dislib

The Distributed Computing Library (dislib) is a Python library built on top of PyCOMPSs that provides distributed mathematical and machine learning algorithms through an easy-to-use interface. Some of the available algorithms are: K-means, DBSCAN, support vector machines, and random forests.

dislib abstracts Python developers from all the parallelization details, and allows them to build large-scale machine learning workflows in a completely sequential and effortless manner. The main concepts around dislib are:

- **Distributed arrays**: A built-in 2-dimensional array that can be operated in parallel, and that is used as the main input for the different algorithms. Distributed arrays store samples and labels in a distributed way that works as a regular Python object from the user point of view.
- **Data handling**: Methods for loading data from files in common formats, such as CSV and LibSVM.
- **Unified interface**: scikit-learn inspired interface for the different algorithms (i.e., fit, predict, etc.). This makes dislib's interface easy to learn to users already familiar with scikit-learn, and allows a smooth transition of existing codes from scikit-learn to dislib.

dislib can be easily integrated into any existing PyCOMPSs application, and can run in any computing platform supported by PyCOMPSs, such as clouds, clusters, and supercomputers. The library can be installed locally via pip with 'pip3 install dislib', and is available at MareNostrum 4 using 'module load dislib'.