Algorithmic Fairness in Clustering

A comps project by Avery Hall, Brie Sloves, Armira Nance, Sophie Boileau, Victor Huang, Muno Siyakurima, and Jeremiah Mensah

Clustering and Fairness

What is k-means clustering?

- K-means clustering is a machine learning algorithm that partitions a dataset into *k* 'clusters'
- Resulting clusters are internally homogenous, giving a sense of the structure of a dataset as well as commonalities and differences amongst portions of its observations
- Clustering is useful in a broad variety of contexts, including market segmentation, traffic flow optimization, image compression, and more

How can a clustering be fair or unfair?

- In applications like loan selection and geographic crime analysis, k-means may form biased clusters with regard to attributes like race and gender, which can produce unfair and potentially damaging real world impacts on social groups
- Fair clustering algorithms aim to neutralize these biases by using variations on traditional k-means approaches
- To ensure fairness, they account for a list of protected attributes whose ratios are intentionally enforced within each cluster

Project Objectives

- Investigate k-means clustering, including its mathematical basis and specific variants
- Implement 3-4 methods of clustering using Python, and run them on our selected dataset
- Draw conclusions about the strengths and weaknesses of different clustering algorithms

Our Dataset

In investigating the effectiveness of different clustering methods, we chose to analyze the dataset **High School Longitudinal Study of 2009** from the National Center for Education Statistics. The study tracked educational data of students in high schools across the U.S. between 2009-2017, including factors such as highest parent's level of education and income per person in the household.

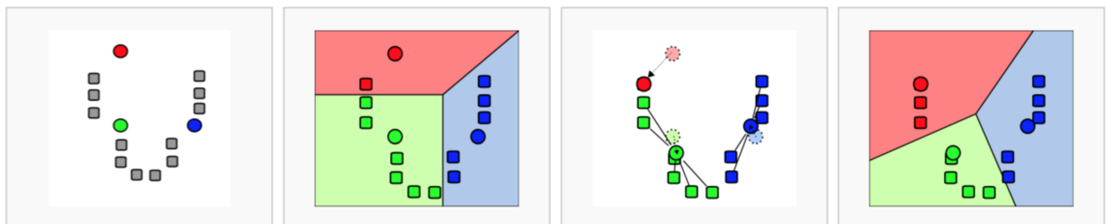
Our Clustering Implementations

In our investigation, we researched a handful of k-means variations that aim to produce fair clusterings, in addition to basic k-means clustering. We worked to implement each of the following algorithms:

Basic k-means

- The basic unsupervised algorithm divides *n* data points into *k* separate internally homogenous clusters, as shown below in Figure 1

Fig 1. K-means clustering steps



Randomly generate *k* initial 'means' (shown in color) within the data (in grey).

Generate *k* clusters by associating each data point with the nearest mean.

The centroids, or mean values of each data cluster, become the new means.

Repeat steps 2 and 3 until the centroids converge.

k++ means Clustering

- Unlike basic k-means, k++ starts by allocating one cluster center randomly, and then searching for other centers given the first one
- Shown to improve runtime and performance

Fair Clustering Through Fairlets

- 'Fairlets' are minimal sets that aim to preserve equal representation of protected attributes (i.e. race, gender) in the dataset
- Rather than being colorblind, this algorithm explicitly uses protected attributes to find a fair solution
- Any dataset can be broken down into fairlets, and then clustered according to traditional algorithms

Socially Fair Clustering (Fair Lloyd's)

- Minimizes the average clustering cost across demographic groups in the dataset
- Exhibits unbiased performance by ensuring that all groups have equal costs in the output k-clustering, while incurring a negligible increase in running time
- Inherits basic Lloyd's inheriting its simplicity, efficiency, and stability, while increasing fairness

PCA: Representing multidimensional data in 2D

- With each of these algorithms, we process multidimensional data that contains dozens of attributes for each observation
- With the Principal Component Analysis algorithm, we are able to reduce the dimensionality of the data for representation and analysis purposes

