Using High Throughput Computing for a Simulation Study on Cross-Validation for Model Evaluation in Psychological Science

Hannah Moshontz, PhD
Background

- Goal pursuit in everyday life
- Machine learning for predicting complex, multiply-determined outcomes

Technical / programming skills

- Advanced training in quantitative methodology
- Proficient with programming languages for statistics — SAS and R
- Prior to this project, very limited BASH, no HTCondor
Project Background

- Lack of norms or guidance on machine learning practices in psychological science
- People incorrectly interpreting / using cross-validated model performance estimates in top journals
- Goal: characterize and give guidance on cross-validated model performance estimates in data contexts typical of psychological science

https://psyarxiv.com/ns9mj/
Simulation Study Overview
Simulation Study Overview

- Entailed running about 144,000 times the number of models in a single ML study
- Within a single ML study, tens of thousands of models are run. Run time is less than a minute to weeks, depending on data context.
- Compute time well over 1 million hours
Simulation Study Overview

- Entailed running about 144,000 times the number of models in a single ML study
- Within a single ML study, tens of thousands of models are run. Run time is less than a minute to weeks, depending on data context.
- Compute time well over 1 million hours

*We didn’t understand the scale of this project initially.*
Our software and scripts

Each job we submitted had:

- R script (.R)
- Arguments (.csv)
- PRE and POST scripts (.sh)
- Submit file (.sub)
- Executable file (.sh)

(We started with DAGs, but didn’t use later)

We used notepad ++, vim, and later created all files in R.
Our software and scripts

Output from each batch

- Zipped outdata - for each job, a summary of the best model performance, plus information about the model (.rds)
- Zipped “job files” - the submit and executable files, the args, the error files, the output files
- The log
### How we tracked jobs

<table>
<thead>
<tr>
<th>nX</th>
<th>nN</th>
<th>UB</th>
<th>Complexity number</th>
<th>Elapsed time range</th>
<th>Memory use (max)</th>
<th>Disk use (max)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>100</td>
<td>B</td>
<td>1</td>
<td>5M to 46M</td>
<td>1075</td>
<td>580917</td>
<td>4 holds, released and ran/completed - kendra</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>B</td>
<td>1</td>
<td>55M to 3H 51M</td>
<td>977</td>
<td>580918</td>
<td>3 holds - released/ran with 1 hold - said size of job was 17090, to resolve needed to up the memory</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>B</td>
<td>3</td>
<td>2M to 55M</td>
<td>849</td>
<td>580914</td>
<td>14 jobs held for memory reasons; upped to 2048 - hannah</td>
</tr>
<tr>
<td>10</td>
<td>500</td>
<td>B</td>
<td>1</td>
<td>11M to 30M</td>
<td>1159</td>
<td>580917</td>
<td>6 hold, released and ran/completes - kendra</td>
</tr>
<tr>
<td>100</td>
<td>500</td>
<td>B</td>
<td>2</td>
<td>15M to 2H 10M</td>
<td>952</td>
<td>580915</td>
<td>16 holds - released and ran/completed - kendra</td>
</tr>
<tr>
<td>1000</td>
<td>500</td>
<td>B</td>
<td>3</td>
<td>4M to 4H 10M</td>
<td>1014</td>
<td>580917</td>
<td>19 holds for fluke memory reasons, all ran fine on release - gayle</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>B</td>
<td>1</td>
<td>16M to 1H 15M</td>
<td>1612</td>
<td>580917</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>100</td>
<td>1000</td>
<td>B</td>
<td>2</td>
<td>24M to 4H 33M</td>
<td>1076</td>
<td>580915</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>B</td>
<td>3</td>
<td>4M to 5H 13M</td>
<td>1177</td>
<td>580921</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>U</td>
<td>1</td>
<td>7M to 17M</td>
<td>900</td>
<td>580917</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>U</td>
<td>1</td>
<td>52M to 4H 19M</td>
<td>1005</td>
<td>580917</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>U</td>
<td>3</td>
<td>3M to 47M</td>
<td>865</td>
<td>580914</td>
<td>8 jobs held for fluke memory reasons (impossible peak usage); ran fine on release - gayle</td>
</tr>
<tr>
<td>10</td>
<td>500</td>
<td>U</td>
<td>1</td>
<td>8M to 22M</td>
<td>1000</td>
<td>580917</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>100</td>
<td>500</td>
<td>U</td>
<td>2</td>
<td>10M to 2H 45M</td>
<td>946</td>
<td>580916</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>1000</td>
<td>500</td>
<td>U</td>
<td>3</td>
<td>4M to 4H 5M</td>
<td>926</td>
<td>580916</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>U</td>
<td>1</td>
<td>9M to 41M</td>
<td>1251</td>
<td>580981</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>100</td>
<td>1000</td>
<td>U</td>
<td>2</td>
<td>29M to 8H 28M</td>
<td>2048</td>
<td>580979</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>U</td>
<td>3</td>
<td>6M to 9H 45M</td>
<td>1120</td>
<td>580952</td>
<td>18 cases, released and ran/completed - kendra</td>
</tr>
</tbody>
</table>
Errors and troubleshooting

