1st MILESTONE - Binary classification of normal and cancer-associated fibroblasts from scRNA-Seq

Vojtěch Melichar

CTU - FIT and UCT Prague

melicvo1@fit.cvut.cz

December 1, 2023

1 Assignment Summary and Data

Single-cell RNA-Seq is a state-of-the-art technique for measuring the transcriptional profile of each individual cell in the tissue sample. Simply put, the transcriptional profile can be seen as a capture of the current state of the cell; it tells us which genes are active and which are turned off. In turn, a particular gene or set of genes can be associated with a disease. However, these *in sillico* results must be validated by further *in vitro* or *in vivo* experiments.

In this particular experiment [1], researchers took samples from normal and cancerous tissue. These two classes are the target variable. The intensity of each gene is measured in each cell. This makes up a count matrix, which is used for training and testing of the model. The features correspond to individual genes, and each row is an individual cell. Due to the nature of the experiment, the count matrix has a high level of sparsity. The dataset contains approximately 25k genes and 29k cells and is publicly available.

2 Methods

The dataset was split into training, validation and testing datasets, with ratio of 50/20/30. The NN was designed in `pytorch`.

There are two main goals in this project. Firstly, create a neural network for binary classification of cells. The size of the input layer is equal to the number of genes in the dataset.

Secondly, rank genes according to importance by recursive feature elimination (RFE). This can return a set of only a few genes that is the best predictor of this type of cancer.

3 Current Results

Preliminary results show that there is a good level of separation between classes because even a fairly simple NN can predict on the test set with an F1 score greater than 0.95.

In the final stages of RFE when there are less than 100 features, I am planning to use a different, more simple machine learning algorithm to properly rank the most important features. NN is not an ideal tool for such a small set of features.

References

[1] Karolína Strnadová et al. "Exosomes produced by melanoma cells significantly influence the biological properties of normal and cancer-associated fibroblasts". In: *Histochemistry and cell biology* 157.2 (Feb. 2022), pp. 153–172. ISSN: 0948-6143. DOI: 10.1007/s00418-021-02052-2. URL: https://europepmc.org/articles/PMC8847298.

[2] Juexin Wang et al. "scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses". In: *Nature Communications* 12.1 (Mar. 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-22197-x. URL: http://dx.doi.org/10.1038/s41467-021-22197-x.

[3] Fan Zhang et al. "Recursive Support Vector Machine Biomarker Selection for Alzheimer's Disease". In: *Journal of Alzheimer's Disease* 79.4 (Feb. 2021), pp. 1691–1700. ISSN: 1875-8908. DOI: 10.3233/jad-201254. URL: http://dx.doi.org/10.3233/JAD-201254.

[4] Yan Zhou et al. "scDLC: a deep learning framework to classify large sample single-cell RNA-seq data". In: *BMC Genomics* 23.1 (July 2022). ISSN: 1471-2164. DOI: 10.1186/s12864-022-08715-1. URL: http://dx.doi.org/10.1186/s12864-022-08715-1.

Repository

Link to the repository:

`https://gitlab.fit.cvut.cz/melicvo1/mvi-sp`