Applied Property Testing
For Drug Repurposing
Proposal + High Level Design

Name: Omer Mualem

Advisor: Dr. Sarel Cohen

Abstract

In this project we intend to achieve better results for a recent Drug Repurposing paper by Cohen et al[PLOS ONE 2023]. The paper's goal was to leverage existing FDA-approved drugs through a process known as drug repurposing. By repurposing drugs that have already undergone rigorous testing for other indications, they seeked to expedite the identification and deployment of treatments for novel viruses, such as the Coronavirus.

Our methodology includes using a different clustering algorithm from the field of applied property testing. The clustering algorithm, first presented in a paper by Michal Pernas, Dar, Ron and Alon in 2003 named “Testing Of Clustering”, uses randomness and presents a property testing algorithm that can be used to find clusters in a dataset of points in an N-dimensional plane with lower complexity than the currently used algorithm, k-means, in the Drug Repurposing paper. As we will show, this algorithm has the ability to get a good approximation of clusters in multiple datasets and we hope that this algorithm will yield better inference results for the predictive model.

First we test our new algorithm on different datasets, SIFT&GIST, to check whether it’s applicative Then we will test it on the Drug Repurposing dataset with new evidence from clinical trials about a variety of drugs specifically in their effect on the SARS-CoV-2 virus. The trials can help us label which drugs are effective on the virus and which aren’t. Then we can evaluate our clustering ability.

Introduction

Background

In the face of a global viral outbreak, the immediate need for effective treatments becomes paramount. The traditional drug development pipeline, with its stringent regulatory processes and extensive clinical trials, is not conducive to addressing the urgent demand for therapies to combat emerging viruses. Time is of the essence during a pandemic, and delays in treatment development can have devastating consequences. Our project aims to circumvent these challenges by leveraging existing FDA-approved drugs through a process known as drug repurposing. By repurposing drugs that have already undergone rigorous testing for other indications, we seek to expedite the identification and deployment of treatments for novel viruses, such as the Coronavirus.

Current Approach

The original paper's approach combines advanced machine learning techniques with a comprehensive understanding of biological interactions to identify potential drug candidates for the treatment of new viruses. In a collaboration of Dr. Sarel Cohen with Potsdam University, they utilized a sophisticated algorithm based on deep learning principles to explore a Knowledge Graph encompassing relationships between viruses, compounds (drugs), and genes. This algorithm, powered by Graph Neural Networks (GNN), analyzes vast amounts of data to predict novel connections between FDA-approved drugs and newly discovered viruses, focusing specifically on the 32 strains of the Coronavirus. By employing Link Prediction algorithms, we generate a matrix representing potential interactions between drugs and the Coronavirus strains. Subsequent post-processing techniques refine this matrix to produce a prioritized list of 100 drugs with the highest potential for efficacy against the Coronavirus. Using the K-means algorithm to cluster each column vector in the matrix they tried to predict the effectiveness of new drugs on coronavirus strands.

Our Goal

In an effort to enhance the model's capabilities we are using newly found information about the effectiveness of drugs on SARS-CoV-2 virus from recent clinical trials and with the new clustering algorithm we hope to achieve better performance on predicting drug effectiveness on coronavirus strands.

Literature Review

Review Of Currently Used Clustering Method

Clustering algorithms are a type of unsupervised learning method used to group similar data points together. There are many different methods of clustering from partitioning methods and hierarchical methods to graph-based or grid-based methods. The original paper uses the k-means clustering method on the embedded representation from the knowledge graph. The k-means algorithm is used to partition a dataset into k distinct, non-overlapping clusters (Figure 1). The goal is to minimize the variance within each cluster and maximize the variance between clusters.
The steps of the k-means algorithm are as follows:

1. **Initialization:** Randomly select k initial centroids from the dataset.
2. **Assignment:** Assign each data point to the nearest centroid, forming k clusters.
3. **Update:** Recalculate the centroids as the mean of all data points assigned to each cluster.
4. **Iteration:** Repeat the assignment and update steps until the centroids no longer change significantly or a predetermined number of iterations is reached.

The algorithm is simple, and with a small number of k it is very efficient because it only calculates the distance from every data point to each of the k cluster centroid in each iteration. On the other hand, for large k's the calculation is cumbersome on large datasets. Furthermore, the algorithm will find exactly k clusters even when the data itself can be clustered in less, which leads to misinterpretations in the model outputs on unseen data.

