Machine Learning in Computational Biology

5/4/2022

Assignment #2

For this unsupervised learning assignment you are provided (check eclass) five single-cell synthetic datasets. Each dataset provides the gene expression profiles (200 Genes) of 200 cells. Your goal is to develop a data analysis pipeline that accepts as input a dataset file and implements the following data analysis stages in a pipeline:

1. **Dimensionality reduction** of the expression data using at least three alternative methods, that should include, but not be limited to, Principal Components Analysis (PCA), t-distributed stochastic neighbor embedding (well known as "tsnee"), and Uniform Manifold Approximation and Projection (well known as UMAP). Selecting the best number of dimensions based on some criterion is part of the needed investigation for this stage. Your choices should be justified.

2. **Clustering** of the dimensionality reduced data into the "best" number of cell "states" (clusters) using Gaussian Mixture Modeling. Your solution should be able to "discover" the optimal GMM model in terms of the number of components and their covariance matrix structure using well established model selection methods (such as the BIC criterion). At the end of this stage every cell should have a posterior distribution to each state and the number of extracted states should be optimal in some sense.

3. **Visualization** of your results (clusters inferred, cell posteriors, cell joint distributions etc.) in the most intuitive manner for a human investigator. The use of python notebooks or R markdowns that facilitate the documentation and validation of your code is highly encouraged. It is your job to select the best methods to present the data, your results and a convincing story to the evaluator.

You should submit a typed report presenting and discussing your results for every stage of your processing pipeline along with your code (R or python allowed only) in a properly organized and documented .zip file with your name and ID number. In your report you should present your pipeline and discuss each one of its stages in detail. Your pipeline implementation should be broken down to a set of well delineated function calls.

You are allowed to use any available package in R or python that you deem suitable for this assignment (e.g., *mclust*). However, you should fully justify the selection of the packages you use and of their specific parameters. You should also provide detailed instructions allowing an evaluator to easily install the dependencies and be able to run your code (act as if you are submitting a paper for review to a journal). If you have a github or similar site you can use it for this purpose (not required).

Bonus part (for up to 100% bonus points): Suggest a drastically different computational pipeline for the same data analysis (e.g. not using GMMs for producing posterior probability estimates) and compare systematically your original and the new pipeline. Also make both pipelines parametric so that it can work with a dataset of any size in terms of the number of cells and genes provided. Test both pipelines using realistic synthetic data that you will generate using GMMs with 2 to 9 states in different state “topologies”.

Deadline: You should upload your code to eclass in a well-organized .zip file by **May 2, 2022.**