Fair Data Clustering Algorithms

"Every day I'm clustering"

Tanvi Bajpai[1] Mik Zlatin[2]

[1]Department of Computer Science
Carnegie Mellon University

[2]Department of Mathematics
Rutgers University - New Brunswick

REU-CAAR Summer 2018

Advisor: Samir Khuller

Outline

1. Motivation
   - Machine Learning
   - Fairness
2. Clustering problems
   - k-Center
3. Research
   - Empirical Motivation
   - Foundational Work
   - Our work

Outline

Machine Learning is important

Machine Learning is really important

Outline

Disparate Impact (an ideal)

- "Protected attributes, such as race and gender, should not be explicitly used in making decisions, and the decisions made should not be disproportionately different for applicants in different protected classes" (Feldmen et al.)

The Reality

- If an unprotected feature, such as height, is correlated with a protected feature, such as gender, then decisions based off of height can still be unfair

The Reality

- If an unprotected feature, such as height, is correlated with a protected feature, such as gender, then decisions based off of height can still be unfair
- In other words, it is not enough to exclude protected attributes from decision making in ML.

Outline

1. Motivation
   - Machine Learning
   - Fairness
2. Clustering problems
   - k-Center
3. Research
   - Empirical Motivation
   - Foundational Work
   - Our work

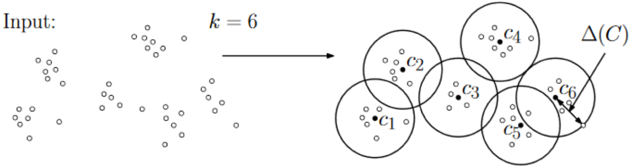
k-Center

Input: Metric space (X, d) on n points, parameter k
Output: Designate k points in the metric space as centers, and assign the remaining points to center to minimize the maximum distance between any point and its assigned center.

k-Center

Input: Metric space (X, d) on n points, parameter k
Output: Designate k points in the metric space as centers, and assign the remaining points to center to minimize the maximum distance between any point and its assigned center.

This is NP-Hard (2-approximation: Hochbaum-Shmoys; Gonzalez)



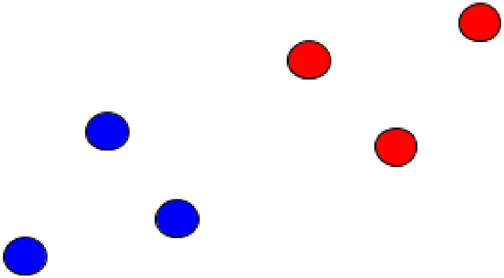
Fair k-Center

Input: Metric space (X, d) on *n colored* points (colors represent protected attributes), parameter k

Output: Designate k points in the metric space as centers, and assign the remaining points to center to minimize the maximum distance between any point and its assigned center, *while satisfying certain fairness conditions to ensure that the clusters are fair*

