Clustering and Visualization of Patient Z's genetic fingerprints

Saurabh Jain

Department of Computer Science, University of Southern California,

Los Angeles, California 90089, USA

(Dated: March 11, 2021)

Abstract

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The principal task of this assignment is to create clusters based on the available data using K-Means clustering method and detecting which cluster patient Z belongs leading to the correct dosages of vaccines to that that cluster. The major outcomes to be noted are that the optimal number of clusters is 5 with 2866 patients in the same cluster as Patient Z beside patient Z. Patient Z belongs to cluster '2' with respect to the plot displayed in the text. The data consists of 14398 patients. PCA is used to visualize the clusters on a plot by reducing their dimensionality. Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

I. INTRODUCTION

This assignment is related to unsupervised learning. The primary task is to identify the main clusters in the data. Next, the task is to decide how many clusters exist and where they are. Next, the task is to find which cluster the patient Z belongs to. People from that cluster are likely to have the same covid-resistance as patient Z. We not have any test-set to self-check how good the predictions are. K-means clustering algorithm was used to cluster the data set. PCA was used to visualize the clusters formed as a result of K-means. [1]

Each genetic fingerprint is represented by a vector of 386 numbers. There are 14,398 patients. PCA was used to reduce its dimensionality from 386 to 2. Various plots were formulated to confirm that the optimal number of clusters is 5.

II. DATA EXPLORATION

Each genetic fingerprint is represented by a vector of 386 numbers. There are 14,398 patients. Standard Scaler was used to scale the data as well. Patient Z is an array of 386 numbers (a genetic fingerprint). The two numpy arrays (one containing 14,398 patients' genetic fingerprint data and the other containing patient Z's genetic fingerprint) were combined and then used to perform K-Means and PCA for visualization purposes.

III. DATA PREPROCESSING

After the data was concatenated (one containing 14,398 patients' genetic fingerprint data and the other containing patient Z's genetic fingerprint), initial PCA was computed and the graph can be visualized in Fig. 1 and Fig. 2. This is a great visualization before performing K-Means.

Before performing K-Means for the essential number of clusters, an essential step is to find the optimal number of clusters than can be used to perform K-Means. An elbow curve and a silhouette score curve was plotted to find the optimal number of clusters by performing K-Means on a for loop. This can be visualized in Fig. 3 and Fig. 4.

After trying PCA with various clusters, a bar plot was created to check the variance % versus PCA features. This can be visualized in Fig. 5.

[1] sklearn library was used to complete various tasks.

