Machine Learning Software Overview

Machine Learning (41204)

Robert E. McCulloch / Mladen Kolar
Vinh Luong / Juan Yrigoyen

Fall 2015

Invitation to Two Parties

One of our key goals in this course is to give you a good deal of exposure to, and ultimately confident working knowledge of, some of the latest and best open-source software that either:

- directly implements complex Machine Learning algorithms; or
- indirectly facilitates obtaining, using and developing such Machine Learning software.

Given such a goal, we have decided to offer course content in not only one, but **two programming languages**, namely **R** and **Python**, the two leading, default go-to open-source software ecosystems for intensive, large-scale scientific computation in general and Machine Learning algorithms in particular.

You have the OPTION of doing your programming work in either language. You are certainly not required to work with both, but are highly encouraged to if you have the time. We may give out some "diligence bonus" when you use both!

Invitation to Two Parties (cont'd.)

In subsequent slides, we briefly discuss the key components of the R and Python ecosystems. You can notice a lot of parallelism between the two.

- **Booth Analytics Club** has compiled a publicly-available **Learning Resources Catalog**, including many online courses and software packages in R and Python. You may check it out every now and then.

The R Machine Learning Ecosystem

In R, we will work with the following:

- Programming Language: **R**
- Integrated Development Environment (IDE): **RStudio**
- Dynamic Documents: **R Markdown**, already embedded in RStudio
- Default Package Repository: **Comprehensive R Archive Network (CRAN)**
- Pre-eminent Machine Learning Package: **Caret**, a highly optimized wrapper around about 200 Machine Learning algorithms

Extras include:

- Revolution Analytics' **doParallel** package for multi-core parallel computation

The Python Machine Learning Ecosystem

In Python, we will cover the following:

- Programming Language: **Python**, particularly through the **Anaconda** pre-packaged distribution by Continuum Analytics
- Integrated Development Environment (IDE): **PyCharm**
- Dynamic Documents: **IPython Notebook**
- Default Package Repository: **Python Package Index (PyPI)**
- Pre-eminent Machine Learning Package: **SciKit-Learn**, which gives you hundreds of Machine Learning algorithms as well as highly efficient data processing tools

Extras include:

- DeepLearning.net's **Theano** package for fast numerical computation, especially beneficial for those of you whose machine comes with an NVIDIA graphics card

Additional Supporting Software

Additionally, throughout the course you will also come across certain supporting tools that are useful for general open-source software development and distribution:

- **GitHub**: a web-based, revision-controlled, code hosting repository; our course materials will be distributed solely through our **GitHub course repository**
  - GitHub is based on version-control software **Git**
  - **SourceTree** is one of the best apps for managing Git and GitHub repositories
- **CygWin** ***(for Windows users only)***: a command-line terminal to run Unix-style commands on Windows