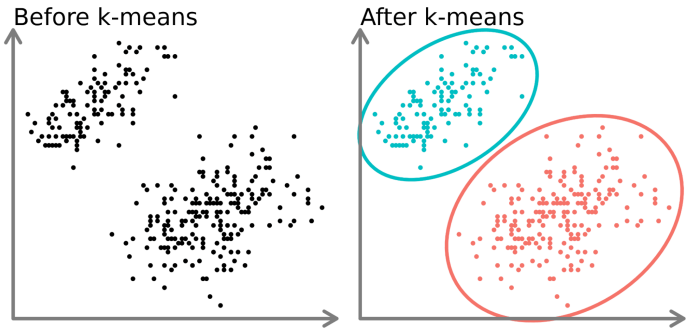

Figure 1: K-means clustering example

Random Algorithms And Property Testing

Random clustering algorithms are a category of clustering methods that do not rely on deterministic processes to form clusters. Instead, they use randomness to guide the clustering process.
Our algorithm comes from the subfield of theoretical computer science called Property testing. Property testing focuses on designing efficient algorithms to quickly determine whether a given object (such as a graph, function, or data set) has a certain property or is significantly different from having that property. In the context of clustering problems, property testing involves developing algorithms that can efficiently assess whether a data set has certain clustering properties without necessarily performing a full clustering and without scanning the entire data.

Methodology

Our Clustering Algorithm

Our new algorithm is a random clustering algorithm first suggested in a paper by Michal Pernas, Dar, Ron and Alon in 2003. The algorithm attempts to get a better approximation on the clustering problem which is an NP-Hard problem, using a Property Testing outlook. In Property Testing our goal is to sample as few points as possible from the dataset, preferably not relying on the dataset's size.

To discuss the algorithm, we need to introduce and define some terms:
Given a set of points X, a max radius b,a max number of clusters k and a distance parameter ε, We define X to be (k,b)-Clusterable if X can be divided to k clusters where the cost of each cluster is $\leq$ b. The cost in our implementation is the radius of points that belongs to the same cluster.
We define X to be ε-far from being (k,b)-Clusterable if at least ε percentage of points from X need to be removed so that X would be considered (k,b)-Clusterable.

The algorithm relies on the assumption that the dataset points hold the triangle inequality and that each cluster in X has a cost $\leq$ b.
In our dataset the data points holds the triangle inequality by being a vector space on R. The second assumption cannot be known because the data is in high dimensions for different kinds of drugs. Therefore, by using the algorithm, we will determine its

applicability on this dataset given the resulting accuracy on clinical trials test data of recently tested drugs on different coronavirus strands.

The algorithm idea is to find each cluster a representative so that the distance between two different clusters is > b (Figure 2). If we find more than k representatives then we will reject the input. As proved in the paper, There is a probability $\geq \frac{2}{3}$ for X to be ε-far from being (k, 2b)-Clusterable when the algorithm rejects the input.

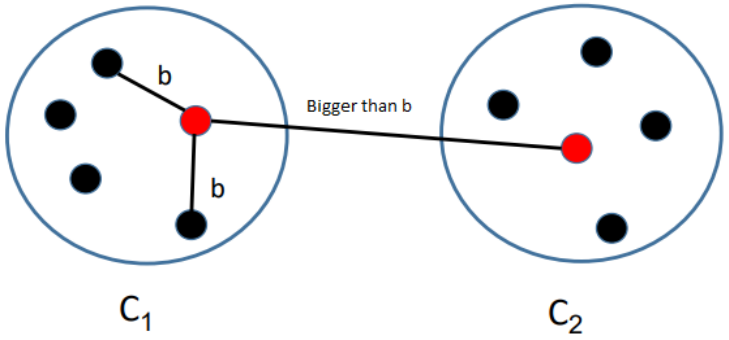

Figure 2: distance between clusters

The steps of the algorithm are as follows:

1. Let Rep_1 be a random point from X

2. i = 1, Find_New_Rep = True

3. For i < k + 1 and Find_New_Rep = True:
   - Choose $\frac{ln(3k)}{\varepsilon}$ points from a uniform distribution on X.
   - If a point x in the sample has a distance > b from all current representors
     then Rep_{i+1} = x and i = i + 1.
   - else, Find_New_Rep = False.

