Parallel Machine Learning and Artificial Intelligence

Dr. Handan Liu

h.liu@northeastern.edu

Northeastern University

Content

- Install Python Environment on Cluster
- Launch Jupyter Notebook on Cluster
- Examples
  - Example 1: Parallelize a Pandas DataFrame
  - Example 2: Parallel tuning C parameter in SVM
  - Example 3: Parallel GridSearch Cross-Validation
- Notes

Install Python Environment on Cluster

- To load anaconda, type **$ module load anaconda3/3.7** => $ which python
- To create your environment, type $ conda create -n py37 python=3.7 anaconda
  - Follow the prompts to complete the Conda install.
- To activate your Conda environment, type $ source activate py37
- To install a specific package, type $ conda install -n py37 [package]
- To deactivate the current, active Conda environment, type **$ conda deactivate**
- To delete a Conda environment and all of its related packages, type
  $ conda remove -n py37 –all

Reference: https://rc-docs.northeastern.edu/en/latest/software/conda.html

Launch Jupyter Notebook on Cluster

- Open OnDemand (OOD) is a web portal to the Discovery cluster.
- Type $ source activate py37
- Type $ conda install jupyterlab to install jupyterlab in your environment.
- Go to https://ood.discovery.neu.edu and sign in with your Northeastern username and password.

Reference: https://rc-docs.northeastern.edu/en/latest/using-ood/interactiveapps.html

Example 1: Parallelize a Pandas DataFrame

- When you parallelize a DataFrame, you can make the *function-to-be-parallelized* to take as an input parameter:
  - one row of the dataframe
  - one column of the dataframe
  - the entire dataframe itself

- Parallelize Pandas DataFrames with multiprocessing.

- The multiprocessing Pool.map, apply, starmap can parallelize Dataframe in their normal ways.
- Pool.imap: imap(*func, iterable[, chunksize]*)
  - A lazier version of map()
  - The chunksize argument is the same as the one used in the map() method. For very long iterables using a large value for chunksize can make the job complete much faster than using the default value of 1.

The pathos

- pathos is a framework for heterogeneous computing. It provides a consistent high-level interface for configuring and launching parallel computations across heterogeneous resources.
- It claims that pathos.multiprocessing is better multiprocessing and multithreading in python ⍰⍰
- Pathos follows the multiprocessing style of:
  - Pool > Map > Close > Join > Clear.
  - Check out the pathos docs for more info:
    - ✔ https://github.com/uqfoundation/pathos
- Install on the cluster:
  - $ source activate py37
  - $ pip install pathos <version 0.2.7 >

Example 2: Parallel tuning C parameter in SVM

- Implement a naive parallel *k*-fold cross-validation algorithm in Python for tuning the "C" parameter in SVC with linear kernel.
- Use Scikit-learn for the SVM algorithm and the test set from the UCI optical handwritten digits dataset; Implement *k*=5

Example 3: Parallel GridSearch Cross-Validation

- Tuning hyperparameters for ExtraTreesClassifier by GridSearchCV
- This class implements a meta estimator that fits a number of randomized decision trees (aka. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- Cross-validation generator is set StratifiedKFold for Classifier.
- Computations can be run in parallel if your OS supports it, by using the keyword n_jobs.

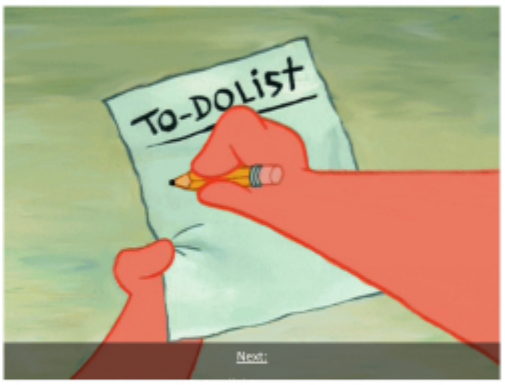
Note: Avoid Oversubscription of CPU Resources

- When using more processes than the number of CPU on a machine, the performance of each process is degraded as there is less computational power available for each process.
- Moreover, when many processes are running, the time taken by the OS scheduler to switch between them can further hinder the performance of the computation.
- It is generally better to avoid using significantly more processes or threads than the number of CPUs on a machine.

Note: Avoid running out of memory

- Multiprocessing works by replicating the same code and memory content in various new Python instances (the workers), calculating the result for each of them, and returning the pooled results to the main original console.
- If your original instance already occupies much of the available RAM memory, it won't be possible to create new instances, and your machine may run out of memory.
  - "Exceed Memory Limit" □ then exit.

- Stay safe!
- See you next class!



Next Lecture will Continue:

Parallel Machine Learning