K-Means Clustering: Fairness and Variants

Brie Sloves, Sophie Boileau, Avery Hall, Victor Huang, Jeremiah Mensah, Armira Nance, Muno Siyakurima

Advised by Layla Oesper

Overview

K-means clustering is a process which takes a dataset of size *n* and groups the data points into *k* clusters, where *k* is a fixed integer. The goal is to cluster the data points such that each cluster has the lowest possible cost; in other words, the data points within each cluster are as close as possible to the centroid, or the mean of the data in the cluster.

We implemented different variants of K-Means Clustering: Lloyd's Algorithm, K-Means++, and Fair K-Center, and analyzed them through the lenses of consistency and fairness, defined later.

Data

Our chosen dataset is "High School Longitudinal Study of 2009 (HSLS:09)" which has data about student performance and demographic information. For our project, we selected 5 attributes, 4 of which were used for clustering and the remaining one for the fairness portion of Fair K-Center. The attributes used for clustering are:

- Highest Parent Education Level
- Socioeconomic status
- Annual Income per Household Member
- Weekly Hours of Extracurricular Activity

Our fairness attribute:

- Sex

Lloyd's Algorithm

Consider a dataset containing *n* data points which we want to group into *k* clusters.

We begin by choosing *k* random data points to be centroids.

Then, we repeat the following steps until convergence is reached (i.e. the clusters are unchanging):

- For each of the *n* data points:
  - Calculate the Euclidean distance between the point and each of the *k* centroids
  - Assign the data point to the cluster with the nearest centroid
- Then, for each cluster:
  - Find the mean of all data points in the cluster
  - Reassign the location of the centroid to the calculated mean

Notably, the random initialization step can result in Lloyd's Algorithm converging at a locally optimal, but globally suboptimal clustering.

Upon running the code 10 times:

- 2 attempts yielded clusterings similar to that of Figure 2a
- 6 attempts yielded clusterings similar to that of figure 2b.
- 2 attempts yielded clusterings similar to the one produced by K-Means++, shown in Figure 3

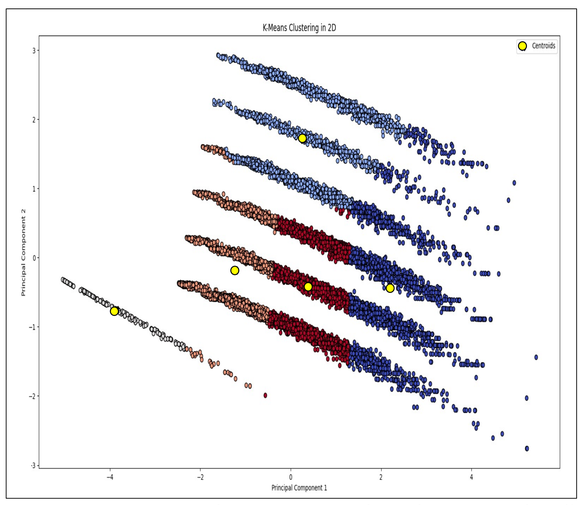

Figure 2a. One clustering produced using Lloyd's Algorithm. Credit: Muno Siyakurima.

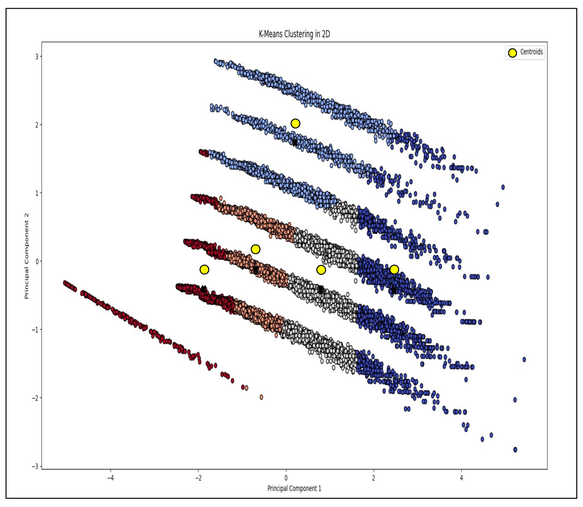

Figure 2b. A second clustering produced using Lloyd's Algorithm. Credit: Muno Siyakurima.

Fair K-Center Results

To implement a fairer clustering, we replicated code outlined in the paper "Fair Clustering Through Fairlets" by Chierichetti et al. Fair clusterings attempt to preserve the ratio of a protected demographic feature from the overall data within each of the clusters. In our code, we chose to use sex as our protected feature, and used the ratio between male and female data instances to assess fairness. The results, displayed below, indicate that the implementation of the code from the paper resulted in more gender-balanced clusterings with little effect on the k-center cost.