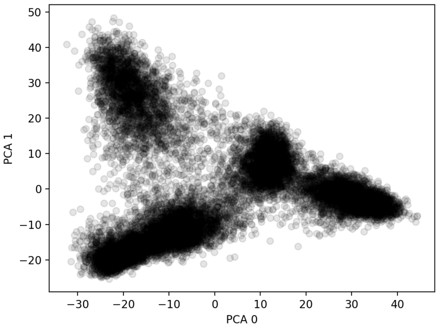

FIG. 1. PCA before K-Means

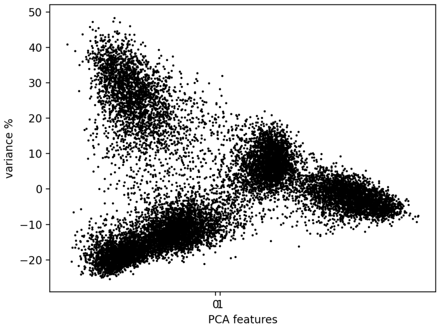

FIG. 2. PCA before K-Means

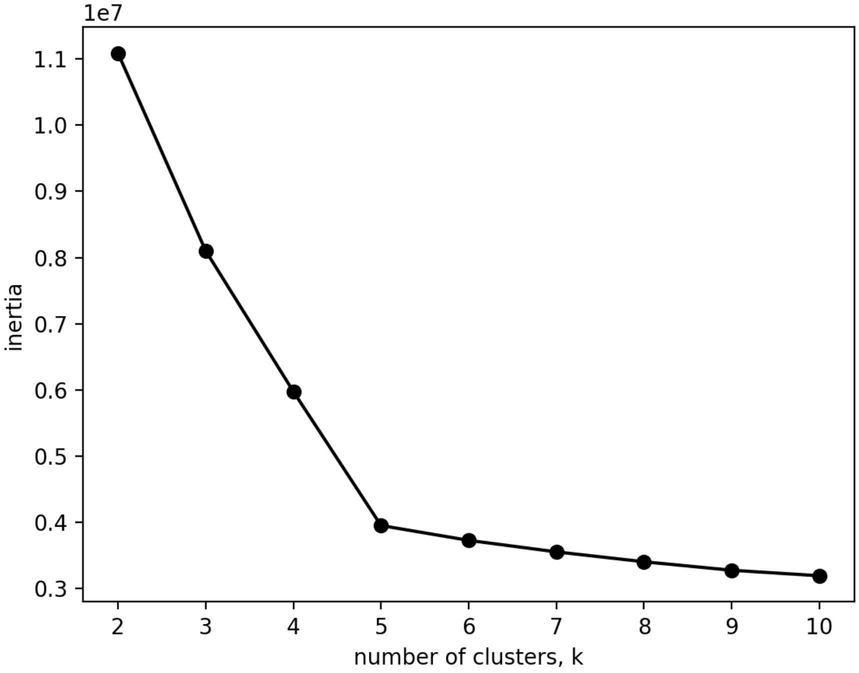

FIG. 3. Elbow Curve

As can be visualized from the graph, the variance % drops on the 4th column, which is essentially the 5th cluster. Therefore, we can conclude that the optimal number of clusters is 5 for this particular data.

IV. MODEL SELECTION

K-Means was now computed using 5 clusters using sklearn's library KMeans. Using this, labels were created to assign each patient to a cluster. This was then used to plot the 5

