Praktikum: Machine Learning & Artificial Intelligence

Mohsan Jameel

(mohsan.jameel@ismll.de, Room: C037**)**

Information Systems and Machine Learning Lab (ISMLL)

University of Hildesheim, Germany

Abstract

The digitization and automation of business processes has increase the amount of digital data many folds over the past decade. Facebook alone ingest **500 - 600 terabytes** of data every day into its storage system (which is **hundreds of Petabytes**), generated through activities like photo upload, Likes, status updates etc. The **physical limitations** of a stand-alone system renders it **impractical** to be employed in **solving large scale data analytic and machine learning problem**. Scaling large scale data analytic and machine learning algorithms to multiple machine is inevitable in-order to make use of huge amount of data.

Reason to scale up

The main reasons for scaling up machine learning algorithms are

- **Large number of data instances:** the number of training examples is extremely large i.e. Facebook has 100s petabytes of data
- **High input dimensionality:** the number of features are very large, may need to partition across features
- **Model and algorithm complexity:** a number of high-accuracy algorithms are computationally expensive either rely on complex routines or nonlinear models etc.
- **Inference time constraints:** applications such as robotic navigation requires real time prediction.
- **Model selection and parameter sweeps:** Tuning hyper-parameters of learning algorithms and statistical evaluation require multiple executions of learning and inference

Objective

The main objectives of this praktikum are

- To provide opportunity to solve one of the large scale distributed machine learning problem.
- Employing Message Passing Interface to parallelize and scale across multiple machines.
  - MPI is the de-facto standard for parallel programming of distributed memory system.
- To implement
  - a method proposed in a recent research paper
  - validating the proposed results &
  - proposing improvements or generating new ideas.

Topics

Some of suggested topics are:

- Matrix Factorization (Recommender system)
  - NOMAND
  - DSGD
  - CCD
- DS-ADMM

Praktikum Organization

- **Today:** MPI tutorial, Choose topics, and make groups
- **27.04.2015:** Topic Introduction Presentation
- **11.05.2015:** First idea talk
- **Weekly:** Please discuss your progress/problems. **(MUST)**
- **Final Talk:** Date will be announced shortly. (*Will be around term end*)
- **Final Report:** Approx. 20 page report.

  (Important: your final grade includes the evaluation of this report)

- **Groups:** 2 – 3 students per topic.
- **Misc:**

Thanks

References and Further Reading

- "Scaling Up Machine Learning – parallel and distributed approaches R. Bekkerman, M. Bilenko and J. Langford , 2012 Cambridge
- **DSGD** R. Gemulla, P. J. Hass, E. Nikamp and Y. Sismanis " Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent" KDD 2011
- **CCD** Hsiang-Fu, Cho-Jui Hsieh, Si Si and Inderjit Dhillon " Scalable Coordinate Descent Approaches to Parallel Matrix Factorization for Recommender System" ICDM 2012
- **NOMAD** H. Yun, H Yu, C. Hsieh, S V N Vishwanathan and I. Dhillion " NOMAN: Non-locking, stOchastic Multi-machine algorithm for Asynchronous and Decentralized matrix completion" VLDB 2013
- **DS-ADMM** Zhi-Qin Yu, Xing-Jian Shi, Ling Yan and Wu-Jun Li "Distributed Stochastic ADMM for Matrix Factorization" CIKM 2014