In-Database Regression in Input Sparsity Time

Rajesh Jayaram
Carnegie Mellon University

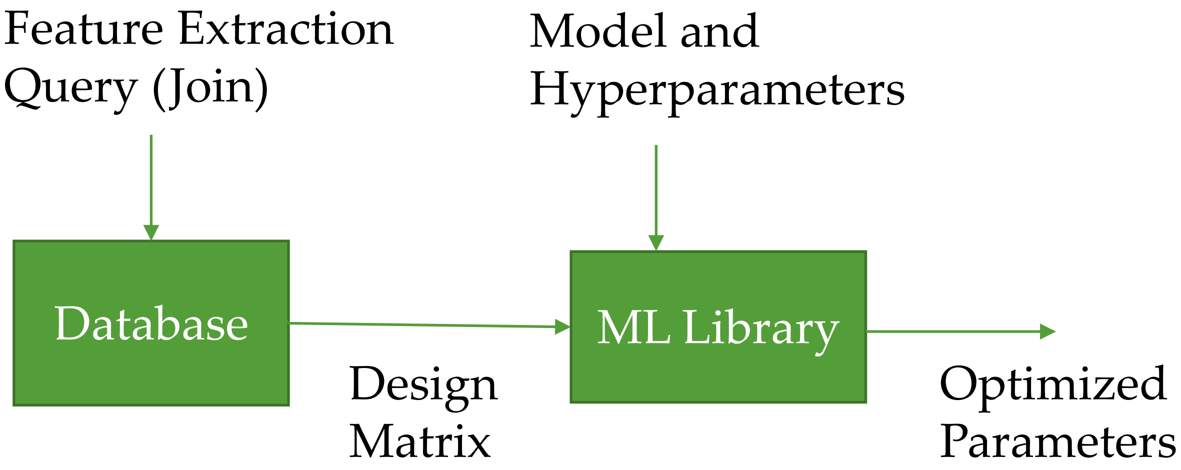
Alireza Samadian
University of Pittsburgh

David Woodruff
Carnegie Mellon University

Peng Ye
Tsinghua University

Motivation

- Often data used by data scientists comes from a database join.
- For example:
  - A grocery store which wants to predict the sales of items
- Companies and Teams working on In-Database Machine Learning:
  - RelationalAI
  - Google's Bigquery ML
- Problems with Naïve approach:
  - Join increases the size
- In-Database ML means solving the problem without explicitly computing the join itself.
  - In-Database ML can be 100 times faster in practice for linear regression.[1]



Feature Extraction Query (Join)

Model and Hyperparameters

Database

ML Library

Design Matrix

Optimized Parameters

Problem Definition

- Given a join query $J = T_1 \bowtie T_2 \bowtie \cdots \bowtie T_n$, we would like to perform linear regression on J.

Prior Works

- Using Functional Aggregation Queries (FAQ) we can solve linear regression by computing $J^T J$ [2,3].
  - For acyclic joins it has the worst-case time complexity $\tilde{O}(d^4 m\, n)$
  - For two table instances with numerical data, it can be improved:
    - $\tilde{O}(d^{\omega-1}$ n)
    - $\omega \approx$ 2.373 is matrix multiplication exponent

Our Results

- We solve the problem using subspace embedding.
- $A' \in \mathbb{R}^{k \times d}$ is a subspace embedding of $A \in \mathbb{R}^{n \times d}$ if for all $x \in \mathbb{R}^d$ we have

$$(1-\epsilon)\|Ax\|_2 \leq \|A'x\|_2 \leq (1+\epsilon)\|Ax\|_2$$

- We design a sketching algorithm that provides a subspace embedding for 2 tables join with probability at least 9/10
  - Standard sketching algorithms need all the data to be present.
  - We design a new sketching algorithm for joins.
  - We divide the join into multiple blocks (without computing the join itself) and for each block we use
    - TensorSketch techniques [4]
    - Leverage score sampling: we sample using a novel sampling scheme without computing the leverage scores for all the rows in the join explicitly.
- Runtime of our algorithm is
  - $\tilde{O}\left(\frac{1}{\epsilon^2}\left((n_1+n_2)d+d^3\right)\right)$ for dense data and $k=\tilde{O}(\frac{d^2}{\epsilon^2})$
  - $\tilde{O}\left(\frac{1}{\epsilon^2}\left(nnz(T_1)+nnz(T_2)+d^5\right)\right)$ for sparse data and $k=\tilde{O}(\frac{d^4}{\epsilon^2})$

Our Results

- We extend our subspace embedding into machine precision regression.
- For linear regression problem, machine precision regression algorithms can obtain x' such that
$$\|Ax' - b\|_2 \leq (1+\epsilon) \min_x \|Ax - b\|_2$$
- Our machine precision regression algorithm has the following runtime:
  - $\tilde{O}\left(((n_1 + n_2)d + d^3)\log(\frac{1}{\epsilon})\right)$ for dense data
  - $\tilde{O}\left((nnz(T_1) + nnz(T_2) + d^5)\log(\frac{1}{\epsilon})\right)$ for sparse data
- For general joins, we can improve FAQ based algorithm for Ridge Regression using TensorSketch [4].

Experimental Results

- We compare our 2-table algorithm on LastFM dataset [5] and MovieLens dataset [6].

Dataset	n_1	n_2	d	$T_{FAQ}(sec)$	$T_{ours}(sec)$	err
LastFM	92834	186479	6	.034	.011	0.70%
MovieLens	1000209	3883	23	.820	.088	0.66%

$$err = \frac{\|Jx' - b\|_2^2 - \|Jx - b\|_2^2}{\|Jx - b\|_2^2}$$

References

[1] Schleich, M., Olteanu, D., Abo Khamis, M., Ngo, H. Q., & Nguyen, X. (2019, June). A layered aggregate engine for analytics workloads. In Proceedings of the 2019 International Conference on Management of Data (pp. 1642-1659).

[2] Abo Khamis, M., Ngo, H. Q., & Rudra, A. (2016, June). FAQ: questions asked frequently. In Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (pp. 13-28).

[3] Abo Khamis, Mahmoud, Hung Q. Ngo, XuanLong Nguyen, Dan Olteanu, and Maximilian Schleich. "In-database learning with sparse tensors." In Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, pp. 325-340. 2018.

[4] Ahle, Thomas D., Michael Kapralov, Jakob BT Knudsen, Rasmus Pagh, Ameya Velingker, David P. Woodruff, and Amir Zandieh. "Oblivious sketching of high-degree polynomial kernels." In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 141-160. Society for Industrial and Applied Mathematics, 2020.

[5] Cantador, Iván, Peter Brusilovsky, and Tsvi Kuflik. "Second workshop on information heterogeneity and fusion in recommender systems (HetRec2011)." In Proceedings of the fifth ACM conference on Recommender systems, pp. 387-388. 2011.

[6] Harper, F. Maxwell, and Joseph A. Konstan. "The movielens datasets: History and context." Acm transactions on interactive intelligent systems (tiis) 5, no. 4 (2015): 1-19.