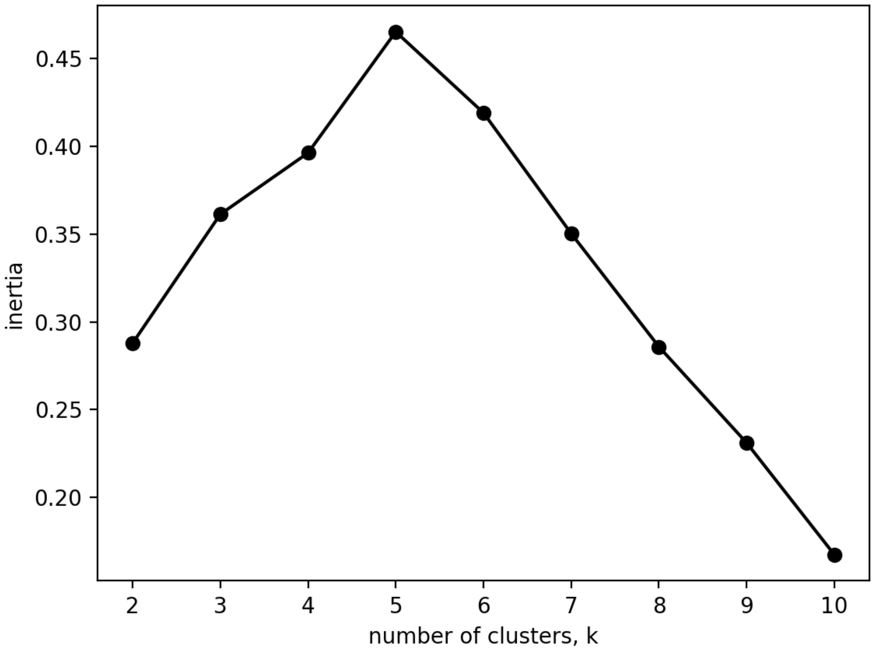

FIG. 4. Silhouette Score Curve

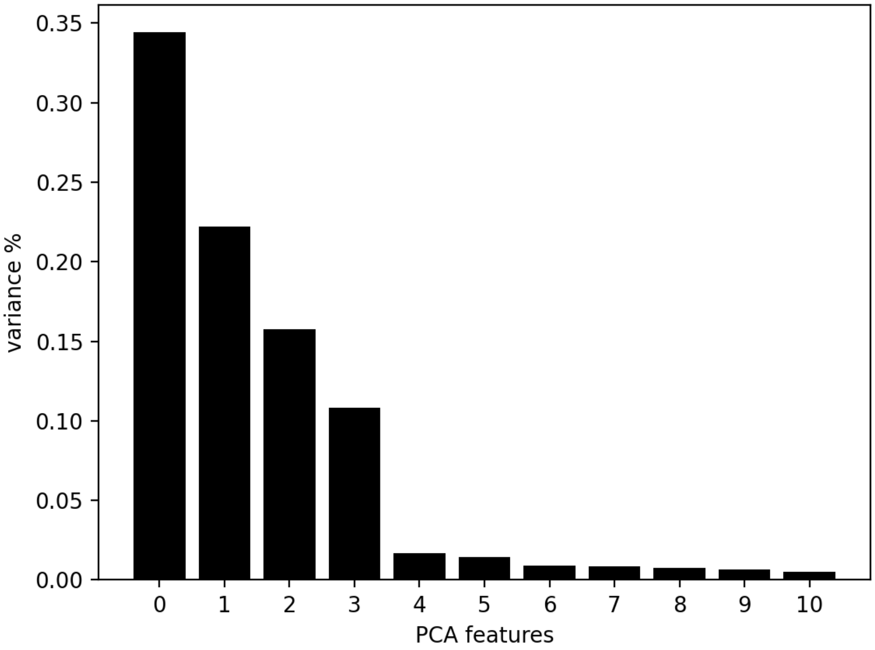

FIG. 5. Variance %

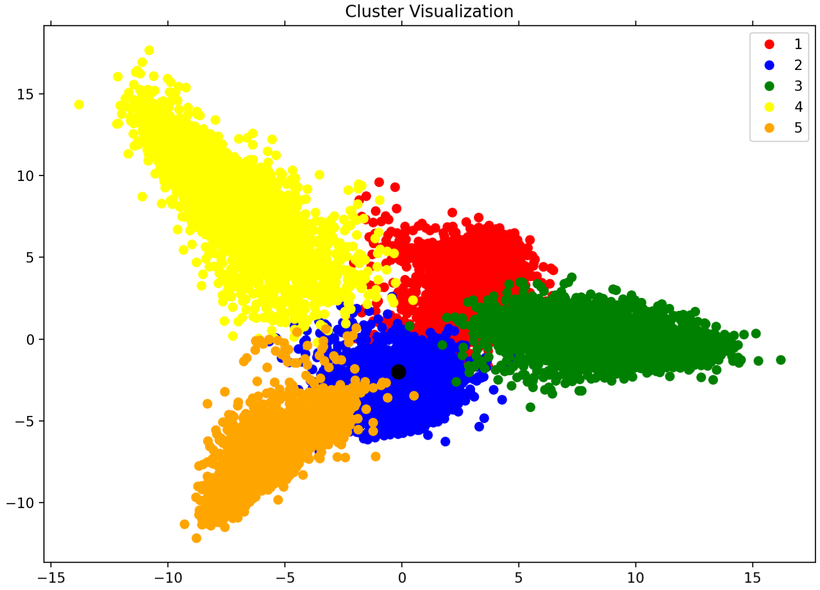

FIG. 6. Cluster Visualization with Patient Z shown using black dot.

clusters using PCA along with plotting patient Z, showing the cluster in which patient Z belongs to. This can be visualized in Fig. 6.

V. MODEL EVALUATION

We can conclude that the patient Z belongs to the blue colored cluster which is numbered '2' in the legend. This can further be visualized in in an enlarged image in Fig. 7 which just plots the patient Z and its corresponding cluster.

