DENSITY ESTIMATION OF SINGLE CELL MASS CYTOMETRY DATA WITH GENERATIVE MODELS

PROJECT DESCRIPTION

BY

FROMSA HERA

SUPERVISED BY

NELLO BLASER

FEB 2020

The University of Bergen

Norway

Overview

Mass cytometry can measure up to forty different parameters for millions of blood cells. The multivariate density of healthy and diseased blood is not well understood. In this project, generative models, such as *Gaussian mixtures, variational autoencoders* and *generative adversarial models* will be used to estimate the multivariate density of blood cell parameters.

Thesis

The thesis of this project is as follows: Does estimating the multivariate density of blood cell parameters give any reasonable generative model; if this is the case, could such model augment the cytometry data in order to train a classifier to distinguish between healthy and diseased blood cells?

Resources

- The cytometry dataset from the Flow Repository.
- Possibly other datasets.
- A GPU processor for training the models. Could be cloud and/or local at the institute of informatics at UIB.

Process

The process consists of these steps as of now;

1. Understand the data.
2. Cyclically
   (a) Run a model on the data.
   (b) Validate the results.
   (c) Tune model parameters.
3. Compare the variability between models.
4. Train a classifier on a data augmented by a generator model.

Timeline

Sem. 1 Write project description.

Set up resources.

Analyse, preprocess data and write a report.

Choose generative models. (explain why they are appropriate)

Sem. 2 Train chosen models and write a report.

Fine tune parameters. (explain choices)

Compare models.

Sem. 2 Write a first draft of the thesis.

If time allotted:

- Choose some classifiers.
- Train and evaluate the classifiers.
- Fine-tune parameters.

Sem. 3 Write and Submit a final draft of the thesis within 1 June 2021.

Expected Results

- The underlying density of the data might be too complicated to estimate.
- The augmentation of the data might not have a desired effect for the classifiers. The quality of the generators' estimation as well the quality of the data will affect the quality of the classifiers.
- No single generative model will be the "best" model: their strength and weakness might vary across different metrics such that no one is optimal across all metrics.

Requirements

1. A complete thesis paper.
2. One or more generative models estimating the density of cytometry data.
3. If time allotted, a classifier that can distinguish between deceased and control patient blood cells.
4. Midway presentation of the project to other master-students.
5. Defend the thesis.