Scalable Inference In Latent Variable Models

Amr Ahmed, Mohamed Aly, Joseph Gonzalez, Shravan Narayanamurthy, Alexander Smola

Yahoo! Research & CMU

Motivations: Data is everywhere

- Social graphs
- Click streams
- Email/Mail
- Documents
- Tuples
- Query streams

Challenges

- Inefficiencies in amount of data
- Labeling is prohibitive
- Space for supervised problems unbalanced data streams
- Limited structure for representation purposes
- Solutions are often specialized to problems

Examples:

- Clustering
- Latent Dirichlet Allocation

Three Basic Inference Problems:

- Global State Synchronization
- Key Distribution and Fault Tolerance

Global State Synchronization

- Background sync
- Copy per machine

Key Distribution and Fault Tolerance

- Dedicated server for variables
  - Select server via consistent hashing
  - Message is 0(1)/per machine
  - Communication is 0(1) per machine
  - Slow sync plus dump state per server

Architecture: LDA

- Distributed (key/value) storage on Hadoop
- Background asynchronous processing
- Single word at a time to avoid deadlocks
- Use read to have joint dictionary
- Use disk, network, CPU simultaneously