Parallel Machine Learning and Artificial Intelligence

Dr. Handan Liu

h.liu@northeastern.edu

Northeastern University

Dask ML Hands-on

Install Dask Machine Learning and XGBoost

- Install in Jupyter,
  - pip install dask_ml
  - pip install dask_ml[xgboost] # also install xgboost and dask-xgboost
  - pip install dask-ml[complete] # install all optional dependencies

Hands-on: Dask ML

- Dask ML

Project Process

- Quiz:
  - Change from 3 times to twice.
  - The next quiz will be held in the week 10.
  - The next quiz will leave 1 hour for you to finish.

Project

- The project package is considered a final exam.
- Project Package: 40%

  Including:
  - Proposal 5%
  - Final Report 10%
  - Coding 10%
  - Presentation Slides and Oral Speech 15%

    -- including the attendance of these two presentation classes:
    - deducting 4 points for two classes, 2 points each class for absence with no reason, and no permission.
    - If you have a legitimate reason, you should contact the instructor beforehand, give the proof and get the approval.

Steps

1. Team and Choose a topic (you)
   - Data and Resource
2. Proposal (you)
3. Review and Advice (me)
4. Coding, Slides and Final Report (you) (I can give advices if necessary)
5. Oral Presentation (you do) (I evaluate and grade including above)

Step 1: Team and Choose a topic

- The project will be completed independently by individuals.
- You decide your team number and choose a topic
  - TA will post a form in Slack for you to fill in the team number and the topic. Everyone needs to fill in these two items lest your topic/content conflicts with others'.
  - The team number will determine the order of presentations:
    - ✔ Team 1 - 8 will do presentation in the 1st class on April 27.
    - ✔ Team 9 - 15 will do presentation in the 2nd class on April 30
  - Everyone shares his/her slides on Zoom and conducts a presentation of the whole project in 12-15 minutes.

Step 1: Team and Choose a topic

- From now on, you can start choosing the topic.
  - You can choose any topic that falls into the (application) field of machine learning and/or deep learning.

- Specific requirements in this course
  - You must do: 1) Choose the real large dataset or large-scale models in the real world; 2) Use parallel methods (2 or more) that you have learned (or will learn) on multiple CPUs or multiple GPUs (on one machine or one node) 3) Compare the parallel speedup of multiple parallel methods on multiple CPUs or GPUs; and analyze the results.

Data and Resource

- Requirement:
  - Dataset size is required to be big enough to get parallel performance. For example:
    - ✔ Image or video/audio dataset is recommended to exceed 1GB.
    - ✔ Numerical and/or string dataset is suggested more than 100MB.
  - The number of samples and features:
- Data Resource
  - Kaggle Machine Learning and Data Science Community
    https://www.kaggle.com/
  - UCI Machine Learning Repository
    http://archive.ics.uci.edu/ml/index.php
- Google: machine learning data sets more open-source datasets

Step 2: Proposal

- Proposal Template

  <Project Topic>

  <Team number and your name>

  Introduction

  <Including: background, motivations and goal, etc.>

  Methodology

  <Including: algorithms and methods you plan to use for parallel machine learning/deep learning in your project> *Don't just list the name of the approaches.*

  Description of your dataset

  Data Sources (download link)

Step 2: Proposal

- Submission Format:
  - MS Office Word
  - 2 - 3 pages

- Submission via email to h.liu@northeastern.edu
  - In email Subject, please clarify your team number, e.g.

    Email Subject: Team 6 Proposal in CSYE7105 Spring 21

- Submission due date: by the end of March 26th

Step 3: Review and Advice

- Instructor Review date: Tuesday class, April 6th

Step 4: work on your project

- Problem
- Solution
- Program
- Visualization
- Final Report

Template: Final Report

<Project Topic>

<Team Member>

Introduction

<Including: background, motivations and goal, etc.>

Methodology

<Including: algorithms and methods you are using for parallel machine learning/deep learning in your project>

Description of Dataset (including dataset download link)

Results and Analysis

<Including: the visualization of your calculation results and analysis of performance on a number of CPUs and/or a number of GPUs>

Conclusion

Reference

<list all resources you use in your project otherwise you will take the risk of plagiarism>

<Anything you want to add in your final report>

- The final report format:
  - MS Office Word
  - 12 font
  - Single line
  - You can use Word Styles like Heading1, Heading2 for section titles, but the body text should use “Normal”, with 12 font and single line.
  - The inserted figures should be set to an appropriate size.

- No page limit, but not less than 15 pages.

Step 5: Submission the Whole Package

- Submit the whole package:
  - The final report
  - The python file(s) and slurm script if used, or Jupyter files, or other files
  - The dataset download link (if not in the final report)
  - The ppt slides for presentation
- Submission via email to h.liu@northeastern.edu
  - In email Subject, please clarify your team number, e.g.

    Email Subject: Team 6 Project in CSYE7105 Spring 21
  - Attach the whole package with zip
  - Or, submit your google drive link
- Submission due date: by the end of April 24th

Step 6: Presentation

- For each team, there is a limit time for presentation.
- I will post the details later when you prepare slides.
- ……

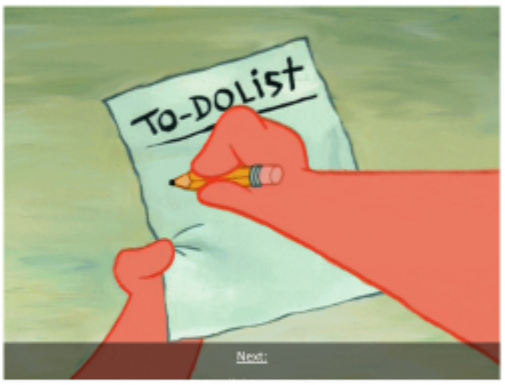
Timeline

Steps	Timeline	
Team and Topic	March 12	
Submit the proposal	March 26	
Review the proposal	April 6	
Final submission	April 24	
Presentation	April 27	April 30

Mid-Semester Course Review

- Most students to give reviews.
- Reviews and comments are very positive ("Strong" and "Very Strong"). Greatly appreciated all reviews and comments!
- Feel free to give me suggestions via email or message (Slack/LinkedIn/others) at any time. I have a Slack app. I can respond you as soon as possible.
- Please remember to spend time to give me reviews on **TRACE** in the end of the semester.

- Stay safe!
- See you next class!



Next Lecture will Continue:

GPU Computing