Reproducible Research Pattern Recognition and Machine Learning

Dr. André Anjos
Idiap Research Institute, Switzerland
http://andreanjos.org
andre.anjos@idiap.ch

Abstract

This is a course on Reproducible Research (RR) [1] for research engineers working with software applications in Pattern Recognition (PR) and Machine Learning (ML) [2]. It motivates and explains concepts behind RR, an increasing trend in scientific publications in this niche, its implications and tools for implementing it on an individual or group levels. It is a hands-on course in the sense students will be required to create their own workflows for selected problems in ML and PR. By the end of this course, students should understand the basic concepts of reproducibility, its importance on their daily practice and how to achieve it with freely available tools and environments.

1 Introduction

One of the key aspects of modern technological research lies on the use of personal computers (PCs) either for the simulation of known phenomena or for the evaluation of data collected from natural observations. Mashups of these data, organized in tables and figures are attached to textual descriptions leading to scientific publications. In the current practice, data sets, code and actionable software leading to those results are excluded upon recording and preservation of articles. This panorama slows down potential scientific development in at least two major aspects: (1) re-using ideas from different sources normally implies on the re-development of software leading to original results and (2) the reviewing process of candidate ideas is based on trust rather than on hard, verifiable evidence that can be thoroughly analyzed [3].

In this course, we introduce the concept of "Reproducible Research" (RR) [1], a term that labels scientific work that provides not only a description of the effort leading to stated conclusions, but points to data, software and instructions that allows readers to reproduce author results locally, with all required details and in a very short time. The promised gains of RR are incredible [4], but it does not come without a cost: in order to boost reproducibility, researchers now need to (re-)organize themselves so as to always be doing RR. This course will walk students through tools and practical exercises in order to implement RR on their daily activities.

Finally, we'll introduce students to the BEAT Platform [5, 6]: a web-based system for Reproducible Research. BEAT provides an all-in-one experience in RR: tools to graphically create workflows, write algorithms, run, log and search for results in a socially interactive way. All complexity of RR and computation is hidden behind an easy-to-use graphical web interface. Experimentation designed inside the platform can be easily transmitted and reproduced in a matter of seconds.

2 Tentative Topics and Outline

The length of each topic will depend on student motivation and discussions. The minimum course time is ∼10 hours. If required, the course can be given in two days (each with, at least, 5 hours of course time).

1. Introduction and Some Programming Background
   - (a) The need for reproducibility
   - (b) Database and protocols: how to do it
   - (c) Tools for RR in the wild
2. Python and Bob [7, 8]:
   - (a) Building database packages (encoding protocols)
   - (b) Using Python and Bob for basic Machine Learning
   - (c) Putting all together
3. Going social [6]:
   - (a) The requirement for a web-based RR tool
   - (b) The BEAT platform
   - (c) Adapting your workflow to the platform
   - (d) From running experiments to publication preparation using only a web-browser

3 Course Requirements

Participants shall understand the basics of Pattern Recognition, Machine Learning and programming. Knowing the Python programming language is a plus. Here is a list of resources which can be interesting:

- Theoretical:
  - Machine Learning Basics [2]
- Practical:
  - Dive into Python (free tutorial) [9]
  - Numerical and Scientific Programming in Python [10, 11]
  - Bob framework for Signal Processing, Machine Learning and Biometrics [12]

3.1 Required Material

Each participant must bring their own laptop, running VirtualBox [13]. I'll provide a virtual image via USB keys in which students will be able to test/implement all exercises.

The room for the course should be equipped with a (normal) digital projector, with XGA resolution (1024x768) or superior and a wireless internet connection.

References

[1] `http://reproducibleresearch.net/`, 2015.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing ed., October 2007.

[3] B. R. Jasny, G. Chin, L. Chong, and S. Vignieri, "Again, and again, and again...," *Science*, vol. 334, no. 6060, p. 1225, 2011.

[4] P. Vandewalle, "Code sharing is associated with research impact in image processing," *IEEE Computing in Science and Engineering*, vol. 14, pp. 42–47, July 2012.

[5] `https://www.beat-eu.org/`, 2015.

[6] `https://www.beat-eu.org/pltatform/`, 2015.

[7] A. Anjos, L. El Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *Proceedings of the ACM Multimedia Conference*, Oct. 2012.

[8] `https://www.idiap.ch/software/bob/`, 2015.

[9] `http://www.diveintopython.net/toc/index.html`, 2015.

[10] http://www.numpy.org/, 2015.

[11] http://www.scipy.org/, 2015.

[12] https://github.com/idiap/bob/wiki/Bob-Starter-Course, 2015.

[13] https://www.virtualbox.org/, 2015.