4. If i $\leq$ k then we accept the representors and parameters, else we decline

We will use the algorithm after using the knowledge graph to get the embedded representation (Figure 3). There we will hope to find better clusters than the k-means algorithm output which we hope will give us better accuracy on new drug types.

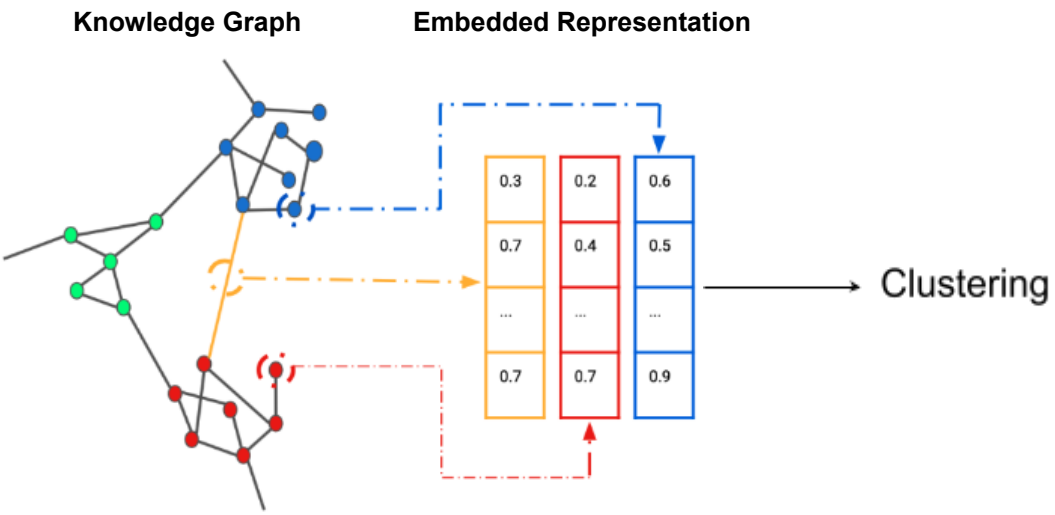

Figure 3: Model's architecture

In order to insert the code cleanly into the existing codebase, we are creating an algorithm object with the same structure as the `sklearn.cluster.KMeans` object and replacing the use of k-means algorithm with our algorithm.
s

Experimental Setup

Dataset

Sarel Cohen work relies on the Drug Repurposing Knowledge Graph (DRKG), which compiles data from different biomedical databases and uses 98 edge types between 4 entity types, namely gene, compound, anatomy and disease. In particular, it contains drugs and substances as compound entities, as well as different COVID-19 variants as disease entities. Finally there are 5000 drug entities and 33 different COVID-19 entities. The edge types include compound-treats-disease edges, which is the kind of edge the model predicts.This is the input for my project.

Data Preprocessing

For data preprocessing we are creating neighborhood graph embeddings (Figure 3), which map fixed-size feature vectors to graph nodes and relations. Finally we get 5000 drug entities and 33 different COVID-19 entities. The edge types include compound-treats-disease edges, which is the kind of edge the model predicts. The nodes and relations graph represented by a matrix is the input to my clustering algorithm.

Algorithm Input

The input graph is represented by a two-dimensional matrix. different COVID-19 variants as disease vs drugs approved by the US Food and Drug Administration as drugs. Every cell i,j in the matrix has a value which is the prediction score of a disease j to be treated by a drug i. The disease feature, as well as the drug feature, is actually a vector consisting of multiple features. These vectors are the inputs for our random clustering algorithm.

Parameters

In our algorithm, there are multiple parameters whose values need to be determined.

- ε - outlier percentage parameter
- b - max radius

- k - max number of clusters

We will test a range of values for each parameter to find the best fit so that every point in the dataset will be included in a cluster with as little outliers as possible. We will test our predictions on new data from clinical trials where drugs affect on COVID-19 variants were tested.

Results Evaluation

On the SIFT & GIST datasets we have labeled test data that we can evaluate the error of our clustering algorithm.

In our Drug Repurposing improvement, we will use clinical trials new data of drugs effectiveness on coronavirus and test if their placement is on the clusters.

Tools and programming languages

In this project we are using python to implement the algorithm. The original work by Dr.Sarel Cohen and Potsdam University is also implemented in python using modules such as scikit-learn, pandas and numpy and our implementation of the algorithm extends these modules with pytorch’s tensorboard for visualizing clusters.

