On Socially Fair Low-Rank Approximation and Column Subset Selection

Zhao Song, Simons Institute of Computing and UC Berkeley
Ali Vakilian, Toyota Technological Institute at Chicago
David P. Woodruff, Carnegie Mellon University and Google Research
Samson Zhou, Texas A&M University

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Unfortunately, real-world machine learning algorithms across a wide variety of domains have recently produced a number of undesirable outcomes from the lens of generalization:

- [BS16] noted that decision-making processes using data collected from smartphone devices reporting poor road quality could potentially underserve poorer communities with less smartphone ownership.
- [KMM15] observed that search queries for CEOs overwhelmingly returned images of white men
- [BG18] observed that facial recognition software exhibited different accuracy rates for white men compared with dark-skinned women.

Biased data or biased algorithms?

- Better training data, fair algorithms

Low-Rank Approximation

- Find rank k matrices $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{k\times d}$ that minimizes $\|UV - A\|_F$
- Finding structure among noise
- Matrix completion problem
- Closed form solution to find optimal $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{k\times d}$ that minimizes $\|UV - A\|_F$
- V is the top k right singular vectors of A
- Can be computed in polynomial time using singular value decomposition (SVD)

$$n \begin{array}{|ccccccccc|} 1 & 3 & 5 & -2 & 7 & 0 & 11 & 4 & -8 \\ 0 & 0 & -1 & 3 & 13 & 2 & 8 & 6 & 2 \\ 2 & 5 & 6 & 1 & 4 & 0 & -7 & 5 & 3 \\ 8 & 7 & 2 & 1 & -1 & -3 & -2 & -4 & -6 \\ -5 & 3 & -4 & -1 & -2 & -1 & 0 & -3 & -1 \\ 7 & 1 & 3 & 2 & 4 & 1 & 0 & 11 & 1 \end{array}$$
$$d$$

Column Subset Selection

- Find matrices $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{k\times d}$ that minimizes $\|UV - A\|_F$, where V is k columns of A
- Finding structure among noise
- Low-rank approximation variant with better interpretability
- NP-hard problem
- Can achieve $O(k)$ approximation by volume sampling or local search
- Approximation algorithms use polynomial time

Social Fairness

- Find rank k matrices $U_1 \in \mathbb{R}^{n_1\times k}$, …, $U_\ell \in \mathbb{R}^{n_\ell\times k}$, $V \in \mathbb{R}^{k\times d}$ that minimizes $\max_{i\in[\ell]} \|U_i V - A_i\|_F$
- Each matrix A_i is the dataset of a protected subpopulation
- Ensure solution is equitable to all subpopulations

$$n_1 \begin{array}{|ccccccccc|} 1 & 3 & 5 & -2 & 7 & 0 & 11 & 4 & -8 \\ 0 & 0 & -1 & 3 & 13 & 2 & 8 & 6 & 2 \end{array} A_1$$
$$d$$
$$n_2 \begin{array}{|ccccccccc|} 2 & 5 & 6 & 1 & 4 & 0 & -7 & 5 & 3 \\ 8 & 7 & 2 & 1 & -1 & -3 & -2 & -4 & -6 \\ -5 & 3 & -4 & -1 & -2 & -1 & 0 & -3 & -1 \end{array} A_2$$
$$\vdots$$
$$n_\ell \begin{array}{|ccccccccc|} 7 & 1 & 3 & 2 & 4 & 1 & 0 & 11 & 1 \end{array} A_\ell$$

Our Results

Lower Bounds

- Fair low-rank approximation is NP-hard to approximation within any constant factor
- Under the exponential time hypothesis (ETH), fair low-rank approximation requires $2^{k^{\Omega(1)}}$ time to approximate within any constant factor
- Recall: Low-rank approximation can be solved in polynomial time

Upper Bounds

- Given accuracy parameter $\varepsilon \in (0,1)$, there exists $(1+\varepsilon)$-approximation algorithm for fair low-rank approximation that uses time $\frac{1}{\varepsilon}\,\text{poly}(n)\cdot(2\ell)^{\text{poly}\left(\ell,k,\frac{1}{\varepsilon}\right)}$
- Given trade-off parameter $c \in (0,1)$, there exists $\ell^c \cdot 2^{\frac{1}{c}} \cdot O\big(k(\log\log k)(\log d)\big)$-approximation algorithm for fair low-rank approximation that uses polynomial time, but with bicriteria rank $O\left(k(\log\log k)\left(\log^2 d\right)\right)$
- There exists $O\big(k(\log\log k)(\log d)\big)$-approximation algorithm for fair column subset selection that uses polynomial time, but with bicriteria rank $O(k\log k)$
- Additional results for fair regression

Algorithm 1 Input to polynomial solver

Input: $\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(\ell)}$, $\mathbf{S}$, α
Output: Feasibility of polynomial system