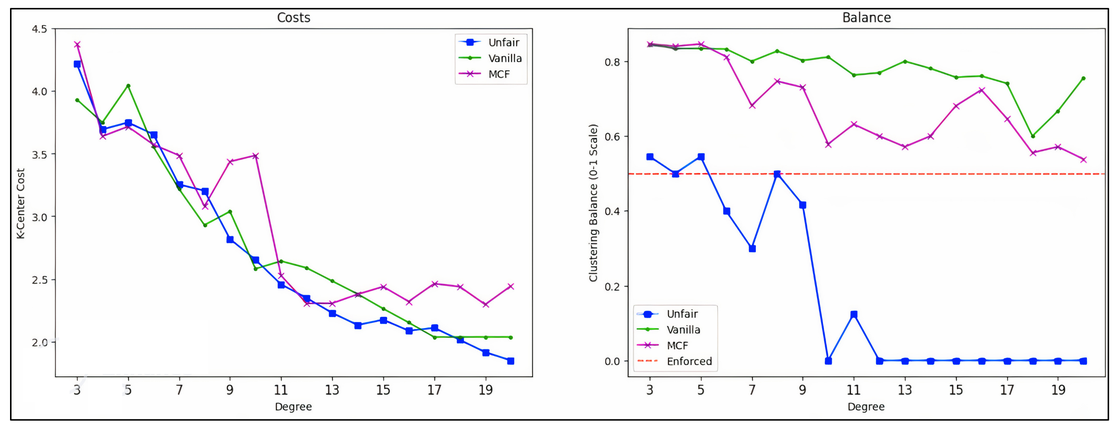

Figure 1. Graphs showing clustering cost and balance for Lloyd's, Vanilla Fairlets, and Minimum Cost Flow implementations. Credit: Victor Huang.

K-Means++ Algorithm

K-means++ clustering is much the same as Lloyd's algorithm, only differing in the initial selection of centroids, outlined below.

Randomly select a data point to be the first centroid. The remaining *k-1* centroids are each selected as follows:

- For each of the *n* data points, find the Euclidean distance from it to the nearest previously selected centroid
- For each of the *n* data points, assign a value proportional to their distance from the nearest centroid (distance/sum of distances)
- Select a centroid randomly from the *n* data points, each point having the probability calculated above as the probability of being selected as a centroid

Once k centroids have been selected, proceed with Lloyd's algorithm as described to the left.

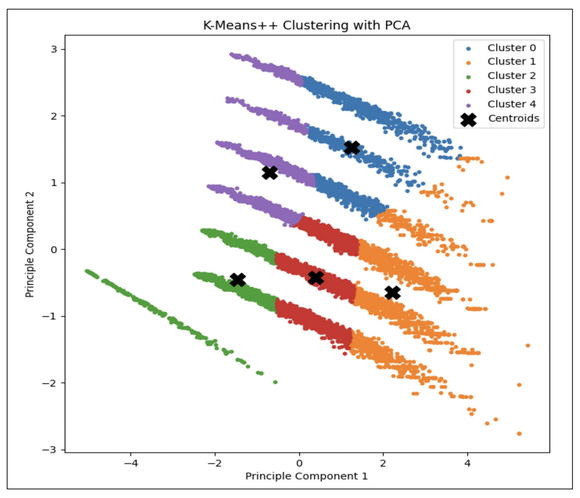

Figure 3. K-means++ clustering of the data. Credit: Jeremiah Mensah.

Upon running the code 10 times:

- 5 attempts yielded clusterings similar to that of Figure 2a
- 3 attempts yielded clusterings similar to that of figure 3.
- 2 attempts yielded other clusterings.

Conclusions

- Both variations of the algorithm outlined in the Fairlets paper led to noticeably fairer clustering while maintaining similar k-center costs.
- Neither Lloyd's and K-Means++ produce consistent clusterings when run repeatedly. This is likely because both have a random initialization step.

References

- Chierichetti, Flavio, et al. Fair Clustering through Fairlets, 2017, dl.acm.org/doi/pdf/10.5555/3295222.3295256.
- "High School Longitudinal Study of 2009 (HSLS:09) - Overview." National Center for Education Statistics, nces.ed.gov/surveys/hsls09/. Accessed 5 Nov. 2023.
- Wohlenberg, Johannes. "3 Versions of K-Means." Medium, Towards Data Science, 2 Apr. 2023, towardsdatascience.com/three-versions-of-k-means-cf939b65f4ea.