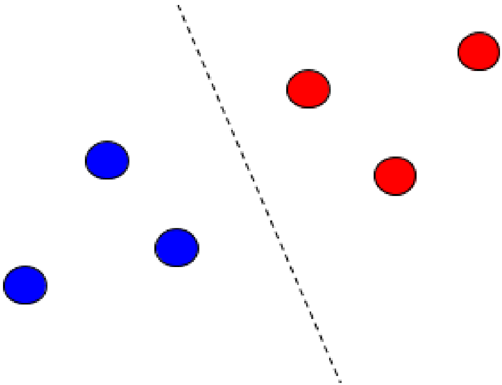
This results in different clusters

$k = 2$



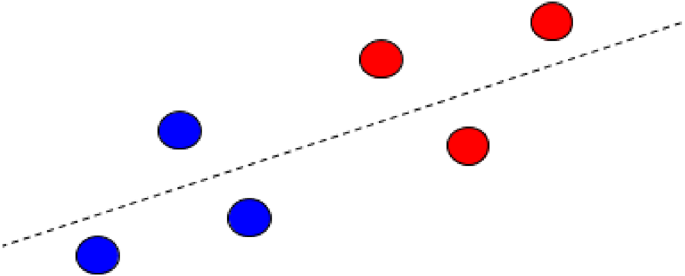
This results in different clusters

$k = 2$



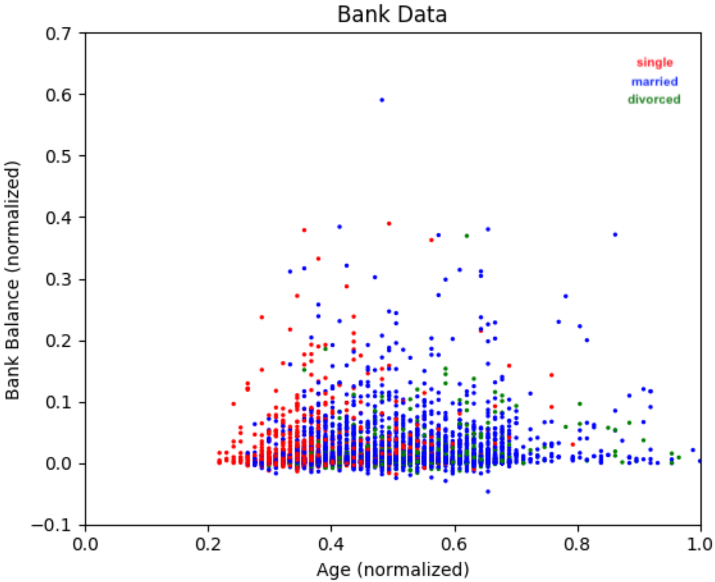
This results in different clusters

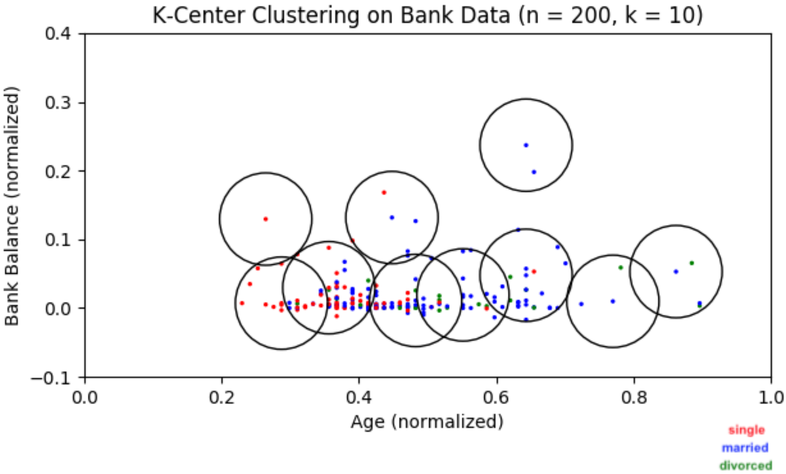
$k = 2$

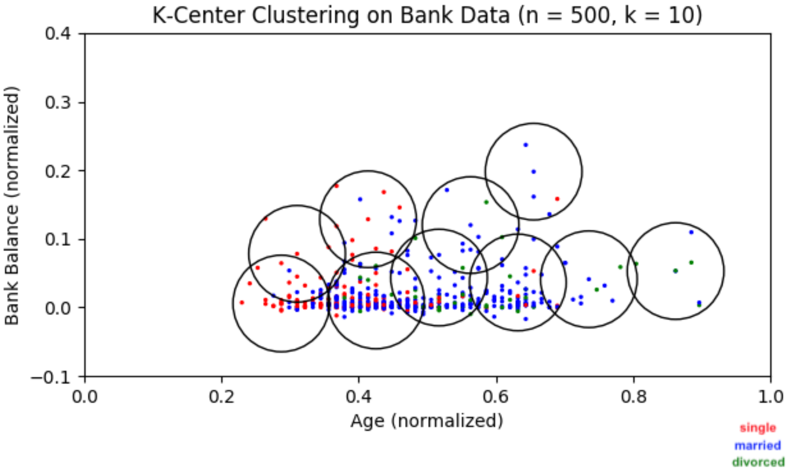


Outline

1. Motivation
   - Machine Learning
   - Fairness
2. Clustering problems
   - k-Center
3. Research
   - Empirical Motivation
   - Foundational Work
   - Our work


