ML and Numerical Software Development

Overview, Logistics, Terminology

Organon Analytics

November 12, 2019

The goals of the course

1. ML goals:
   - ML theory, concepts, terminology
   - Building (supervised) ML models
   - ML pipelines: Technologies, tools, methods, paradigms
2. (Numerical) Software Development Goals:
   - How to develop software from scratch
   - Patterns, principles and practices
   - Data access, CRUD ops, SQL
   - Matrix computations, cache optimization, parallelism
   - Cross cutting concerns: Validation, logging, exception handling
   - Development of an industrial grade ML algorithm

Why this course? Why free?

- ML is multi-disciplinary and relatively young. University curricula are yet to catch up
- Gaps exist between what is taught and what is encountered in industry
- Show what is out there. Expose the necessary skills. Right people will respond
- We have to give so that we can take

The contents and the execution

Why this content?

- CS people don't know math, Math people don't know how to develop software
- Industry needs hybrid individuals: hybrid of Math and Computations
- Coding is easy, software development is hard
- Expect 7-10 years to master the material
- Jump-start bright/junior people for the long road ahead
- AI/ML: Tremendous potential for people with the right background&skill-set&experience

Execution method:

Part-I (Math): Lay out the mathematical foundation

Part-II (Computations): Develop ML software. Just-in-time flashbacks to the material in Part-I

Organon Analytics

- Founded in 2011
- Vision: Become a global service provider in AI/ML
- Mission: Automate building of AI/ML pipelines
- Motto: AI for everyone
- Organon AI platform: proprietary auto-ML software
- Verticals: CyberSecurity, Healthcare Analytics, E-commerce Analytics, Credit Risk Analytics
- Core competencies: Statistics, Machine Learning, Deep Learning, Big Data, Software Development
- 2 teams: Data Science & Software Development

Instructor credentials

- Ph.D. in Applied&Computational Mathematics
- A Hybrid: Math + Computations
- 20+ years of academic&industry experience in applications of Mathematics
- Founder&CTO of Organon Analytics
- Hair started to turn gray

Career paths in ML/AI

1. **Data Scientists**
   - Background(s): Statistics, Math, Engineering
   - Knowledge base: Probability, Statistics, AI/ML/DL theory and practice, Databases, SQL, Python
   - Job definition: Formulate the business problem, prototype the solution, and design the final ML pipeline
2. **ML Engineers**
   - Background: Computer Scientists
   - Knowledge base: Core computer science, AI/ML coursework, software development
   - Job definition: *Develop* AND/OR *implement* ML pipelines: Access data, prepare data, run algorithms, deploy models, troubleshoot, develop if necessary, maintain

Course Logistics

- When: Tuesdays and Thursdays, 18:00-20:00
- Office hours: Wednesdays, 14:00-16:00, Organon Office Etiler
- Where: Bogazici University, Computer Science Building, A6
- E-mail: mlcourse@organonanalytics.com
- Web-page: http://www.organonanalytics.com/en/career-detail/free-course-for-ml-numerical-software-development/54
- GitHub address: https://github.com/Organon-Analytics
- Start Date: 12 November 2019
- Duration: 10 weeks, 20 days
- Calendar: Slight modifications possible

Some reminders

- Attendance, assignments are NOT compulsory
- All artifacts (code, slides, documents) will be posted to the GitHub address. Check it out regularly
- Ask anything and everything. ML is a broad subject
- You are welcome at the office hours
- Use course e-mail
- Assignments: Reading assignments, problems, coding assignments
- Push/challenge the instructor
- Try to get the most out of the course: Study!

