technische universität dortmund

Data Science Center

SFB 876 Providing Information by Resource-Constrained Data Analysis

Data and dimensionality reduction for large scale statistical data analysis

Alexander Munteanu | 01.10.2019

Massive data analysis

Data collection

- Social media
- Online services
- Consumer electronics
- Physical experiments

Massive data analysis

Data collection

- Social media
- Online services
- Consumer electronics
- Physical experiments

MASSIVE DATA

Data analysis

- There is great value in understanding the data
- Statistics, Machine Learning, Artificial Intelligence

Massive data analysis

Scalability remains a challenge

- Often not considered or only heuristically
- Crucial for any useful machine learning approach

Massive data analysis

Scalability remains a challenge

- Often not considered or only heuristically
- Crucial for any useful machine learning approach

Contribution:

- Theoretical foundations for massive data analysis
- Methods for
  - performing statistical data analysis on
  - massive data, data streams and distributed data

Massive data analysis

Scalability remains a challenge

- Often not considered or only heuristically
- Crucial for any useful machine learning approach

Contribution:

- Theoretical foundations for massive data analysis
- Methods for
  - performing statistical data analysis on
  - massive data, data streams and distributed data
- Limitations
  - Lower bounds for data reduction
  - Lower bounds memory and communication

Massive data analysis

Sketch and solve paradigm

$$
\begin{array}{ccc}
X & \xrightarrow{\quad\Pi\quad} & \Pi(X) \\
\downarrow & & \downarrow \\
f(\beta \mid X) & \approx_\varepsilon & f(\beta \mid \Pi(X))
\end{array}
$$

Massive data analysis

Sketch and solve paradigm

$$\begin{array}{ccc} X & \xrightarrow{\Pi} & \Pi(X) \\ \downarrow & & \downarrow \\ f(\beta \mid X) & \approx_\varepsilon & f(\beta \mid \Pi(X)) \end{array}$$

Canonical approach

1. Data reduction $X \to \Pi(X)$, where $|\Pi(X)| \ll |X|$
2. Time- and space efficient calculations on $\Pi(X)$
3. Approximation guarantee: solution is close to optimal

Our contributions for large or high-dimensional data

Massive data domain

1. **Bayesian regression** with Geppert, Ickstadt, Quedenfeld, and Sohler, Statistics and Computing 2017
2. **Graphical models** and **GLMs** with Molina and Kersting, AAAI 2018
3. **GLMs** with Schwiegelshohn, Sohler, and Woodruff, NeurIPS 2018
4. **Survey** on Coresets, with Schwiegelshohn, KI 2018

Our contributions for large or high-dimensional data

Massive data domain

1. **Bayesian regression** with Geppert, Ickstadt, Quedenfeld, and Sohler, Statistics and Computing 2017
2. **Graphical models** and **GLMs** with Molina and Kersting, AAAI 2018
3. **GLMs** with Schwiegelshohn, Sohler, and Woodruff, NeurIPS 2018
4. **Survey** on Coresets, with Schwiegelshohn, KI 2018

High-dimensional domain

1. Probabilistic **Smallest Enclosing Ball** with Krivosija, SoCG 2019
2. Global **Bayesian optimization** with Nayebi and Poloczek, ICML 2019
3. **Polygonal curves** (and **time-series**) with Meintrup and Rohde, NeurIPS 2019

Our contributions for large or high-dimensional data

Massive data domain

1. **Bayesian regression** with Geppert, Ickstadt, Quedenfeld, and Sohler, Statistics and Computing 2017
2. **Graphical models** and **GLMs** with Molina and Kersting, AAAI 2018
3. **GLMs** with Schwiegelshohn, Sohler, and Woodruff, NeurIPS 2018
4. **Survey** on Coresets, with Schwiegelshohn, KI 2018

High-dimensional domain

1. Probabilistic **Smallest Enclosing Ball** with Krivosija, SoCG 2019
2. Global **Bayesian optimization** with Nayebi and Poloczek, ICML 2019
3. **Polygonal curves** (and **time-series**) with Meintrup and Rohde, NeurIPS 2019

Thanks for your attention!