Project 2

October 16, 2019

1 INF367 - Project 2

- Deadline: Sunday, 17.11.19, 23.59
- Submission details: Jupyter-notebook
- Deliver here `MittUIB.no/assignments`

Projects are a compulsory part of the course. This project contributes 20% to the final grade. The grade will be based good choice of methods, correctness of answers, clarity of code and thoroughness and clarity of reporting.

Deliverables: A jupyter notebook containing all the code to reproduce your work and a report of all your methodological choices and results. Please "restart and run all" before submission, so that you submit a clean version.

Code should be documented and special tricks (e.g. to avoid division by zero, to make sure it takes finite time to run, etc.) should be reported. The rational behind all steps in the code should be clear from the report. In particular, if you use subsampling, you should report it, and you should consider for each step how much subsampling is appropriate.

NOTE: This project is a learning experience. If we see that you have copied your answers from online resources, you will get 0 points. This is an application project, that means you may use any freely available library for the application tasks of the project.

Model selection is an important part of the task and will be graded accordingly. Before applying machine learning algorithms, you should always consider (and report) what results you expect. When you have successfully applied machine learning algorithms, you should always comment on how well the results match your expectations.

1.1 Task 1 - Preprocessing

(10 points)

In the next task, you will prepare the mass cytometry dataset for analysis.

The datasets contain information on 20,000 blood cells of 20 rheumatoid arthritis patients and 20 healthy controls. The first two columns identify the patient and the patient group. The remaining columns are the cell markers measured.

In this task you summarize and visualize the data and prepare it for analysis.

- Check for any missing values and handle these appropriately.
- Find the ranges and basic statistics of the features and rescale them if appropriate. For similar data, scaling using `arcsinh(x/5)` has been used successfully.
- Visualize the univariate densities of all features using your favorite density estimator.

- Calculate basic bivariate statistics, such as correlations.
- Perform any other appropriate preprocessing steps.
- Discuss the results of your summaries and visualization efforts and explain your preprocessing choices (not doing any preprocessing is also a choice).

1.2 Task 2 - Dimensionality reduction

(15 points)

- Visualize the mass cytometry dataset using at least three different representation learning algorithms.
- Explain your choices of algorithms.
- For each algorithm, explain your choice of parameters.
- For each dimensionality reduction, describe the main features you see and discuss if these features come from the data or the dimensionality reduction technique.
- Discuss the differences and similarities of your dimensionality reductions.

1.3 Task 3 - Clustering

(25 points)

- Train at least five cluster algorithms discussed in class on the mass cytometry dataset.
- Explain your choices of algorithms.
- For each algorithm, explain how your choice of parameters.
- Check the clustering performance using two different internal cluster validation measures (explain your choices).
- Discuss the performance (computation time, internal validation) of different methods.
- Visualize the three best clusterings using the dimensionality reduction from above. Use the coordinates of the dimensionality reduction and color points by the cluster they belong to. Use a qualitative color scale.
- Discuss how the dimensionality reduction and the clustering algorithms agree with each other.

1.4 Task 4 - External validation

(10 points)

- For each patient, calculate the cluster sizes of the best clustering.
- Visualize these cluster sizes using a simple representation learning algorithm (explain your choice).
- Use two supervised learning methods with the cluster sizes as predictors and patient group as outcome. The dataset is very small, so report on cross-validation accuracy and AUC.
- Explain your choices of algorithms.
- Explain how you performed the cross-validation step.
- Discuss how this classification validated your clustering in the previous step.