Computation resources

- A laptop for coding (Nice to have: multi-core, RAM$\geq$8GB)
- Visual Studio Community Edition
- Math library: Intel MKL library
- Logging: log4net
- Database: PostgreSQL
- Unit Testing: NUnit
- More as the course unfolds

Book references

- Artificial Intelligence: A Modern Approach, 3rd edition
- The Elements of Statistical Learning, 2nd Edition
- Deep Learning, by Ian Goodfellow, et al.
- Introduction to Probability Models, by Sheldon Ross
- Pattern Classification, 2nd edition
- Neural Networks for Pattern Recognition, by Christopher Bishop

Syllabus

Day-1: Overview, Logistics, Terminology

Day-2: Probability Theory

Day-3: Statistics-I

Day-4: Statistics-II

Day-5: Data&Databases

Day-6: Machine Learning-I

Day-7: Machine Learning-II

Day-8: Gradient Boosting Machines (GBM) algorithm

Day-[9, 20]: Development of GBM algorithm

Probability and Information Theory

- ML is interested in data generating processes (DGP)
- DGPs of ML are unknown
- ML deals with uncertain/random events and quantities
- Probability theory is the tool to express uncertainty and make computations
- Important concepts&techniques:

  Sample space, events
  Conditional probability
  Random variables
  Expectation, conditional expectation
  Bayes' rule. Incorporating prior knowledge
  Distributions, densities
  Law of Large Numbers, Central Limit Theorem
  Pseudo Random Number Generation and MC simulations

Statistics: The original data analysis discipline

- Statistics is application of probability theory to finite data
- Primary goals: Estimation, Understanding, Prediction
- Old-fashioned cousin of Machine Learning
- Statistics $\sim$ Small sample, parametric, linear, emphasis on understanding, works on structured data
- ML $\sim$ Big Data, non-parametric, non-linear, emphasis on prediction, works on structured/non-structured data
- Important concepts& techniques:

  Statistics: Estimation, Hypothesis Testing, Prediction
  Important statistics: location, dispersion, asymmetry, tail
  Hypothesis Testing: Heart of Inductive Reasoning
  Types of errors: Type-I, Type-II
  Measures of statistical dependency
  A taxonomy of (statistical) data: Categorical, ordinal, scale
  Maximum Likelihood Estimation: The go-to method of Statistical Estimation
  Multivariate Linear Regression: Grandpa of all statistical and ML algorithms

Data: The fuel that drives the AI/ML revolution

- The dawn of a new age: The data as the proxy for the truth
- Data in the wild: Structured, semi-structured, non-structured
- Databases: SQL, No-SQL. Use cases, tools
- Distributed filesystems and log data
- Transactional processing vs. batch processing
- What is Big Data? Use cases, tools
- Data Quality: Hated kid in the family. Poisons everything if not treated

Machine Learning: Building (supervised) ML models

- ML tasks: Unsupervised, supervised, and reinforcement learning
- Supervised Learning: Classification, Regression
- Training, validation, and testing of a supervised ML model: A recipe
- Generalization and regularization. Bias-Variance trade-off
- Hyper-parameters and their optimization
- Performance of a model
- Transparency and accuracy trade-off
- (Supervised) Algorithms you need: GAMs, GBMs, ANNs

Machine Learning: Building (supervised) ML pipelines

(Supervised) ML pipeline:

1. Access and integrate heterogeneous data
2. Measure data quality, take corrective actions
3. Prepare data (sampling, feature extraction, etc.)
4. Run ML algorithm, build and save the model
5. Deploy ML model
6. Document the results
7. Monitor the model

Methods, tools, technologies, examples

Gradient Boosting Machines: Current champion

- Gradient descent algorithm
- Derivation of GBM
- Insights behind the success of GBM: Iterative, incremental, local
- GBM pseudo-code
- Building a GBM model on an example dataset. Examination of the resulting model
- References: Articles, current software implementations

Numerical software development

- Objective: Build ML software from scratch. Observe its evolution
- Software development process
    - *i* Requirements gathering, analysis, design
    - *ii* Waterfall versus Agile
    - *iii* Software as a living organism
- Architectural pattern: Layered architecture
- Work at Application Layer: Define shell services. Build the scaffold
- Work at Domain(Business) Layer: Primary data and algorithm classes
- Matrix Computations: Linear Algebra. BLAS and LAPACK
- The journey of data from disk to the registers. Cache optimization

Numerical software development

- Parallelization: Data parallelism, task parallelism. Tools, techniques, guidelines
- Work at Data Access Layer: R/W data in databases. Techniques, guidelines, rules-of-thumbs
- Work at Application Layer: Fill out the shell classes
- Cross cutting concerns: Validation, Logging, Exception handling
- Testing: Unit Testing, Integration Testing
- Performance Profiling and Optimization

