Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Distributed machine learning with dislib

Javier Álvarez, Rosa M. Badia, Javier Conejero, Jorge Ejarque, Daniele Lezzi, Francesc Lordan, Nihad Mammadli, Cristian Ramon-Cortes, Salvi Solà

PATC 2020

Barcelona

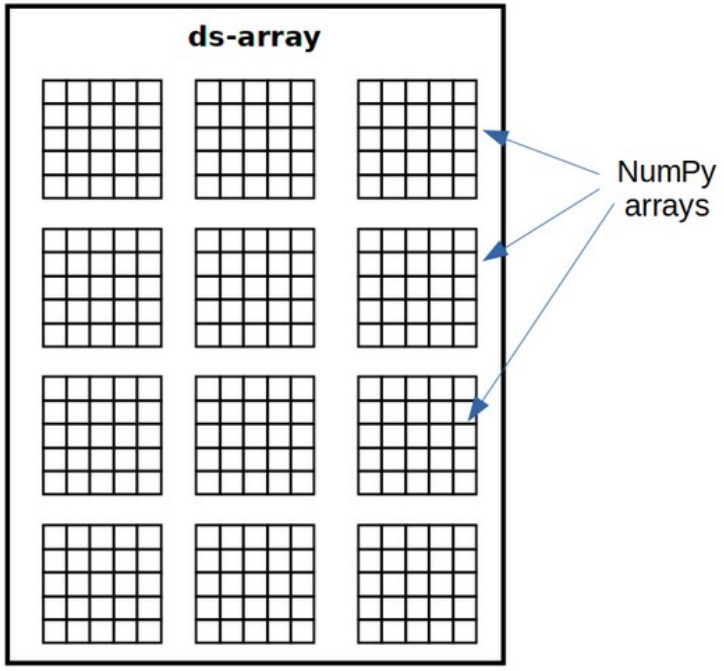
28 Jan 2020

dislib | distributed computing library

- Built on top of PyCOMPSs
- Distributed array
  - similar to NumPy
- Distributed machine learning models
  - similar to scikit-learn

Distributed arrays

- 2-dimensional structure (i.e., matrix)
  - Divided in blocks (NumPy arrays)
- Work as a regular Python object
  - But not stored in local memory!
- Internally parallelized with PyCOMPSs:
  - Loading data (e.g., from a text file)
  - Indexing (e.g., x[3], x[5:10])
  - Operators (e.g., x.min(), x.transpose())

