Bias in Clustering Systems

Assessing and Mitigating Bias and Discrimination in AI

I Introduction

What is Clustering?

Unsupervised learning method that groups similar objects together. Objects in a cluster are more 'similar' to each other than to objects in other clusters.

Centre-Based Clustering

Splits data into k groups. Each group is represented by a centroid. Points in dataset are assigned to group with closest centroid.

K-Means Algorithm

- Randomly choose k points as initial centroids
- Measure distance between each point and each centroid
- Assign point to cluster with smallest centroid distance
- Update centroids to be the mean of the points in each cluster
- Repeat until centroids and cluster assignments cease to change

Given k clusters and n data points, $C \in \{c_1, c_2, ..., c_k\}, c_i \in \mathbb{R}^d$ is set of centroids and $a \in \{a_1, a_2, ..., a_n\}, a_i \in \{1, 2, ..., k\}$ is set of cluster assignments. Optimal set of cluster centroids $\hat{C}$ and cluster assignments $\hat{a}$ is:

$$\hat{C}, \hat{a} = \arg\min_{C,a} \sum_{i}^{n} \|x_i - c_{a_i}\|^2$$

The set of clusters and assignments that provide the lowest total squared error is the optimal set of cluster centroids and assignments.

- Easy to implement and interpret
- Guaranteed to converge
- sensitive to initial choice of centroid and outliers

Elbow Method

Train several k-means clustering models and measure each one's inertia (total sum of squared errors). Then plot inertia as a function of number of clusters and look for the 'elbow' in the plot. Elbow is where the decrease in inertia begins to slow down, where adding more clusters does not noticeably improve model performance.

II Bias in Clustering Systems

Group Fairness

Groups should be treated equally. No group should be adversely affected or given preferential treatment by decisions made by an algorithmic system.

Common Group Fairness Metrics in Clustering:

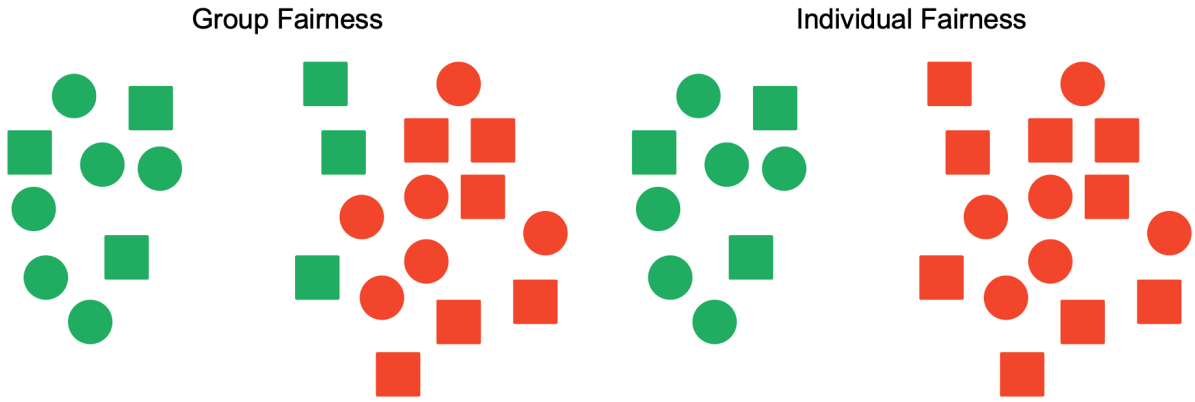
- Balance
- Cluster Distribution KL Divergence
- Social Fairness
- Silhouette Difference

Individual Fairness

Similar individuals should receive similar treatment from an algorithmic system. Similar individuals should be clustered together.
Common Individual Fairness Metrics in Clustering:

- Proportionality
- Aggregate Fairness

Prioritizing group fairness can negatively affect individual fairness and vice versa:



III Measuring Bias

Balance

For m protected groups and k clusters we have r: proportion of samples in the entire dataset that belong to protected group b, r_a proportion of samples in cluster a that belong to protected group b. Balance is defined as:

$$\min_{a\in\{k\},b\in\{m\}} \min\{R_{a,b}, \frac{1}{R_{a,b}}\}, R_{a,b} = \frac{r}{r_a}$$

Balance can take values between 0 and 1. The closer the balance is to 1, the more fair the clustering is.