Bank Data
single
married
divorced
Bank Balance (normalized)
Age (normalized)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0


K-Center Clustering on Bank Data (n = 200, k = 10)
Bank Balance (normalized)
0.4
0.3
0.2
0.1
0.0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
Age (normalized)
single
married
divorced


K-Center Clustering on Bank Data (n = 500, k = 10)
Bank Balance (normalized)
0.4
0.3
0.2
0.1
0.0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
Age (normalized)
single
married
divorced

Outline

What's been done (Bercea,Khuller, and Kumar)

- Fair Clustering is NP-Hard (Even just the assignment phase of fair clustering is hard)
- Approximation Algorithms
  - Bicriteria (3,4)-approximation
  - 6-approximation

Outline

Clustering with Outliers

k-Center with Outliers

Input: Metric space (X, d) on n points, parameters k and p
Output: Cluster points into k clusters that cover at least p points (i.e., we can allow $n - p$ points to be uncovered)

Clustering with Outliers

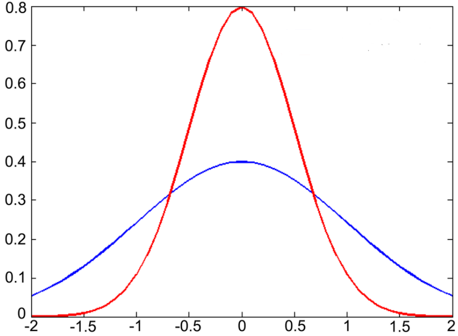
k-Center with Outliers

Input: Metric space (X, d) on n points, parameters k and p
Output: Cluster points into k clusters that cover at least p points (i.e., we can allow $n - p$ points to be uncovered)

NP-Complete; Greedy 3-approximation (Charikar, Khuller et al.); LP-based 2-approximation (Chakrabarty et al.)

Colourblind Outliers

- Generated red and blue points from different normal distributions
- Ran colourblind approximation algorithm with outliers

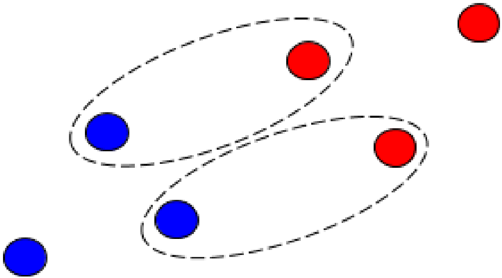


Fair Clustering with Outliers

Fair k-Center with Outliers

Input: Metric space (X, d) on n *colored* points, parameters k and p, specified fairness conditions
Output: Cluster points into k clusters that cover at least p points and satisfy the balance conditions

$k = 2$
$p = 4$



Clustering with Outliers

Fair k-Center with Outlier Clusters

Input: Metric space (X, d) on n *colored* points, parameters k and ℓ, specified fairness conditions
Output: Cluster points into k clusters that cover all n points, but allow ℓ of those k clusters to violate balance conditions

Result

We have a (3,4)-approximation to solve the outlier cluster instance, assuming the number of outlier clusters is $O(1)$.

Result

We have a (3,4)-approximation to solve the outlier cluster instance, assuming the number of outlier clusters is $O(1)$.

What if the number of outlier clusters is not $O(1)$?

Other Future Work

- Implement Fair Algorithms
- Improve (3,4)-approximation
- Develop other algorithms for outliers

Conclusion

- Thanks to Samir Khuller for being a great advisor
- Ioana Bercea for very helpful discussions
- DoD Sponsors
- Questions!