Generalized Leverage Scores: Geometric Interpretation and Applications

Bruno Ordozgoiti[1], Antonis Matakos[2], Aristides Gionis[3]

[1]Queen Mary University of London, UK
[2]Aalto University, Finland
[3]KTH Royal Institute of Technology, Sweden

ICML 2022

Context: leverage

Leverage in regression: $\min_x \|Ax - y\|_F^2$.
Useful to detect outliers.

Context: leverage

Leverage in regression: $\min_x \|Ax - y\|_F^2$.
Useful to detect outliers.

Leverage in the column subset selection problem (CSSP):

$$\min_C \|A - CC^+ A\|_F^2, \quad \text{where } C \text{ consists of } k \text{ columns of } A.$$

Useful to pick columns for the CSSP.

Use of leverage in the CSSP

Boutsidis et al. (2009).
Sample columns of A w.p. essentially proportional to leverage and refine.

$$\|A - CC^{+}A\|_F^2 \leq O(k^2 \log k)\|A - A_k\|_F^2,$$

where A_k is the best rank-k approximation.

Use of leverage in the CSSP

Papailiopoulos et al. (2014).

Sort columns by $\ell_i^{(k)}$ and pick r leading ones, so that $\sum_{i=0}^{r} \ell_i^{(k)}(A) \geq k - \epsilon$.

$$\|A - CC^{+}A\|_F^2 \leq (1 + 2\epsilon)\|A - A_k\|_F^2,$$

where A_k is the best rank-k approximation.

If $\ell_i^{(k)}$ are concentrated, few columns provide good approx.

Generalized column subset selection

Generalized column subset selection (GCSS).
We are given A, B.

$$\min_C \|B - CC^+B\|_F^2, \quad \text{where } C \text{ consists of } k \text{ columns of } A.$$

Equivalently,

$$\max_C \|CC^+B\|_F^2.$$

Leverage-score sampling not applicable: e.g. it may be that $V_k \in \ker(B)$.

Question

Can we extend leverage-based techniques for GCSS?

Generalized leverage scores and geometric bounds

Result #1

Consider a matrix $A \in \mathbb{R}^{m \times n}$ and its singular value decomposition $A = U\Sigma V^T$. Consider a column sampling matrix $S \in \mathbb{R}^{n \times r}$ and write $C = AS$. Then

$$\|CC^{+}U_k\|_F^2 \geq \|V_k^T S\|_F^2.$$

Generalized leverage scores and geometric bounds

Result #1

Consider a matrix $A \in \mathbb{R}^{m \times n}$ and its singular value decomposition $A = U \Sigma V^T$. Consider a column sampling matrix $S \in \mathbb{R}^{n \times r}$ and write $C = AS$. Then

Cosines of p. angles between U_k and C.

$$\|CC^+ U_k\|_F^2 \geq \|V_k^T S\|_F^2$$

Leverage scores.

Generalized leverage scores and geometric bounds

Result #1

Consider a matrix $A \in \mathbb{R}^{m\times n}$ and its singular value decomposition $A = U\Sigma V^T$. Consider a column sampling matrix $S \in \mathbb{R}^{n\times r}$ and write $C = AS$. Then

Cosines of p. angles between U_k and C.

$$\|CC^+U_k\|_F^2 \geq \|V_k^T S\|_F^2$$

Leverage scores.

Result #2

Consider a matrix A and its singular value decomposition $A = U\Sigma V^T \in \mathbb{R}^{m\times n}$. Consider an arbitrary index set R and a column sampling matrix $S \in \mathbb{R}^{n\times r}$ satisfying $\|V_R^T S\|_F^2 \geq |R| - \frac{\epsilon\sigma_\mu^2}{2\sigma_\omega^2}$, and write $C = AS$. Then

$$\|CC^+U_R\|_F^2 \geq \|V_R^T S\|_F^2 - \epsilon$$

where $\sigma_\omega = \max_{i\notin R} \sigma_i(A)$ and $\sigma_\mu = \min_{i\in R} \sigma_i(A)$.

Generalized leverage scores and geometric bounds

Result #1

Consider a matrix $A \in \mathbb{R}^{m \times n}$ and its singular value decomposition $A = U\Sigma V^T$.
Consider a column sampling matrix $S \in \mathbb{R}^{n \times r}$ and write $C = AS$. Then

Cosines of p. angles between U_k and C.

$$\|CC^+U_k\|_F^2 \geq \|V_k^T S\|_F^2$$

Leverage scores.

Result #2

Consider a matrix A and its singular value decomposition $A = U\Sigma V^T \in \mathbb{R}^{m \times n}$.
Consider an arbitrary index set R and a column sampling matrix $S \in \mathbb{R}^{n \times r}$ satisfying $\|V_R^T S\|_F^2 \geq |R| - \frac{\epsilon \sigma_\mu^2}{2\sigma_\omega^2}$, and write $C = AS$. Then

$$\|CC^+U_R\|_F^2 \geq \|V_R^T S\|_F^2 - \epsilon$$

Generalized leverage scores.

where $\sigma_\omega = \max_{i \notin R} \sigma_i(A)$ and $\sigma_\mu = \min_{i \in R} \sigma_i(A)$.

Application to GCSS

Application to GCSS

Deterministic generalized leverage score sampling for GCSS

Let $C = AS$, where S is the matrix output by deterministic GLS sampling. Then

$$\|CC^{+}B\|_F^2 \geq (1-\epsilon)(1-\delta)\|B\|_F^2.$$

If $\ell_i^{(k)}$ are concentrated, few columns suffice (off from Papailiopoulos et al. (2014) by $\sigma_\omega^2/\sigma_\mu^2$).

Experiments

Experimental results:

- Greedy algorithm better alternative overall.
- Our approach outperforms it in some cases.

Thanks!

References I

Boutsidis, C., Mahoney, M. W., and Drineas, P. (2009). An improved approximation algorithm for the column subset selection problem. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, page 968–977, USA. Society for Industrial and Applied Mathematics.

Papailiopoulos, D., Kyrillidis, A., and Boutsidis, C. (2014). Provable deterministic leverage score sampling. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 997–1006.