Fundamentals of Operations Research

Project: Clustering Using Linear Programming and Network Flow

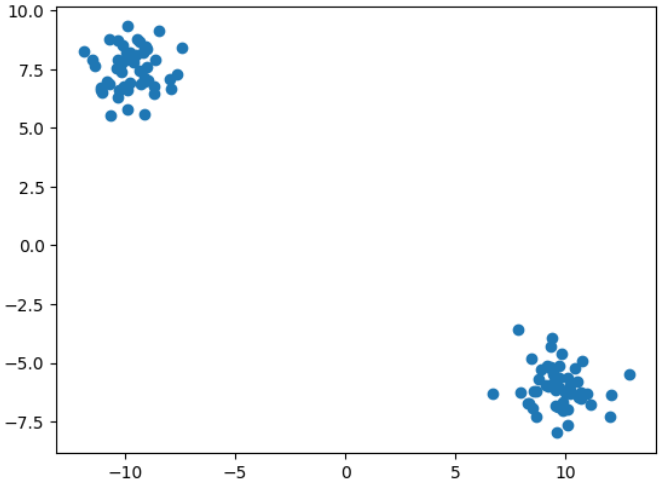
Shiraz University – CSE Department

Introduction

Clustering is a fundamental task in unsupervised machine learning that involves grouping similar data points together based on their characteristics. It aims to identify inherent patterns and structures within the data. One popular clustering algorithm is K-means, which partitions the data into K distinct clusters. In this project, you will be implementing the K-means clustering algorithm in two different ways. You can also try to implement the this task with other algorithms like DBSCAN instead of K-means.

Your dataset is a well clustered data in 2 dimensions. The number of clusters are known in phase 1 and 2.

For example in the following figure you can see a two clusters of data (the number of data in this example is 100) and you task is to find each data is belongs to which cluster.



To recap your project is mainly as follow:

Input: The x_1 and x_2 for each data point with exact number of clusters in a file.

Output: determine each data is belongs to which cluster.

Phase 1: Designing Using Linear Programming Model

In this phase, you will design the K-means clustering algorithm using a linear programming model. You can leverage its optimization techniques to find the optimal cluster assignments.

To implement the designed model, you will use the MiniZinc language. MiniZinc is a free and open-source constraint modeling language that allows you to express optimization problems declaratively. By formulating the K-means clustering problem in MiniZinc, you can utilize its solvers to find the optimal solution.

So your task in this phase is to design and implement the K-means clustering algorithm using a linear programming model in MiniZinc. You will need to define the objective function, constraints, and decision variables to accurately represent the problem. Additionally, you should consider how to handle the initialization of cluster centroids(meas) and convergence criteria within the linear programming framework. You can also implement a model that not only it can find the optimal centroids, but also the summation of all distance from all centroids is minimized.

Phase 2: Implementing Clustering Using Network Design and NetworkX Library

In the second phase of this project, you will implement the K-means clustering algorithm using network design techniques and the NetworkX library. NetworkX is a powerful Python library that provides tools for creating, manipulating, and analyzing complex networks or graphs.

To begin, you will represent the clustering task as a graph using NetworkX, where you will use the concept of similarity or distance, data points, clusters and everything else to build the graph. This graph representation allows you to leverage network design techniques for clustering.

For better understanding the problem explain how the graph representation captures the relationships between data points, and analyze how different network design (Assignment, Transportation,...) techniques impact the clustering results. Try to find the simplest model you can.

Phase 3: Number of Clusters Estimation (Bonus)

Moving forward, the third phase of this project will focus on estimation of the optimal number of clusters for a dataset that has an unknown number of clusters. This will enhance the overall quality and flexibility of the clustering process. In this of this project, the objective is not only to do the clustering task for a dataset but also to find the optimal number of clusters. Determining the correct number of clusters is crucial for achieving meaningful and effective clustering results.

The optimal number of clusters can greatly impact the interpretability and usefulness of the clustering algorithm. A poor choice of the number of clusters may result in either an oversimplified representation of the data or overly fragmented clusters that obscure underlying patterns.

Your task in this phase is to devise or use an approach or metric that helps estimate the appropriate number of clusters for a given dataset. As you can understand you have to use the previous parts codes to do this part.

Please be mention that:

- It is suggested to read some articles about clustering and K-means algorithms.
- You can do the task in the form of a two persons team.
- In form of teams, you can consulting in designing the LP or network model but the cheating in code is not accepted.

Good Luck