- Simple errors we made
- Issues related to software when flocking/gliding
- Issues related to errors with our R script
- Issues related to how we broke jobs up / optimization

CHTC staff offered *tremendous* support, both via direct contact and via the excellent HTCondor manual and other online documentation and resources.
Simple errors with HTCondor

- Unix line endings
- Typos in our submit or executables
- Not moving files to the submit server
- Not running PRE.sh
- For big jobs, zipping too many at once
Issues related to flocking, gliding, and R

- Needed to download a support tar (SLIBS) from the SQUID servers
- We changed the version of R we were using, and had issues with package dependencies in our package tar
- There was a set of machines on UW’s campus that were having odd issues with base R
- Jobs would get booted when flocking/gliding (many of our jobs were near or just over 8 hours)
Issues related to the project / R scripts

- Adjustments to how we simulated data - reflection and reviewer feedback
- A few contexts had so few positive cases that the models failed, which didn’t produce the output that we were expecting
Issues related to optimization

- Making a single zipped file that took hours to unzip, not inspecting contents before unzipping
- We wrote the script in a way that was well-suited to being broken down, but in rigid ways
- We starting running jobs before we had tested the most complex contexts and hit a floor in how simple the jobs could be (the most simple was still taking over 72 hours with a particular algorithm)
- We didn’t understand that small efficiencies scale, and are important (e.g., ranger vs RF)
Helper scripts

- **Meta-script**
  - created all files that needed to run a batch of jobs, including changing line endings, making the R script, making the args file for completed jobs, summarized the log, made and ran the other helper scripts

- **Check / unzip script**
  - checked the outdata and unzipped
  - produced an args file for any missing jobs

- **Aggregation script**
  - Made data comparable no matter how jobs were broken up
Advice

- Automate what you can to prevent errors & save time
  - Make files and folders descriptive and machine-readable
- Document everything well
  - keep detailed notes about testing and completed jobs
  - save HTC files
- Be mindful of the resources you will use and are using
  - computing hours can be abstract and hard to estimate! check how many hours you / your team has used
  - reevaluate the scope of your project periodically
  - avoid waste through preparation and testing
Advice

- **Use the HTCondor manual**
  - there are so many useful functions and so much information that you have access to

- **Become a pro troubleshooter**
  - learn to systematically rule out basic issues and diagnose the issue you have
  - reach out for help with detailed information about what you have done, and with jobids, logs, and other documentation

- **Describe the time/resource constraints that informed your research**
  - reviewers may not appreciate these constraints unless explained
Closing thoughts

- We couldn’t have conducted this study in my lifetime without HTC
- The CHTC staff are an incredible resource, and this project wouldn’t have been completed without them and the HTCondor manual
- A great training / learning experience re: general programming skills (e.g., BASH, troubleshooting / problem-solving)
Thank you