Preliminary Work and Progress

Implementation Stages

1. Create a general pipeline for a clustering algorithm
2. Test the pipeline on k-means with artificial data and display results using tensorboard
3. Create a code implementation of our algorithm
4. Test the pipeline on our algorithm with artificial data and display results using tensorboard
5. Test our algorithm on SIFT & GIST 1M datasets and compare results to other clustering algorithms.
6. Replace current Drug Repurposing paper's k-means algorithm with our new algorithm
7. Test the clustering on new clinical trials data.
8. Infer about the use of the algorithm on the problem.

Current Status

Currently we are on stage 5 in the implementation stages.
We trained our clusters on the SIFT 100,000 training samples and now we need to compare the clustering to the test data labels.
For steps 2 and 4 we generated 200 data points on a 2D plane divided into 5 clusters each with a standard deviation of 1.
We got the following results for step 2 with k-means algorithm:

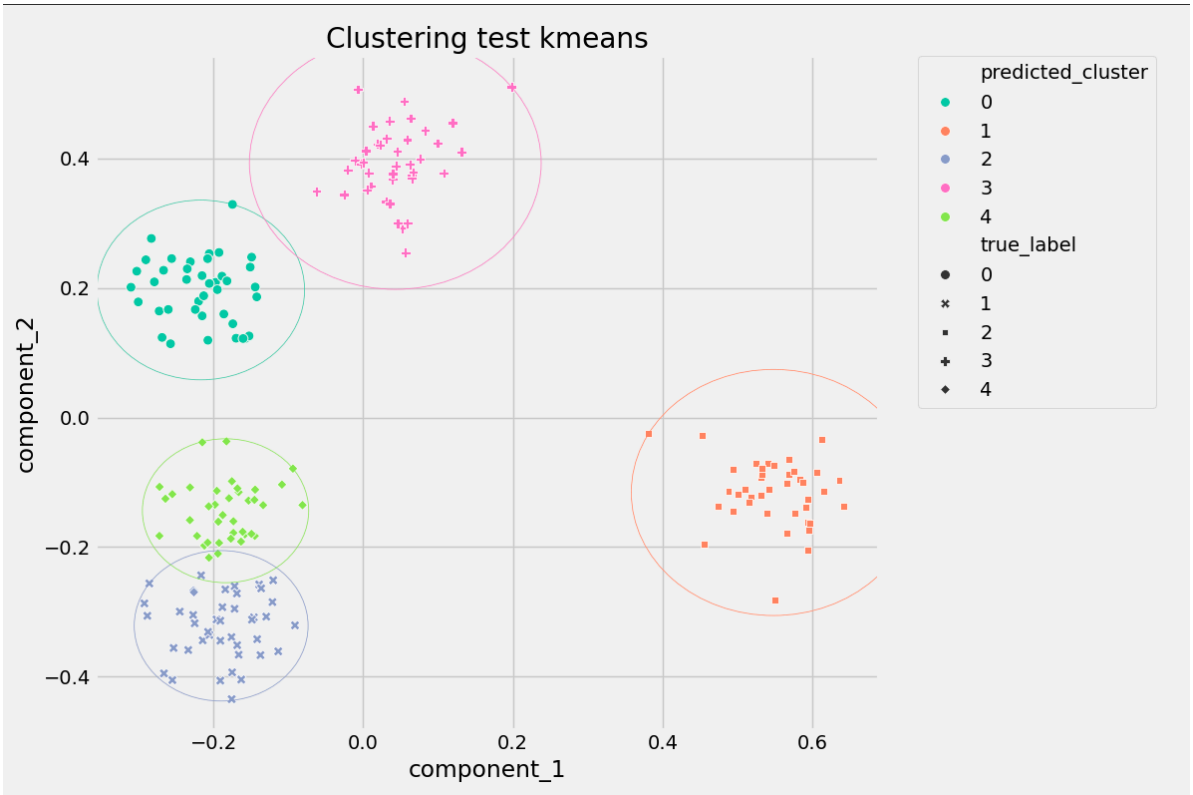

Figure 4: K-means clustering test on artificial data.