VI. INTERPRETATION

It can finally be interpreted what cluster patient Z belongs to. There are 2866 other patients in that cluster and there are 5 total clusters. See Fig. 8

VII. CONCLUSIONS

The data was clustered using K-Means and visualized using Principal Component Analysis. The optimal number of clusters is 5. With total number of 14,399 patients including

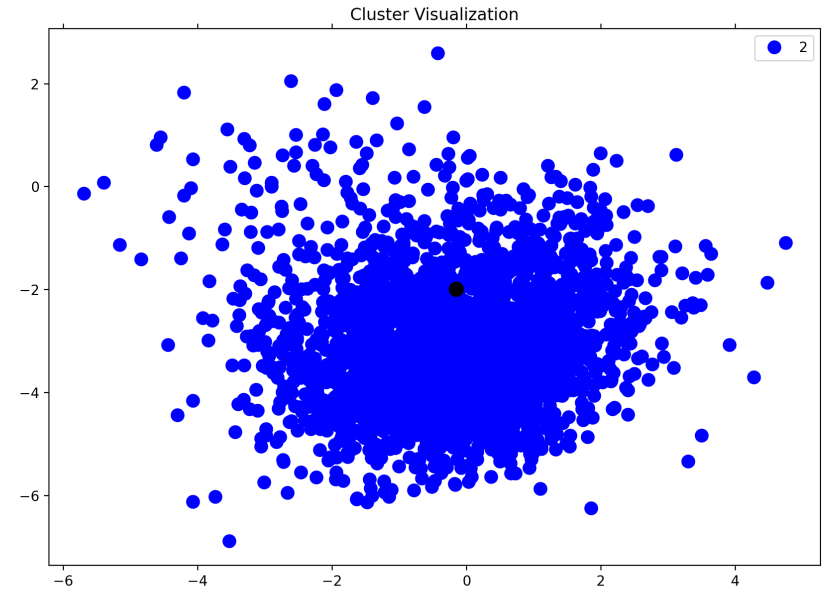

FIG. 7. Cluster comprising Patient Z.

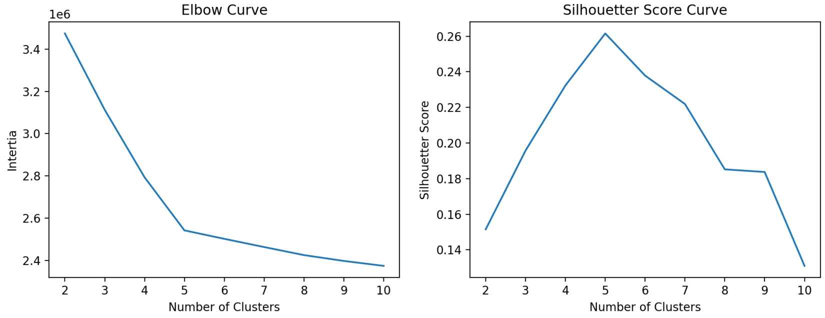

FIG. 8. Graph to compute the optimal number of clusters.

Patient Z, the cluster belonging to patient Z contains 2866 other patients. Each patient has 386 numbers for the genetic fingerprint which was reduced using PCA for dimension reductionality, for plotting and visualizing purposes. The score for Kaggle competition for this model is 0.97538, which suggests that this model and the ideal model have 97.538 patients in common.

DATA AVAILABILITY

Data is available at blackboard namely:

- ps3_patient_zet.npy and : Patient Z genetic fingerprint data (shape: (386,))
- ps3_genetic_fingerprints.npy: Population genetic fingerprint data (shape: (14398, 386))

CODE AVAILABILITY

Code is available at `https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-`

- hw3.py: K-Means, PCA, visualization
- hw3_visualizations.py: some other visualizations

ACKNOWLEDGMENTS

I would like to thank professor Kristina Lerman, professor Keith Burghardt, the TAs and the graders of the course DSCI 552 for all their efforts towards this assignment.

Appendix A: References

1 Wikipedia contributors. "Principal Component Analysis." Wikipedia, 8 Mar. 2021, en.wikipedia.org/wiki/Principal_component_analysis.

1. Wikipedia contributors. "k-Means Clustering." Wikipedia, 8 Mar. 2021, en.wikipedia.org/wiki/k-Means_clustering.