Lab 4: Reproducibility

2025-09-26

Table of contents

Reproducibility

Today's lab is about reproducibility, an important skill to professionals in industry and academia alike. Intimately related to communication and documentation, the main goal of reproducibility is to ensure others (and your future self) are able to reach the same results you did by following a sequence of steps. Our focus in the context of this course will be to attempt to reproduce the results from a recently published papr, pondering along the way whether we would have done anything different and assessing our overall experience. The paper in question is Comparative Advantage of Humans versus AI in the Long Tail (2024) by Agarwal, Huang, Moehring, Rajpurkar, Salz and Yu.

Setup

- Find a link to "Replication Package" on the paper's page and click on "Download this project". Authentication is required; we recommend to "access through your institution" and fall back to the other options in case this fails. By the end of this step, you should have a file named `202185-V1.zip` on your machine. If that also does not work, you may try `wget www.stat.berkeley.edu/~paciorek/transfer/202185-V1.zip`.
- Unzip `202185-V1.zip` and enter `Rad_AI_Longtail/`. (if you are unzipping from the mirror instead, after running `wget`, create a directory `Rad_AI_Longtail/` and unzip the file inside it.)
- Investigate which files are likely to contain reproducibility instructions and give it a quick read.
- We now need to use the specific python version mentioned by the authors (3.11) to install the packages necessary to reproduce the results, which should be straightforward given the `requirements.txt`. But there are several caveats here. Try installing those packages and document your experience. Creating a conda environment using the provided `requirements.txt` file may be a reasonable attempt: `conda create --name lab4 python=3.11 --file requirements.txt`. If it fails on your system, try to identify why.

- As an alternative, you may try the following command to generate an adjusted requirements file that you can then use to install relevant packages:

```
find . -type f -exec grep "import " {} \; |
    sed 's/"//g; s/,//g; s/\\n//g; s/^ *//; s/ *$//' |
    cut -d ' ' -f 2 |
    cut -d '.' -f 1 |
    sort |
    uniq |
    grep -F -f - requirements.txt |
    grep -v '^#' |
    cut -d '=' -f 1,2 |
    sed 's/=/==/g' |
    grep -v '_' |
    grep -v '\-base' \
    > requirements_adjusted.txt
```

- Running the command above incrementally might be useful to better understand what each step does if it is not immediately clear. Installing the packages should then be less prone to error by using this `requirements_adjusted.txt` file. You will also need `nbformat==5.9.2` and `jinja2==3.1.2` to be installed. Think in particular about the following modification `sed 's/=/==/g'` and what are the implications.

Replicating the results in the paper

The goal is now to try to replicate the paper, pondering about important questions along the way. Format your reasoning to the points below in a pdf file and submit it to Gradescope by the end of the lab.

- With a proper environment set up, proceed to identify what is the minimal set of input files that is needed to run all the scripts. Are all input files used? Are there any files provided that you wouldn't expect to be input?
- Start following the instructions to replicate the results. Try to locate the "radiology experiment data" mentioned nd report back. If you are unable to find it, you may download it via `wget https://www.stat.berkeley.edu/~paciorek/transfer/data_public.txt` instead. Discuss the implications.
- Check the file permissions of the shell script via `ls -l make.sh`. You may change the permissions via `chmod` if desired. Discuss.
- Explore the `make.sh` script and discuss how it helps or hinders reproducibility. In addition, pay attention to how failures are handled via bash in the script and discuss what you think the motivation was.
- Note that `make.sh` invokes both python and ipython. Double check both executables are on the correct version. If you do not have ipython installed, you may do so via `conda install ipython` or `pip install ipython` depending on what kind of virtual environment you set up.
- In what follows, we are going to run scripts that output text to the terminal. Document along

do way your opinion about the text that is purposefully printed (informative or uninformative; excessive or lacking; etc.) and about any other text printed to screen (are there warnings, for example?).

- Run the `make.sh` script up to the "main calculations" part. That step takes hours because the number of "bootstrap replicates" is large. Identify how that parameter is set and decrease it. Discuss your experience in making this change and whether you would have programmed this differently.
- Even though we adjusted the "bootstrap replicates" parameter to reduce the time the script takes, we are going to skip this step and conveniently download the files that would be generated by the script with "bootstrap replicates" set to 51. Run `wget www.stat.berkeley.edu/~paciorek/transfer/data-analysis-bootstrap.zip` and `wget www.stat.berkeley.edu/~paciorek/transfer/data-analysis.zip` to do so. Proceed to unzip the files and ensure the files are in the correct directories.
- Run the last two scripts in `make.sh` to generate the plots. Do they look similar to the ones in the paper? If not, what could explain the difference?
- Note the contact information given in the instructions file. Are the two options feasible ways to reach the authors?