On step 4 we tested multiple parameters and these are some of the results:

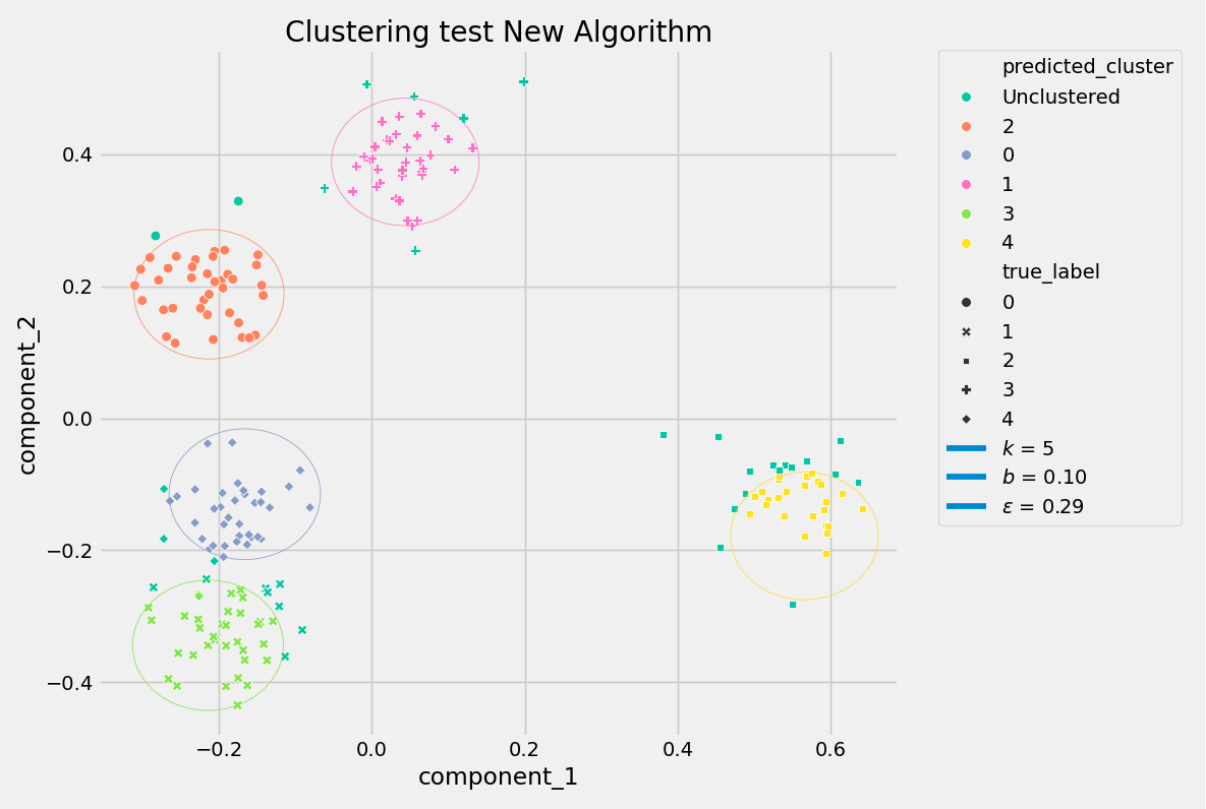

Figure 5: New algorithm clustering test

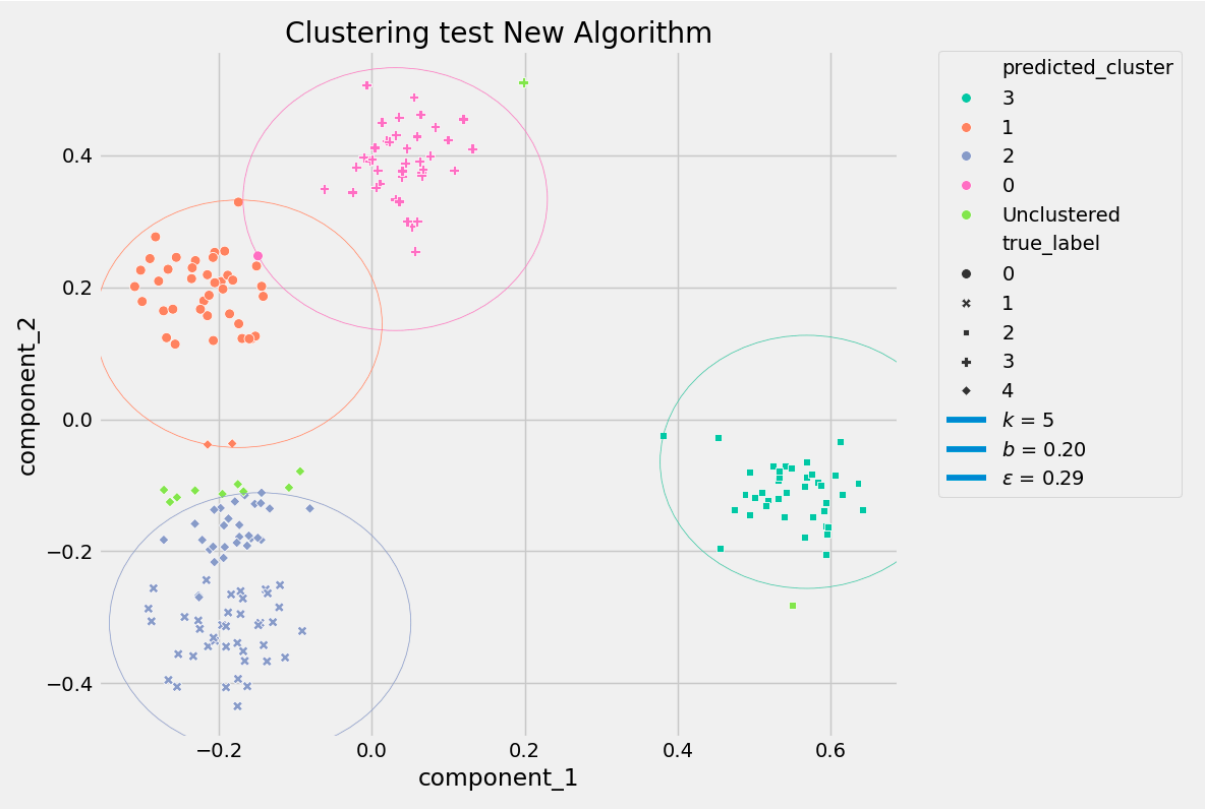