Cluster KL Divergence

For $m = 2$ protected groups and k clusters, we can calculate the distribution of cluster assignments for each group. Then we can calculate the KL Divergence between the two distributions:

$$KL(PQ) = \sum_k P(k) \log \left(\frac{P(k)}{Q(k)} \right)$$

$P(k)$ represents cluster assignment distribution for the first group and $Q(k)$ for the second. The more similar the distributions are, the closer the KL divergence will be to 0, making the system fairer.

Social Fairness

The clustering cost O for the set of cluster centroids $U = \{U_1, U_2, ..., U_k\}$ and input dataset X is defined as:

$$O\left(U, X\right) \sum_{x \in X} \min_{u \in U} \|x - u\|^2$$

For m protected groups let X_a be the samples of X that belong to protected group a, then the social fairness cost is:

$$\max_{a \in m} \frac{O\left(U, X_a\right)}{|X_a|}$$

The goal is to minimize this metric to encourage fairness.

Silhouette Difference

For $m = 2$ protected groups and k clusters, the silhouette difference is the difference in silhouette scores for each protected group. For a given group, the silhouette score is:

$$\frac{d_b - d_a}{\max\left(d_a, d_b\right)}$$

where d_b is the mean nearest-cluster distance and d_a is the mean intra-cluster distance. Silhouette score is a measure of how similar an object is to its own cluster. Silhouette scores can take values between -1 and 1

- 1: Clusters are spaced well apart from each other and clearly distinguished.
- 0: The distance between clusters is not significant.
- -1: Clusters are assigned in the wrong way.

The silhouette difference is then:

$$\frac{S_a - S_b}{2}$$

To get the silhouette difference we average the silhouette scores for each group and take the difference. We must also divide by 2 to keep it bound between -1 and 1. Negative silhouette differences indicate that group a is being treated unfairly while positive differences indicate that group b is being treated unfairly.

Individual Fairness Metrics

Proportionality: For n data points and k clusters, any n/k points are entitled to form their own cluster if there is another centre that is closer in distance for all n/k points.

Aggregate Fairness: Requires the distance of a point from its centroid to be at most α times the average distance of the points in its cluster to its centroid.

IV Mitigating Bias

Levels of Bias Mitigation

- Pre-Processing: Occurs **before** training. Requires original dataset to be modified. The algorithm is trained on a new dataset to make predictions that meet fairness requirements.
- Occurs **during** training. Requires the model or learning process itself to be modified. Without changing the original dataset, modify the algorithm to meet fairness requirements.
- Occurs **after** training. Requires the outputs of the model to be modified. The results from the modification themselves must meet fairness requirements.

Pre-Processing

- Makes changes to the training dataset.
- Original dataset X is transformed to X'.
- The algorithm A remains the same but the application of it to the transformed data results in fair clusters.
- Fairlet Decomposition [1]

Fairlet Decomposition

- The goal is to find fairlets within data that meet fairness requirements.
- Fairlet: micro-clusters that aim to have equal representation of each group.
- The centres of the fairlets are then used as a new dataset to perform clustering.
- Since the fairlets themselves are balanced, the results of the clustering is as well.

In-Processing

- Changes the model by either altering the clustering objective or the algorithm itself to output fair clustering.
- The clustering algorithm A is modified to a new algorithm to A'.
- Need to optimize between clustering cost and fairness trade-off.
- Variational Fair Clustering [2]

Variational Fair Clustering

- Introduces a penalty term based on KL divergence to encourage fairness.
- Combined objective measures the trade-off between the clustering cost and fairness.
- Aims to find clusters with specified proportions of different protected group.

Post-Processing

- Does not modify original data or algorithm.
- Use clustering algorithm A on inputs X to get clusters C C is transformed to get fair clusters C'.
- Post-processes clustering centres such that every group is represented through centres equitably.
- Making Existing Clusterings Fairer [3]: applies regular clustering and uses outputs to compute a new set of clusterings that are close to original and meet fairness requirements.

References

[1] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair clustering through fairlets," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] I. M. Ziko, J. Yuan, E. Granger, and I. B. Ayed, "Variational fair clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 11 202–11 209.

[3] I. Davidson and S. Ravi, "Making existing clusterings fairer: Algorithms, complexity results and insights," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3733–3740.