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, you will use generative models, such as *Gaussian mixtures, variational autoencoders* and *generative adversarial models* to estimate the multivariate density of blood cell parameters.

Purpose

The purpose of this thesis to find out if estimating the multivariate density of blood cell parameters will give any reasonable generative model. If that is the case, these generative models might be used for augmenting the 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 data
2. Run basic models on the data
3. Compare models
4. Validate the results
5. Tune model parameters.
6. Compare and understand the variability
7. Train a classifier on an augmented data using 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. (note training time)

Fine tune parameters. (explain choices)

Compare models.

Sem. 2 Write a first draft of the thesis.

If there is time left expand the task to include the training of classifiers.

Choose classifiers.

Train and evaluate the classifiers.

Fine-tune parameters.

Sem. 3 Write and Submit a final draft of the thesis!

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 masterstudents.
5. Defend the thesis.

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.