CS690A: Assignment 1 - Clustering of Single-cell Sequencing Data

Hamim Zafar
hamim@iitk.ac.in

Indian Institute of Technology Kanpur — August 23, 2024

Introduction

In this assignment, you will be performing clustering of a single-cell RNA-seq dataset. The dataset has been generated from human immune cells. There are two tasks associated with this dataset as described below. **This assignment needs to be done individually. No grouping will be allowed**.

Dataset Description

The assignment contains one file titled 'dataset.h5ad' which contains the single-cell gene expression matrix. You can read this file using Scanpy (`https://scanpy.readthedocs.io/en/stable/`) library.

The file can be accessed from the kaggle competition page.

Tasks

Task 1.1 (35 points)

Perform clustering of the dataset using Leiden clustering method. Before clustering, you need to perform QC, normalization, feature selection and PCA. For feature selection, use the top 2000 highly variable genes. Vary the resolution parameter and select the clustering which you think is the best for your data. After clustering with optimal resolution, differential expression analysis for each cluster is performed to determine the marker genes for each cluster. Based on the marker genes, annotate the clusters you have obtained. The following genes may be of interest to you - *HBD*, *CD3D*, *CD79A*, *NKG7*, *CD3*, *CD4*, *CD8*, *CD197*.

Task 1.2 (15 points)

Instead of highly variable genes, use deviance for feature selection. You can use pipecomp (`https://github.com/plger/pipeComp`) for this. Perform clustering on this set of genes. Compare your clustering results against that of task 1.1.

Task 2 (50 points)

You need to come up with a better clustering algorithm for clustering the given dataset. You can evaluate your clustering algorithm through kaggle. The leader board for the challenge will be maintained based on the performance on the given data. You can number your clusters in the range $[1, K]$ where K is the number of clusters inferred by your method. Consider the points below when preparing your solution.

- You need to be creative in designing your clustering algorithm. You can adopt clustering algorithms recently published in conferences such as Neurips and ICML. Specifically, you should consider clustering methods developed for high-dimensional, sparse datasets.

- Dimension reduction is an important step for clustering such datasets. You can experiment with deep learning-based nonlinear dimension reduction techniques for better clustering of the dataset.
- Some clustering methods can have certain parameters which can affect the clustering results. Make sure to experiment with such parameters to improve clustering accuracy.

Listing 1: Format of output csv file

```
Id, Label
0, 1
1, 2
2, 3
3, 2
4, 1
```

Notice: In case we require a change in the format of the csv file, we will notify you. Keep an eye on the announcements.

Kaggle Leader board: You can submit the csv files multiple times and check your performance.

Deliverables

The deliverables for the assignment are the following

1. Clustering prediction for the three tasks on the dataset. These results will be evaluated and the leader board will be maintained based on the scores in evaluation. For task 1.1, you will be graded based on the successful execution and experimentation of Scanpy's Leiden clustering and other steps. For task 1.2, you will be graded based on the successful execution of deviance-based feature selection and clustering. For task 3, you will be graded based on your position on the leaderboard.
2. Runnable code (in Jupyter Notebook) for all three tasks.
3. Scripts for running your code to generate the predictions. TAs will run these scripts to reproduce the csv files you submit for the challenge
4. A short report describing the steps taken to solve the challenge. Describe in brief the algorithms you have used, any dimension reduction you have performed, training process, training accuracy, etc. The report should also contain the feature plots (based on the obtained clusters) for each task. The writeup should mention your name and roll number.

For submission, all the deliverables should be zipped in a single file and the zip file should be named as `Name_CS690_assignment_1.zip`, `Name` should be replaced by your first name. Also, each file in the zip folder should start with the phrase `Name_` (`Name` replaced by your first name). The file should be submitted via hello.iitk portal.

Submission Deadline

August 30^{th} 11:59 PM.

Kaggle Competition link

`https://www.kaggle.com/t/8182488ead4b45b0a5f314c7ddb4b666`