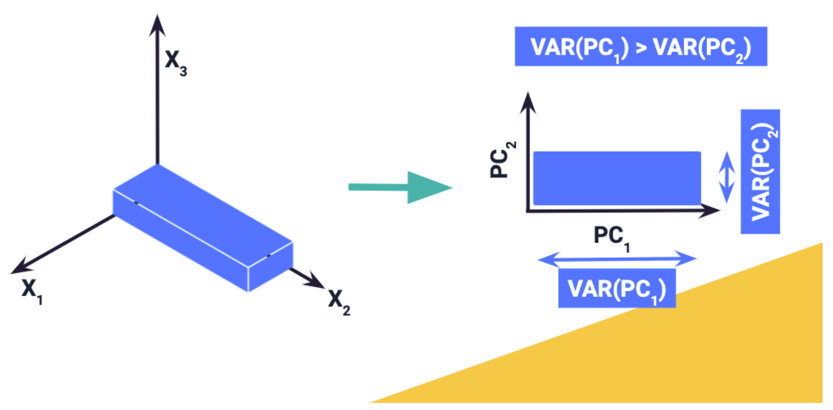


Fig 2. Principal Component Analysis visualized

Results

- We produced and analyzed sample clusterings for each of our completed implementations, as well as a graphical comparison between the different outputs and their key features

