Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Machine Learning for Genomics
Spring Semester 2023

Project 2: Single cell imputation and clustering

Assigned on: **5:00pm on 05.04.2023** Due by: **11:59pm on 03.05.2023**

1 Exercise 1 - Theory

We will use the term **imputation** in Project 2 for inference of a proxy of the transcript expression for genes with zero observed counts in single-cell data (*i.e.*, inference of dropout events).

Please answer these questions in the final report slot on the submission website. You will not be formally graded on these questions but they should help you perform better on the practice section.

1.1 Question 1

What is the ”dropout” effect in single-cell data? What are the major causes for the inflated number of zeros in single-cell RNA sequencing?

1.2 Question 2

What should a good imputation method achieve (Please list 2 objectives)? What should a good imputation method avoid doing (Please list 2 consequences to avoid)?

1.3 Question 3

Samples from different patients may be processed by different technicians and/or at different time points. How could this possibly affect the single cell RNA sequencing data? What type of method is supposed to correct for these potential confounding effects?

2 Exercise 2 - Practice

In this section you will have to impute missing values in a real-world dataset and find meaningful clusters of cells. You will be studying pancreatic cells coming from healthy individuals and individuals with Type 2 Diabetes Mellitus. There are a total of 7 patients in the dataset ($n = 4$ for the train set, $n = 3$ for the test set).

We wish to know if the cellular composition of pancreatic cells differs between healthy and unhealthy patients, as well as if we can find reliable biomarkers of the cell types across patients. To do so, we want to obtain clear clusters in our dataset that represent the underlying biology. You will perform two separate tasks on the dataset. For the first task, you will have to impute missing data in single-cell expression to correct for technical dropout. For the second task, you will have to perform clustering on your data to find biologically meaningful clusters; the clustering output that you will submit can be performed on the raw or imputed counts - you choose! Other than imputation, you are free to perform whatever transformations on your data you wish.

You will be provided with a train dataset, for which you will have single cell RNA sequencing data (raw counts), paired bulk data (raw counts), the patient of origin and the individual cell types provided. You will be evaluated on a test dataset, for which you will only get the single cell RNA sequencing data (raw counts) and the patient of origin.

You will be evaluated according to the two following tasks:

- **Correlation with paired bulk data** We will be evaluating how correlated your data are to paired bulk data. The expression you discover in your single cells should be reflected in the bulk sequencing, where technical dropout does not occur as much. For this purpose we will use
  - Spearman's ρ between the "bulkified" data and bulk data.

    By "bulkified" data we mean here the patient-wide average of gene expression. We will compute the Spearman's correlation with the log-transformed expression of bulk data. We will compute the average ρ for all patients in the test set.
- **Clustering performance** Your clustering method must identify as best as possible the different ground-truth cell types present in the dataset. This would allow you later on to get meaningful biomarkers through differential gene expression analysis. You will thus have to pick a clustering algorithm to apply to whichever transformation of your data you choose. You will be evaluated according to 3 metrics:
  - The silhouette score SSC in the Principal Component Analysis (PCA) transformed space (number of Principal Components PCs 50)
  - The Adjusted Rand Index ARI
  - The V-measure score V

You will pass the project if you beat both baselines:

- ρ for the imputation task,
- $\frac{1}{3}SSC + \frac{1}{3}ARI + \frac{1}{3}V$ for the clustering task.

A bonus of +0.25 on the semester grade will be given to the 10% students who perform the best. To choose groups getting the bonus, we will compute the Z-score associated to both subtasks (imputation and clustering) and sum them up. The 10% of students with the highest sum of Z-scores will get a bonus. As a reminder, bonuses are non cumulative for the class (e.g., if you have already received a bonus for project 1, you will only get +0.25 total even if you get a bonus for project 2).

You will have to provide the following files:

- The code you wrote for your analysis. Your code must be commented, readable and must run. You must provide a *requirements.txt* file listing the packages used in your analysis.
- For every .csv file provided, you have to ensure your rows are indexed $(0, 1, ..., n-1)$

- You will have to provide a document named *imputed_bulkified.csv* containing the bulkified version of the imputed data, meaning a matrix M where M_{ij} corresponds to the imputed value of gene i for patient j. To get the gene expression for patient j, you must simply average the library size-corrected values of the gene expression over all cells of the same patient. The first column of the file should be the gene names in the same order as they were provided in the test single cell data. The file is expected to have a header with the column names ”index”, and then all the patients in the test dataset. There should be as many rows in this file as there are genes in the provided dataset (excluding the header row).

- A file *cluster_membership.csv* containing two columns, the first the indices of the cells in the test dataset, the second the cluster membership of the cell (please ensure your cluster membership indices are 1-indexed, **not** 0-indexed). The file is expected to have a header with the column names ”index”, ”cluster”. There should be as many rows in this file as there are cells in the provided test dataset (excluding the header row).

- A file *PCA.csv* containing the coordinates of the cells in the test dataset in the 50 first PCs computed on your data (make sure you compute your PCA correctly!). This file should contain 51 columns, the first the indices of the cells in the test dataset, the 50 next corresponding to the 50 first PCs. The columns should be named ”index” and then ”PC1”, ”PC2”, ... ”PC50”. There should be as many rows in this file as there are cells in the provided test dataset (excluding the header row).

These files should be saved in a .zip file named *LastName_FirstName_Project2.zip*. We have provided an example of the output files you have to upload in the Project 2 archive.

You will also have to fill in on the submission website page a quick description (10000 characters max) of your work and the steps you took to perform the project. In this report, we will expect

- The answer to the three theory questions of Exercise 1,
- A quick justification of the different steps you took to perform the two tasks,
- A quantitative analysis of the influence of imputation on the clustering tasks (ie difference in SSC, ARI and V score), using the clustering method that you chose,
- Your specific contribution to the project.

Code and files can be shared across members of the team, however this description **must be individualized**.