Introduction

This assignment will assess your critical thinking skills in the fields of machine learning and bioinformatics. This is not a purely technical test and we are looking to understand your data exploration and methodology intuition.

We are looking at your interpretive journey so there is no “correct” solution. Rather, we aim to look at how you interact with the data and your ability to communicate your results effectively.

Final Deliverable: Written report (2 pages max. excluding references):
Due Date: August 15 2023
The project report is formatted as a mini scientific paper. It should include:

- Introduction with background, principal references, problem statements
- Results: a few sections with major findings with a few figures with figure captions
- Discussion or Conclusion section
- References
- Methods/Rationale (why you choose this method)
- Github repository with code used
- Ensure you adhere to the page limit

Evaluation Criteria:

- Ability to generate an interesting research question and draw insightful conclusions
- Demonstrated understanding of current bioinformatic and machine learning methods & application in biomedicine
- Application of concepts for evaluating and enhancing generalizability of learning procedure in formulating and / or evaluating the described methodology
- Appropriate approaches for enabling reproducibility

Resources for getting started with programming ML/AI applications in python/R

- https://pytorch.org/tutorials/
- https://pytorch-lightning.readthedocs.io/en/latest/
- https://www.youtube.com/watch?v=M3ZWfamWrBM&ab_channel=freeCodeCamp.org
- https://keras.io/examples/
- https://scikit-learn.org/stable/auto_examples/index.html
- https://scikit-learn.org/stable/tutorial/index.html
- https://lgatto.github.io/IntroMachineLearningWithR/an-introduction-to-machine-learning-with-r.html
- https://bradleyboehmke.github.io/HOML/intro.html

Assignment: Single-cell cell-type classification using RNA sequencing data

Project description:

Background

Single-cell RNA sequencing technology quantifies transcriptomic expressions at single-cell level, leading to discoveries rooted in cell-to-cell heterogeneity. Cell-type classification, or assignment of cell type labels to single cells, is an important step in analyzing the RNA sequencing data. For example, correctly identifying cells within the tumor microenvironment can provide oncologists with a snapshot of how a patient's immune system reacts to the tumor.

Aim

Develop an ML-based classifier for single-cell cell-type classification.

Suggested methods/experiments

- Leverage different ML/DL models (e.g. deep neural network, SVM, random forest)
- Ablation experiments
- Be creative!

Datasets:

https://github.com/10XGenomics/single-cell-3prime-paper/tree/master/pbmc68k_analysis

Pbmc68k is the "gold standard" dataset for evaluating cell-type classification. The provided GitHub link includes R scripts that conduct analysis on PBMC68k including dimension reduction via PCA, visualization via t-SNE, and cluster analysis via k-means clustering, etc. You can use the scripts to pre-process the dataset.

Computational resources required:

GPU is recommended

System Memory > 8GB is recommended

If extensive hyperparameter tuning is performed, HPC recommended. If access to HPC is unavailable, please indicate this on your report. **Note**: If HPC access is unavailable, you may reduce the features in the dataset in order to enable model development and downstream experiments, but ensure you discuss the steps taken for dimensionality reduction in the report.

If you have any questions please email: senthujan.senkaiahliyan@uhn.ca