AI, ML, DL: definitions

Artificial Intelligence: The umbrella term for mimicking human intelligence by programs (software as mind) and machines (robots as body)

- Reasoning, problem solving (e.g. solve Rubik's cube with a robot hand)
- Knowledge representation (objects, relationships, contents): Databases, ontologies. Content based indexing and information retrieval (Think of ontologies in an e-commerce site or medical dictionaries)
- Expert systems: Domain specific rule-based systems (e.g. Medicine, Marketing)
- Learning: Unsupervised, supervised, reinforcement

Artificial Intelligence cont'd

- Natural language processing: Text understanding, question answering, machine translation (Alexa, Siri, Google translator)
- Perception: Object recognition, speech recognition, facial recognition, emotion recognition
- Robotics: Motion and manipulation (e.g. Boston Dynamics)
- Human/computer interfaces (e.g. NeuraLink)

Machine Learning

Machine Learning: A subset of AI

- Focus on learning. Specifically, learning from data (Inductive Learning, from specifics to general, from examples to the model)
- Three tasks:
    1. *Unsupervised* learning: Learning with no guidance (Segmentation, anomaly detection)
    2. *Supervised learning*: Learning with guidance (Computer Vision, Speech recognition, machine translation, Credit Risk Scoring, Recommendation Systems, Demand Forecasting)
    3. *Reinforcement learning*: Learning to choose the best action to maximize a delayed reward (Deep Blue, Alpha Zero, driverless transportation, marketing automation)

Deep Learning

A subset of ML with focus on (Deep) Artificial Neural Networks for solving the ML problems

- *Computer Vision*: Convolutional Neural Networks (object recognition, object detection, facial recognition, emotion detection, video prediction, etc.)
- *Speech understanding*: Speech recognition, speech-to-text
- *Natural Language processing*: Text understanding, machine translation (Alexa, Siri, Google translator)
- *Reinforcement learning*: Playing games, recommendation systems (Alpha Zero, StarCraft, driverless cars, YouTube rec. system)

The purposes of ML and DL are the same. Algorithms differ.

Artificial Intelligence is inter-disciplinary

- Philosophy: Rationalism, deduction/induction, dualism, determinism, logical positivism, mind as a machine
- Mathematics: Logic, Probability, Statistics, Algebra, Analysis
- Computer Science: Hardware, software, data structures & algorithms
- Economics: Decision theory, rational behaviour as utility maximization
- Neuroscience: Nervous system as a model for artificial neural networks
- Control Theory and Robotics
- Linguistics

Drivers of the AI revolution

1. **Hardware**:
    - Storage technologies improved. Costs dropped
    - Sensor technologies (cameras, IoTs, etc.) improved. Costs dropped
    - Mobile communications: Improved infrastructure, interactions, data collection
    - Internet infrastructure investment in dotcom boom (late 90's)
    - Multi-core chips
    - GP-GPU: Graphic cards used for general computations (e.g. Nvidia)
    - FPGAs: Field programmable gate arrays (e.g. Google TPUs)
    - Cloud computing (Amazon, Azure, Google, IBM)
    - HPC: Exascale computing has arrived

Drivers of the AI revolution: cont'd

2 **Software**:
- Distributed processing: Google File System, HDFS, Map-Reduce
- Big Data eco-system
- Python as a language and as an ecosystem
- Deep Learning frameworks: TensorFlow, PyTorch
- Open source libraries

3 **Research**:
- Support Vector Machines: 1990s
- Boosting algorithms: 2000s
- Deep Learning algorithms: 2010s

4 **Money**:
- Companies: Google, IBM, Apple, Amazon, Microsoft, Nvidia
- Governments: USA, China, EU, Israel

What Next?: From lab to production

- Quantum Computing
- Switch from silicone to graphene
- Exascale computing
- Automate blue-collar jobs (transportation jobs, call center jobs, delivery jobs, etc.)
- Automate white-collar jobs (call center jobs, front-office jobs, concierge jobs, etc.)
- Machine translation
- Driverless transportation
- Drones&robots warfare, cyber warfare, surveillance state
- Preventive medicine, drug discovery
- Entertainment industry (VR, AR)
- Marketing automation