Algorithmic Machine Learning

Introduction to the Course

Pietro Michiardi

Eurecom

Overview

Objectives of the Course

- **Gain hands-on experience on real-life Data Science projects**
- **Use knowledge from "theory" courses and put them into practice**
- **Use knowledge from "systems" courses**
- **Develop a methodology to address challenges such as:**
  - Data preparation
  - Data exploration
  - Algorithm / model selection
  - Experimental validation and evaluation

Notebooks, not Lectures!

- **Essentially, there will be no traditional lectures**
  - Introduction to machine learning and advanced statistical inference
  - Distributed systems and cloud computing
  - Basic computer science skills are necessary
- **Notebooks**
  - A self-contained studying and development environment
  - Contains text, reference material, code, questions, graphs
  - Each Data Science project will be your own project!
- **Publish your Notebooks!**
  - Create a GitHub account and push your Notebooks there
  - High-visibility of your own Data Science projects
  - This is a sort of on-line CV

Notebooks Content – Schedule (1)

- **Lab 1 [3/8/2017]**
  - Introductory laboratory: getting familiar with Notebooks, Python, Numpy, Pandas, PySpark, Data Frames and more
- **Lab 2/3 [3/15/2017 - 3/22/2017]**
  - Recommender Algorithms Project: work with real data from an Internet music streaming service, and recommend new music to users
- **Lab 4/5 [3/29/2017 - 4/5/2017]**
  - Regression Algorithm Project: using random forests to predict airplane delays, using real data from the U.S. DoT

Notebooks Content – Schedule (2)

- **Lab 5/6 [4/12/2017 - 4/26/2017]**
  - Estimating Financial Risk through Monte Carlo Simulation
- **Lab 8/9 [5/3/2017 - 5/10/2017]**
  - Clustering Algorithms Project: Anomaly Detection in Network Traffic with k-means Clustering
- **Industrial Lab [5/24/2017 - 5/31/2017]**
  - Industrial Project from SAFRAN Analytics
- **Lab 10/11 [6/7/2017]**
  - Analyzing Neuro-imaging Data with Thunder

Industrial Notebooks

- **Great opportunity to be exposed to real industrial problems**
  - People from industry supervise the laboratory
  - Main goal: hiring!
- **SAFRAN Analytics**
  - `http://www.safran-group.com/`
  - Distribute goodies
  - Select best student(s) to participate to a SAFRAN event

How to Be a Successful Student (1)

- **Do not underestimate this course!**
  - Be independent and dare to explore and expand your Notebooks
  - Study or revise the theory: students are assumed to be comfortable with machine learning material, and to follow advanced statistical inference courses
  - Follow links on the Notebooks. They contain reference material and research papers that: *i)* provide the necessary background; *ii)* offer starting point to improve algorithms
- **Is this a course about algorithm design?**
  - Sort of: in many cases, Notebooks rely on standard libraries that offer a variety of machine learning algorithms implemented in an efficient way.
  - Notebooks will illustrate the main algorithmic concepts behind a selection of tools available in such libraries
  - Advanced (and optional) approaches to those proposed in the Notebooks are more than welcome!

How to Be a Successful Student (2)

- **Does this course make me a Data Scientist?**
  - Sort of: it is the whole track that provides student with the necessary knowledge to start a Data Science career. This course aims at “learning the hard way” and put into practice theoretical concepts
- **Do I need to know how to program?**
  - Yes, and this is mandatory
  - We will focus on Python, but knowledge of additional languages is definitely a plus

Grading

- **Grading the laboratories / projects**
  - Two-person groups are considered the norm
  - Each Notebook/project is evaluated and graded
  - Grading metrics
    - Answer to Notebooks questions: this allows you to arrive at 10/20
    - Depth of answers
    - Originality of answers and approaches
    - Additional points to innovative material in each Notebook
- **Final exam**
  - Depends on class behavior and performance
  - Example: given a real world Data Science problem, outline an approach to address it, including:
    - Relevant data exploration questions
    - Relevant data cleaning warnings
    - Model and algorithm selection
    - Performance considerations
    - Model validation