1: **Polynomial variables**
2: Let $\mathbf{Y} = (\mathbf{VS}) \in \mathbb{R}^{k\times m}$ be mk variables
3: Let $\mathbf{W} = (\mathbf{VS})^\dagger \in \mathbb{R}^{m\times k}$ be mk variables
4: Let $\mathbf{R}^{(i)} \in \mathbb{R}^{k\times k}$ for each $i\in[\ell]$ be ℓk^2 variables
5: **System constraints**
6: $\mathbf{YWY} = \mathbf{Y}$, $\mathbf{WYW} = \mathbf{W}$
7: $\mathbf{A}^{(i)}\mathbf{SWR}^{(i)}$ has orthonormal columns
8: $\alpha \ge \|(\mathbf{A}^{(i)}\mathbf{SWR}^{(i)})(\mathbf{A}^{(i)}\mathbf{SWR}^{(i)})^\dagger\mathbf{A}^{(i)} - \mathbf{A}^{(i)}\|_F^2$
9: **Run polynomial system solver**
10: If feasible, output $\mathbf{V} = (\mathbf{A}^{(1)}\mathbf{SWR}^{(1)})^\dagger\mathbf{A}^{(1)}$. Otherwise, output $\perp$.

Algorithm 2 $(1+\varepsilon)$-approximation for fair low-rank approximation

Input: $\mathbf{A}^{(i)} \in \mathbb{R}^{n_i\times d}$ for all $i\in[\ell]$, rank parameter $k>0$, accuracy parameter $\varepsilon\in(0,1)$
Output: $(1+\varepsilon)$-approximation for fair LRA

1: Let α be an ℓ-approximation for the fair LRA problem
2: Let $\mathbf{S}$ be generated from a random affine embedding distribution
3: **while** Algorithm 1 on input $\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(\ell)}$, $\mathbf{S}$, and α does not return $\perp$ **do**
4: Let $\mathbf{V}$ be the output of Algorithm 1 on input $\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(\ell)}$, $\mathbf{S}$, and α
5: $\alpha \leftarrow \frac{\alpha}{1+\varepsilon}$
6: **end while**
7: Return $\mathbf{V}$

Lower Bounds

- Given vectors $v_1, \ldots v_n \in \mathbb{R}^d$, minimize the distance from these points to a $(n-1)$-dimensional subspace
- Subspace$(n-1,\infty)$ problem is NP-hard to approximate within any constant factor
- Exponential time hypothesis: The 3-SAT problem requires $2^{\Omega(n)}$ runtime.

Upper Bounds

- Theorem [BPR96]: Given a polynomial system in $x_1, \ldots x_n$ over real numbers and m polynomial constraints of degree d with coefficients at most B bits, there exists an algorithm that determines whether there exists a solution to the polynomial system in time $(md)^{O(n)}\cdot\text{poly}(B)$.
- Find "good" approximation α to fair low-rank approximation and then repeatedly decrease α by $(1+\varepsilon)$ and check polynomial system solver using
  - Use dimensionality reduction to improve polynomial solver runtime

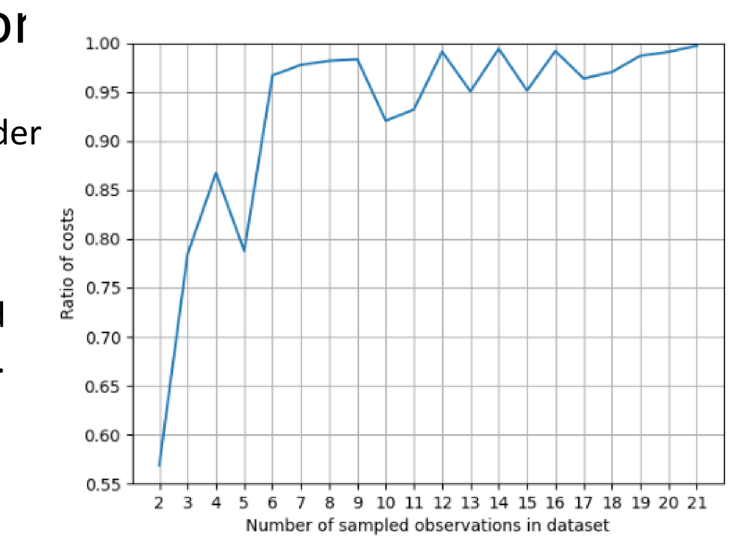
Bicriteria algorithm. $\|x\|_\infty = (1\pm\varepsilon)\|x\|_p$ for large p, so instead minimize over V:

$$\left(\sum_i \left\|A^{(i)}V^\dagger V - A^{(i)}\right\|_F^p\right)^{1/p}$$

- Use Dvoretzky's Theorem to embed into L_p

Experiments

- Credit card dataset with 30,000 observations, 23 features, e.g., previous payment statements and delays, upcoming bill statement, and whether they default. Gender used as the protected attribute
- Baseline: SVD for standard low-rank approximatioı



References

[BS16]: Solon Barocas and Andrew D Selbst. Big data's disparate impact. California law review, pages 671–732, 2016
[KMM15]: Matthew Kay, Cynthia Matuszek, and Sean A. Munson. Unequal representation and gender stereotypes in image search results for occupations, CHI 2015
[BG18]: Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification, FAT 2018
[BPM965]: Saugata Basu, Richard Pollack, and Marie-Fran¸coise Roy. On the combinatorial and algebraic complexity of quantifier elimination. J. ACM, 43(6):1002–1045, 1996

https://github.com/samsonzhou/SVWZ24