Olivier Grisel

Ingénieur Inria

Membre du comité technique de scikit-learn

Calcul distribué & scikit-learn

Pourquoi ?

Temps de traitements (apprentissage) longs

Opérations parallélisables :

- Cross-validation / recherche de paramètres
- Forêts Aléatoires
- ...

Mutualiser un cluster de machines entre plusieurs Data Scientists

Parallélisme avec Joblib

Joblib peut:

- utiliser plusieurs threads / processus sur 1 machine
- se brancher sur un backend externe (nouveau !)

n_jobs = -1 dans les classes scikit-learn

File Edit View Run Kernel Hub Tabs Settings Help

Demo Nested Parallelism.ip

Code

Python 3

Single host, multicore parallelism with scikit-learn and joblib

```
[ ]: from joblib import cpu_count

     print(cpu_count())
```

```
[ ]: from sklearn.datasets import fetch_california_housing
     from sklearn.model_selection import train_test_split


     calhousing = fetch_california_housing()

     X_train, X_test, y_train, y_test = train_test_split(
         calhousing.data, calhousing.target, test_size=0.1,
         random_state=0)
```

```
[ ]: from sklearn.model_selection import RandomizedSearchCV
     from sklearn.ensemble import RandomForestRegressor

     param_grid = {
         'n_estimators': [30, 50, 100, 200],
         'max_depth': [5, 8, 12, None],
         'max_features': [.5, .8, 1],
     }

     base_estimator = RandomForestRegressor()
     search = RandomizedSearchCV(base_estimator, param_grid
```

Options de déploiement de dask

dask-kubernetes (cloud ou localement)

dask-yarn (Hadoop)

dask-jobqueue (SLURM, PBS, SGE…)