Useful References

- *"An Introduction to Statistical Learning"*, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

  Available for download:
  `http://www-bcf.usc.edu/~gareth/ISL/`
- *"Advanced Analytics with Spark"*, by Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills

  Available here:
  `http://shop.oreilly.com/product/0636920035091.do`, also available in the Library
- *"Understanding Machine Learning: From Theory to Algorithms"*, by Shai Shalev-Shwartz and Shai Ben-David

  Available for download: `http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/`

Infrastructure

EURECOM Cloud Computing Platform

- **Private datacenter, hosted at Eurecom**
  - Few hundres of server slots
  - Pretty immutable network configuration
  - No service level agreements
- **Cloud Computing Platform**
  - Hybrid system: VM-based and container-based
  - $O(1000)$ cores, $O(2TB)$ RAM, $O(200TB)$ storage
  - No service level agreements

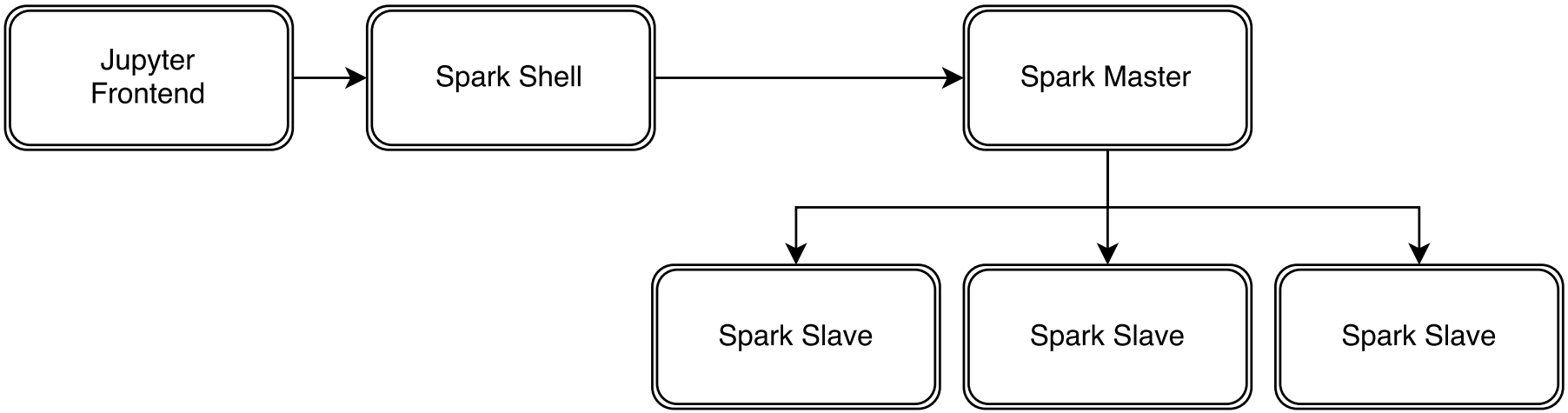
Zoe Analytics

- **Towards datacenter operating systems**
  - Cluster scheduler, in the family of Borg, Mesos, and K8s
  - Geared toward Analytics applications
  - Scheduler and Resource allocator
  - Based on Docker containers
- **Eurecom Open Source project**
  - You can contribute!
  - A lot of interest from many companies
  - A platform for research

Jupyter Notebooks

The Jupyter Notebook is a web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

Zoe Jupyter Applications



Working in the Lab

- **Clone or Fork the AML-course repository**
- **Working on your Notebook project**
  - Upload your Notebook to the Zoe Jupyter application
  - Work on your Notebook
  - Download your Notebook as an iPython notebook
    - ★ This allows you to continue to work on your project during subsequent laboratory sessions, or eventually to work from home on a local installation
    - ★ It is strongly suggested to use GitHub!!
- **Submitting your Notebook for evaluation**
  - Download your Notebook as an html page
    - ★ Be careful! You need to save after you execute all cells!
  - Send by email the html version of the Notebook