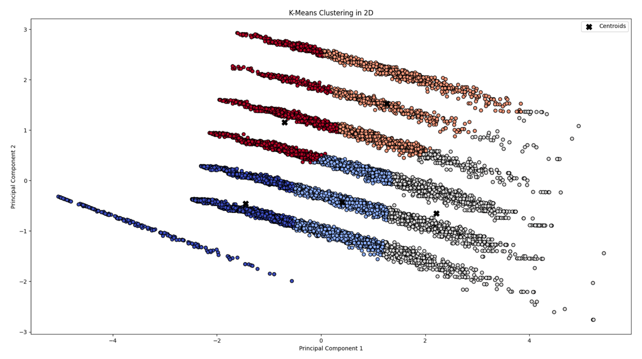


Fig 3. A basic k-means clustering for k=5. Credit: Muno Siyakurima

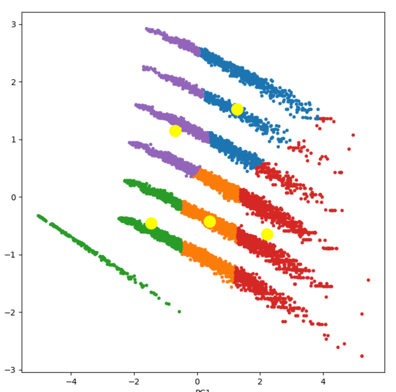


Fig 4. A k++ means clustering for k=5. Credit: Jeremiah Mensah

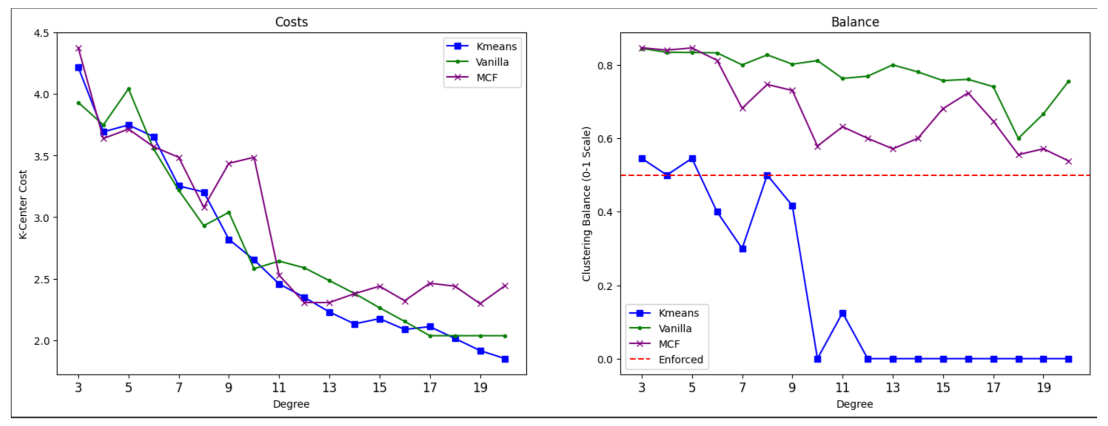


Fig 5: Basic, k++ and MCF compared by cost and balance

- In Figures 3 and 4, we see sample clusterings using basic k-means and k++ means respectively with centroids marked. There appears to be little difference between the clustering shapes produced by each of these methods.
- While costs are relatively correlated between the three variations in Figure 5, we see that vanilla and MCF lie above the center line in the balance graph, indicating that these variations produced fairer clusterings than basic k-means, which lies mainly below the center line.
- Figure 6 shows a breakdown of race percentages within output clusters for a basic k-means run. The ratios within each cluster are not far from proportional to one another, but still have gaps that would not occur in a Fairlets or Socially Fair clustering.

```
--------------------- Cluster 0 ---------------------
Total size: 4719
61.09% White, non-Hispanic
8.07% Black/African-American, non-Hispanic
11.38% Asian, non-Hispanic
8.14% More than one race, non-Hispanic
9.96% Hispanic, race specified
0.57% Amer. Indian/Alaska Native, non-Hispanic
0.57% Native Hawaiian/Pacific Islander, non-Hispanic
0.21% Hispanic, no race specified
0.0% Missing

--------------------- Cluster 1 ---------------------
Total size: 304
50.66% White, non-Hispanic
11.51% Black/African-American, non-Hispanic
9.54% Asian, non-Hispanic
8.22% More than one race, non-Hispanic
13.82% Hispanic, race specified
0.0% Amer. Indian/Alaska Native, non-Hispanic
0.0% Native Hawaiian/Pacific Islander, non-Hispanic
5.26% Hispanic, no race specified
0.99% Missing

--------------------- Cluster 2 ---------------------
Total size: 2169
69.02% White, non-Hispanic
5.03% Black/African-American, non-Hispanic
11.62% Asian, non-Hispanic
```

Fig 6: Sample output for my Python CSV parser that produces demographic data given a variable (i.e. race) for a given clustering arrangement, in this case produced by basic k-means

Conclusions

- Fair clustering algorithms can be effective in increasing fairness of results by minimizing disproportionality of protected attributes like race, gender and family income between clusters
- Basic k-means clustering is insufficient in many real world use cases, and more nuanced fair clustering methods should be adopted across the board. This is especially the case where conclusions drawn from clustering analyses have high potential for real social impact, and specifically the ability to unfairly advantage or disadvantage specific demographics.
- While clustering variations like k++ may be most effective at maximizing efficiency and accuracy of clusterings, others like Fairlets and Socially Fair are most effective at producing partitions relatively free of algorithmic bias.
- Despite the demonstrated strengths of these algorithms, there remains continual room for improvement in runtime, quality of clustering, and fairness in terms preserved ratios of protected. These features are constantly the subject of tradeoffs in clustering algorithms. While respective weaknesses remain, each algorithm can be used in applied scenarios that are most suited to it, in order to maximize efficiency when there is little risk of biases causing harm, or maximizing fairness when the opposite is true.

References

Landman, N. (n.d.). K-means clustering. Brilliant Math & Science Wiki. https://brilliant.org/wiki/k-means-clustering/.

Concept: Principal component analysis (PCA). Concept | Principal Component Analysis (PCA) - Dataiku Knowledge Base. (n.d.). https://knowledge.dataiku.com/latest/ml-analytics/statistics/concept-principal-component-analysis-pca.html

Chierichetti, Flavio et al. "Fair Clustering Through Fairlets." 31st Conference on Neural Information Processing Systems, 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/978fce5bcc4eccc88ad48ce3914124a2-Paper.pdf.

Ghadiri, Mehrdad et al. "Socially Fair k Means Clustering." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, https://dl.acm.org/doi/pdf/10.1145/3442188.3445906.

Acknowledgements

Thank you to my teammates for their passion, engagement with, and dedication toward this comps project. It has been a joy to work with you all.

Thank you also to our advisor, Layla for her consistent guidance and support throughout this process.