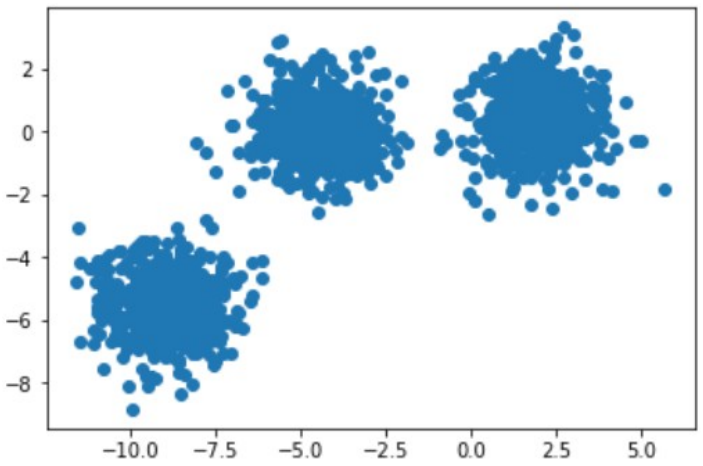


Machine learning basics

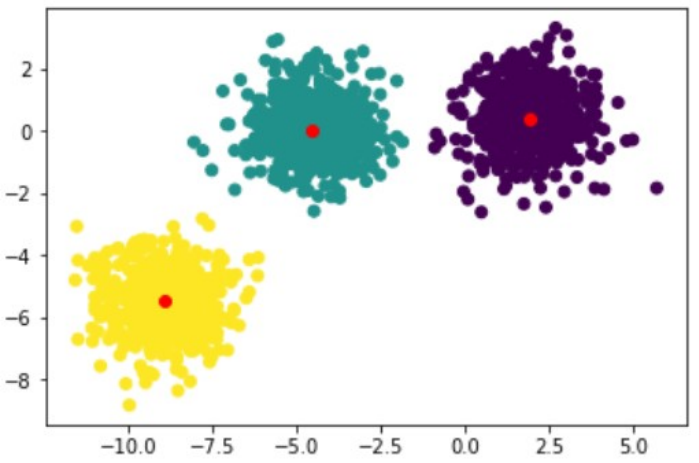
- Unsupervised:
  - Find unknown patterns in (unlabeled) data
  - Example: clustering

- Supervised:
  - Learn a decision function from labeled data
  - Example: classification

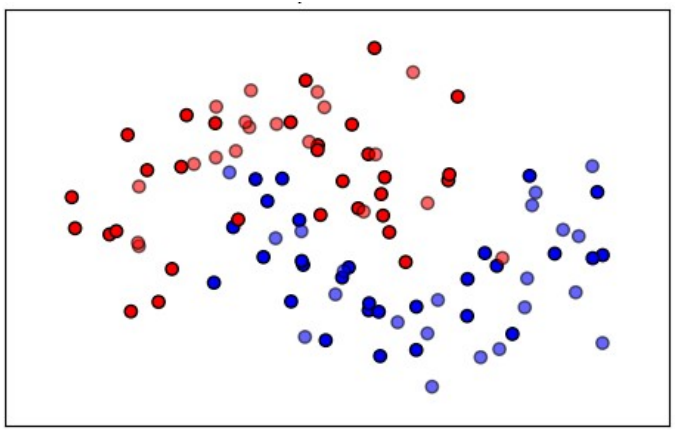
Clustering



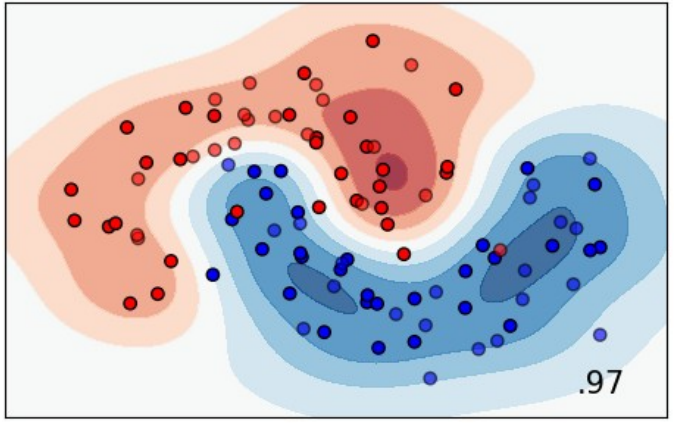
Unlabeled data



Classification



Labeled data



Estimators

- Based on scikit-learn
- Estimator = anything that learns from data (labeled or unlabeled)
- Two main methods:
  - `fit` → learns something from data (e.g., a decision function)
  - `predict` → provides new information based on a fitted model (e.g., labels data based on the computed decision function)

Typical workflow

1. Read input data from file/s

```
x = load_txt_file("train.csv", (10, 780))
x_test = load_txt_file("test.csv", (10, 780))
```

2. Instantiate estimator with parameters

```
kmeans = KMeans(n_clusters=10)
```

3. Fit estimator with training data

```
kmeans.fit(x)
```

4. Make predictions on test data

```
kmeans.predict(x_test)
```

Supported algorithms

- Supervised:
  - Support vector machines
  - Random forests
  - Linear regression
  - ALS
- Unsupervised:
  - K-means
  - DBSCAN
  - K-nearest neighbors
  - Gaussian mixtures
  - PCA

dislib notebook

```
git clone https://github.com/bsc-wdc/dislib.git
cd dislib
pycompss init -i compss/compss-tutorial:2.6
pycompss jupyter notebooks
```