The Challenges of Bringing Machine Learning to the Masses

Alice Zheng and Sethu Raman
GraphLab Inc.
NIPS workshop on Software Engineering for Machine Learning
December 13, 2014
Self introduction

Microsoft Research

GraphLab

ML Research

Tool building
GraphLab

• Started as a research project at CMU in 2009
• Now a Seattle-based startup

GraphLab Create™
Build scalable predictive applications, fast

From inspiration... ...to production
Why build tools?

• Machine learning is transforming many lives across many industries
  • Personalized medicine, internet search, social networks, advertising, etc.
• So much potential …
• … locked in the brains of relatively few
• Building intelligent apps takes a lot of expertise
  • ML, software engineering, system integration, etc.
  • Unattainable to most
• Good tools help people do more things better
Building a predictive app

Key pains:

Noisy Space of Tools

Data scientists use a variety of tools, across different programming languages... require a lot of context-switching... affects productivity and impedes reproducibility.

Hiring Talent

35% Shortfall in data-savvy workers needed to make sense out of big data by 2018 [McKinsey 2011]

Was using 217 business rules hoping world doesn’t change

Have an inspiring idea to reinvent their business
Great sentiment, but…

- ML is hard
  - Lots of models
  - Lots of tuning knobs
  - Lots of tricks in training
  - What features to use?
  - Do I even have the right data?
  - Ok I have a prototype, now what?
Databases

• A good database is hard to build.
  • SQL, Oracle… billion dollar investments
• Why?
Database engine components

- Storage
- Storage engine
- Query execution
- Query optimizer

Diagram showing the interconnections between storage and the query components.
Database engine components

Complex but self-contained, has clean API, only changes when there’s new hardware.
Database engine components

Complex bag of tricks, no formalism, constantly changing to adapt to data, query, disk characteristics.
ML engine components

- Feature engineering
- Model definition
- Training evaluation

Bags of tricks, expert knowledge, experience, lots of trial and error
Advances in databases

• Databases became usable in the 80s
  • Relational db: reasonable abstraction, fast implementations
  • OK to have a mediocre execution plan when the computation engine became fast enough
  • The burden was shifted from the db application writer to the db platform
To advance ML platforms

• ML will be end-user friendly when the platform is clever enough to handle less-than-optimal directions from the user

• What needs to happen?
  • The complexity needs to be automated and wrapped away with neat interfaces between components
What can we do now?

• Automate where possible
• Give plenty of guidance where the user still needs to make a choice
GraphLab Create ML Toolkits

- Business Task
  - Recommender, Target, Social Match, ...
- Machine Learning Task
  - Regression, Classification, Data Matching, ...
- Algorithms & SDK
  - SVM, Matrix Factorization, LDA, ...
- Developers
- Savvy Dev & Data Sci.
- ML experts
Demo: GLC recommenders
Join the revolution!

• Research methods to make the following efficient and automatic:
  • Feature engineering
  • Model selection
  • Problem formulation
  • Model debugging
• Develop novel algorithms on top of our SDK
  • Back by scalable on-disk data structures
  • Integrated tables and graphs
• We’re hiring!