Distributed Machine Learning

Instructor: Ananth Grama

Course Overview: Typical machine learning applications execute in parallel and distributed environments -- from edge/ IoT platforms to large-scale cloud based settings. These platforms pose a diverse set of challenges for learning algorithms -- edge/ IoT devices are highly constrained in their computing and communication resources, distributed systems trade-off data movement costs with communication associated with learning algorithms, parallel platforms often integrate accelerators (GPUs/ FPGAs) and pose different tradeoffs in communication and computation rates. The programming models for these disparate platforms are correspondingly diverse as well -- ranging from CUDA, MPI, and threading on parallel platforms, messaging and microservices on cloud platforms, and fine grained RPC calls on edge devices. Corresponding to these disparate hardware and systems platforms, learning algorithms must themselves be adapted to yield high performance and generalizability. In this course, we address various aspects of learning in parallel and distributed environments.

Course Format: Instructor lectures and paper presentations by students.

Student grades will be determined by:

- Paper presentation(s)
- Semester-long project
- Course participation

Course Content:

1. Basics of Parallel and Distributed Platforms
   - Hardware models
   - Software frameworks
   - Application considerations (data distribution, real-time constraints, resource constraints)
2. Overview of Machine Learning Techniques
   - Unsupervised methods
   - Supervised methods
   - Linear algebraic techniques
   - Graph-based methods
   - Deep learning
3. Optimization for Machine Learning
   - Convex optimization
   - Non-convex optimization
   - Discrete optimization
4. Learning on Edge/ IoT Platforms
   - Applications of Edge/ IoT-based learning
   - Programming models for Edge/ IoT platforms
   - Communication and computation constraints
   - Algorithms for learning
   - Software support, implementation, and performance
5. Learning in Distributed/ Cloud Environments

- Data substrates in distributed/ cloud environments
- Programming models for distributed/ cloud platforms
- Distributed algorithms for learning
- Performance tradeoffs and scalability

6. Learning in Parallel Environments
   - Algorithms for threaded and GPU-based formulations
   - Multi-GPU formulations and scaleout
   - Data, convergence, and accuracy
7. Application Case Studies
   - Important applications of different learning formulations in parallel and distributed environments.