA to reduce the dimensionality of the dataset. This approach is recommended if there are numerous variables, manual variable selection is impractical, or segmentation results are unsatisfactory.

3. Unsupervised Learning

Silhouette Method

The Silhouette Method is a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its cluster compared to other clusters. by providing a succinct graphical representation of how well each object has been classified.

K-Means

K-Means Clustering is a type of Unsupervised Machine Learning algorithm that organizes an unlabeled dataset into distinct clusters. This method assigns data points to one of the K clusters based on their proximity to the cluster centers. Initially, cluster centroids are randomly placed in space. Then, each data point is assigned to the nearest cluster centroid. Subsequently, new cluster centroids are calculated. This iterative process continues until it converges on well-defined clusters.

DBSCAN

DBSCAN is an unsupervised clustering algorithm, offering an alternative to KMeans and hierarchical clustering. It relies on two key parameters: Epsilon (ε), defining the neighborhood radius, and Minimum Points (minPts), specifying the minimum number of points to form a cluster. Epsilon determines the similarity threshold between points, influencing cluster size, while minPts affects cluster robustness and noise handling. Balancing these parameters is crucial for effective clustering without splitting valid clusters or aggregating unrelated points.

Utilizing the silhouette method, determine the optimal number of clusters for the K-means method and the optimal input parameters (minPnt, eps) for the DBSCAN method. Based on the obtained values, store the best result from each method in a CSV file containing only the columns: id_encounter, kmean_label, and dbscan_label.

Questions

1. What preprocessing steps did you perform on the dataset? Provide clear reasons for each decision made.
2. What portion of the dataset did you retain during dimensionality reduction, and which variables were retained? Could you elaborate on the rationale behind this decision?
3. Include a plot illustrating the silhouette coefficient plotted against the input parameters for each clustering method within the report file.
4. How can we determine the optimal number of clusters in K-Means?
5. How can we determine the optimal epsilon value and minPts in DBSCAN?
6. When would you recommend using K-Means, and when would you suggest using DBSCAN instead?

Notes

- Upload your work as a zip file in this format on the website: DS_CA6_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- We will run your code during the project delivery, so make sure your results are reproducible.