The PLearn machine learning library

by Pascal Vincent, Nicolas Chapados

and many other contributors

Project URL: www.plearn.org **Open Source license used:** BSD

PLearn is a large C++ machine learning library and accompanying set of tools (mostly python-based) that has been Open Source and under active development since 1999. While it hasn't been much advertised, it has been used extensively for the research and development of novel algorithms at Yoshua Bengio's LISA Lab at University of Montreal, and has been incorporated and deployed in several large scale industrial applications. It has also served as the core software foundation for a successful startup company that regularly contributes to its public code base.

PLearn offers the following capabilities:

- **At the low C++ infrastructure level:** matrix and vector classes with interface to lapack for linear algebra; automatic memory management through smart pointers; powerful object serialization/deserialization mechanism in both human-readable and efficient binary format, easy remote method call mechanism and support for MPI parallelization.
- **For C++ machine-learning algorithm development:** base classes for *virtual datasets* (not limited to fitting in memory), dataset *splitters* (to generate train/test, k-fold, or sequential validation *"splits"*) and generic learning algorithms (*learners*); easy specification of functions of matrices with automatic gradient backpropagation; efficient gradient based optimization algorithms.
- **For machine-learning experiments:** easy specification of complex experiments through *python scripts*; possibility to *chain learners* to include automated preprocessing steps; powerful *testing framework* (cross-validation, etc...); powerful *hyper-parameter optimization framework*.
- **For deployment or embedding in applications:** can be called upon as a C++ library, a command-line tool, or a standalone computation server; *python interface* allows rapid development of python-based GUIs.
- **PLearn's Python-based tools:** powerful compilation system, PLearn IDE, graphical display of experimental results, plotting of decision surfaces, density landscapes, etc...
- **Library contains both classical and cutting-edge machine learning algorithms** for classification, regression, density estimation, dimensionality reduction, clustering and statistical language modeling. E.g.: k-NN, Parzen Windows, Gaussian Mixtures, k-means Clustering, Neural Networks (flexible architecture), PCA, LLE, Isomap, spectral clustering, linear and kernel regression, regression tree, AdaBoost, Hinton's Restricted Boltzmann Machine and Deep Belief Networks ... It also has an easy interface to Torch (for SVM, ...).

In addition to explaining and demonstrating some of the above capabilities of the PLearn platform, we will talk about the experience of designing a large machine learning library, and stretching its limits beyond the initially foreseen uses over 7 years... We will explain how this has (re)shaped our view on the central points that need to be payed careful attention to, in such a design (and what we would now do differently). We will in particular discuss how the pattern of allowing the flexible *combination* of basic building blocks (or "modules") popped up in many areas, and the different ways that have been used to address this, with their pros and cons in different machine learning settings.