Figure 6: New algorithm clustering test

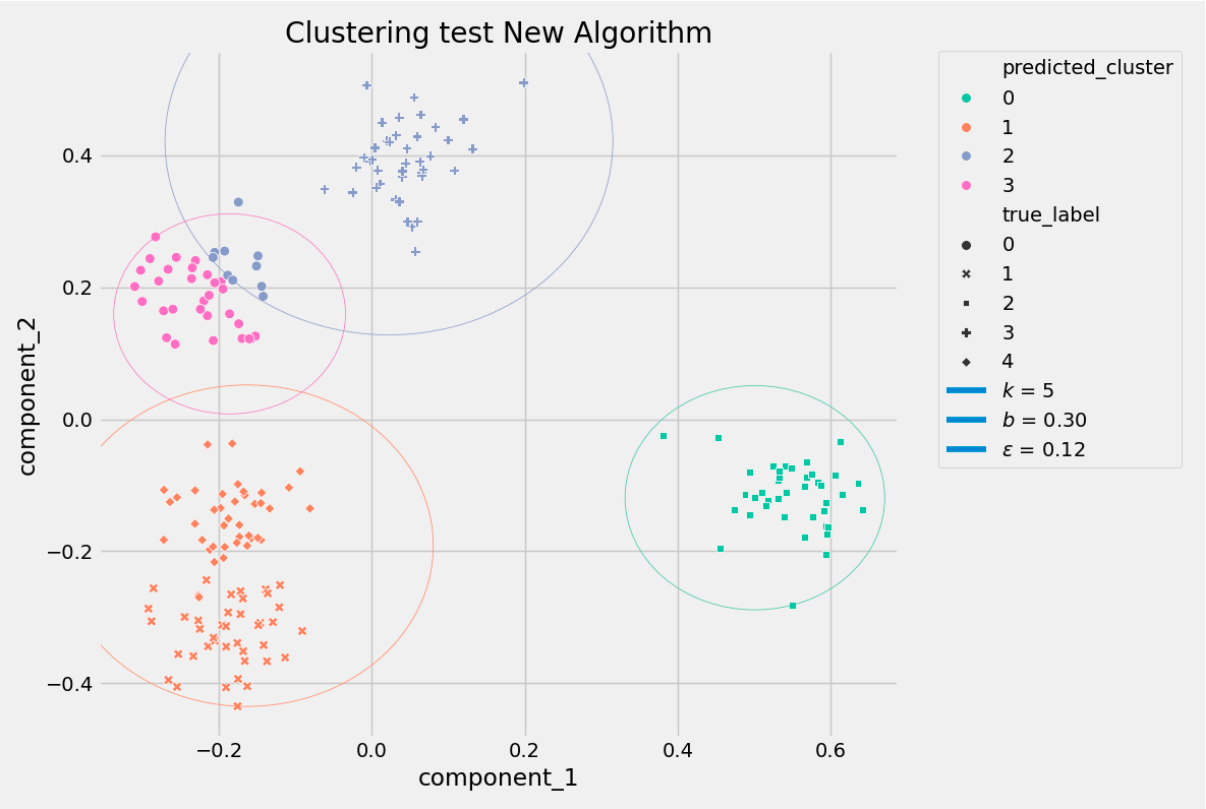

Figure 7: New algorithm clustering test

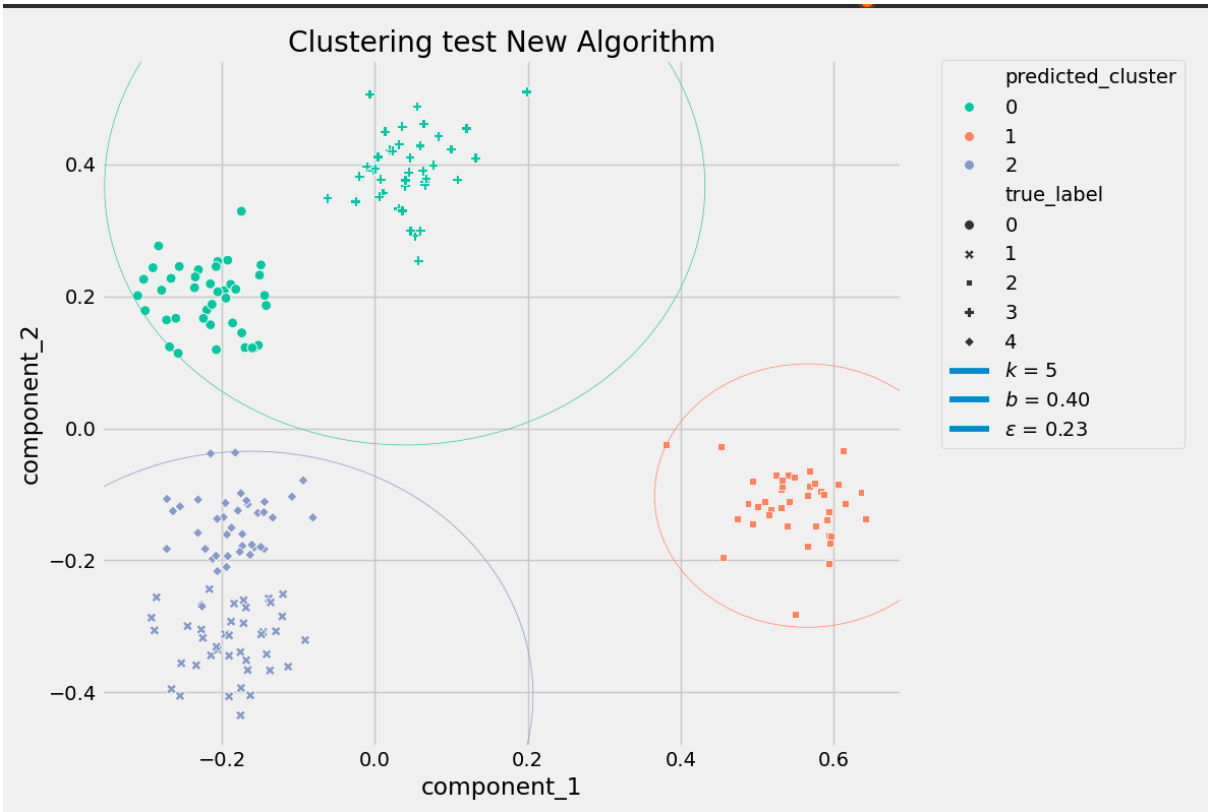

Figure 8: New algorithm clustering test

References

Figure 1: https://www.datacamp.com/tutorial/k-means-clustering-python

Figure 2: Clustering algorithm powerpoint

Figure 3: Avlas Kfir project proposal