Machine Learning for Healthcare - Technical Tips

0368-4273

Computing Resources

SLURM

If your project requires computing resources (e.g., GPUs for training deep networks), you can use the **school's SLURM cluster**. SLURM is a resource manager and job scheduling system that allows you to run your code on powerful servers equipped with multiple CPUs or GPUs.

To use the cluster:

1. **Connect** to one of the **client servers c-[001-009]**

2. **Write a submit file** (e.g., `mlhc.slurm` file) that specifies:

   - SLURM job parameters using `#SBATCH` (e.g., time, memory, partition, etc.).
   - Shell commands to execute your job - e.g., activate your conda environment and run your Python code.

3. **Submit the job** to the cluster with:

   ```
   sbatch mlhc.slurm
   ```

For full instructions, see cs.tau.ac.il/system/slurm.

Partition. In SLURM, a **partition** is a logical group of compute nodes that differ in hardware, runtime limits, or access control. You can use `studentkillable` or `studentbatch` partitions, i.e.:

```
#SBATCH --partition=studentkillable
```
Max runtime: 1 day

Or:

```
#SBATCH --partition=studentbatch
```
Max runtime: 3 days

If you encounter a technical problem, open a ticket by email to system@cs.tau.ac.il and they will try to assist.

Storage. The course has a shared 400GB storage space. Each student has a personal directory under `/home/yandex/MLFH2025/`. You are welcome to utilize it, but please be mindful, as the limit is shared among all the course projects. If additional space is required, please inform us.

If you encounter any issues with compute or storage resources, please let us know.

Google Colab

If suitable, you can continue working with Google Colab, as done in the course assignments.

Here are some useful tips:

1. **Handling large files**: When dealing with large files like models, saving and loading them in Colab can be time-consuming. To simplify this process, you can utilize the following methods:

   – **Mount Drive:** Grant the notebook access to your drive, enabling seamless interaction with your files. see more.

   – **Direct Download from Drive:** Download files directly from the drive without the need for mounting. see more.

2. **GPU and space:** You have 100GB of drive space in your TAU email. You can also use your TAU account to partially overcome the GPU and parallel session limit in Colab.

SQL Queries

Accessing BigQuery from Python

In order to access BigQuery outside Google Colab (e.g., PyCharm), install the BigQuery API Client Library, link it to your Google account, and use:

```python
from google.cloud import bigquery
client = bigquery.Client('YOUR BQ PROJECT')
```

instead of:

```python
from google.colab import auth
from google.cloud import bigquery

client = bigquery.Client('YOUR BQ PROJECT')
auth.authenticate_user()
```

New Queries

In order to add new modalities to your project you will probably need to write new SQL queries. You could alternatively download the raw CSV files at physionet and process them directly but it is much less recommended.

We suggest utilizing the BigQuery console to write and debug these queries before implementing them in your code. MIMIC-III has a well-documented repository at https://github.com/MIT-LCP/mimic-code with different SQL queries. You can use this repository **with proper credits**. If you rely on their code or queries:

   – Review it carefully and make sure you understand every step.

   – Verify it - Don't consider them as ground truth.

   – Add credit in the code documentation.

Good luck!