Term Project Phase III: Compute the Cure

(100 points)

Assigned Date: Tuesday, April 28, 2015

Due Date:

Deadline: 3:00 PM Thursday May 5, 2015

Educational Goal

Validate the machine learning algorithms using ground truth.

Requirements

At this phase, you will validate the results of your experiments with the results reported by [Brunet 2004]. Here is everything provided for the Medulloblastoma data set and Leukemia data set used by Brunet et al. Please go to GitHub and download the nimfa-master zip file and looking within datasets. Here is example using Medulloblastoma reproducing Brunet results. Here is the Leukemia example reproducing Brunet results. The Leukemia data information which is named ALL_AML according to two major well known subtypes of leukemia. The ALL subtype also has T and B subtype within it.

List of links:

Github: https://github.com/marinkaz/nimfa

Medulloblastoma Example: http://nimfa.biolab.si/nimfa.examples.medulloblastoma.html (in Python)

Leukemia Example: http://nimfa.biolab.si/nimfa.examples.all_aml.html (in Python)

The data is located at
http://www.cs.umb.edu/~ding/classes/438_697/homework/term_project//brunet/

or in nimfa's github:

http://nimfa.biolab.si/nimfa.datasets.html

Reference paper:

[Brunet 2004] Metagenes and molecular pattern discovery using matrix factorization, Brunet et al., in Proceedings of the National Academy of Sciences of the United States of America, http://www.pnas.org/content/101/12/4164.full

[R 2010] R, see "A flexible R package for nonnegative matrix factorization", http://www.biomedcentral.com/1471-2105/11/367

1. Using the NMF code you have used in the previous phases and the Medulloblastoma data set and Leukemia data set used by [Brunet 2004], reproduce Figure 4 and Figure 6 in [Brunet 2004].
2. If you cannot reproduce the same results using your NMF implementation, please explain the reason.
3. If needed, you should re-run the experiments for Phase I.
4. How are the Medulloblastoma data set and Leukemia data set different from our Compute the Cure data set? Please explain whether such differences can impact the performance of NMF.

Submission Requirements

1. One submission per team.
2. Submit all the scripts you used for this project.
3. Prepare PPT slides in PDF file of **10 minutes presentation** to report your results in Phase III. Name the PDF of your presentation slides as **teamleadlastname_firstname_team#_ph3.pdf**.
4. Submit a single zipped file of all the files of this assignment through your UMassOnline account. Submit the paper copy including the slides and scripts. Paper copy should be bound firmly together as one pack (for example, staple, but not limited to, at the left corner). 5 points will be deducted for unbounded homework.
5. Name your file with teamleadlastname_firstname_team#_ph3. For example, team 1 of lead John Smith should name their file as Smith_John_team1_ph3.zip.
6. No hard copies or soft copies results in 0 points.