Term Project Phase I: Compute the Cure

(100 points)

Assigned Date: Tuesday, March 31, 2015

Due Date:

Deadline: ~~3:00 PM Tuesday, April 7, 2015~~

3:00 PM Thursday, April 9, 2015

Educational Goal

Become familiar with supervised and unsupervised learning using real cancer data.

Requirements

In this project, you will be using unsupervised methods to discover genetic mutation patterns for specific tumor types. All data is real and collected at Dana Farber. The nature of the data is very noisy, and finding patterns will be difficult.

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

Details:

- Each row represents a patient, and each column represents a gene.
- Matrix cells are nonzero if the column's gene contains a mutation for the row's patient.
- The exact value of each cell is the number of mutations divided by the length of the gene.
- There are all cancer patients in 5 tumor types.

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

1. For each tumor type, generate 2 figures similar as Figure 1 and Figure 4 in [Brunet 2004]:

a. Figure 1 style: Visualize A, W, and H including metagenes, samples, and metagene profile using the best K identified in our empirical study.

b. Figure 4 style: Visualize the Cophenetic Correlation plot for various K values and the Consensus Matrix of the best K.

You may want to clean the data. The suggestions in this section are purely optional, but may help you get better results.

- Remove all 0 rows and all 0 columns.

- Binarize the data. That is, make all cells either 0 or 1 (1 for nonnegative values).

- Remove columns which are not mutated more than 10% of the time.

You are free to clean the data in any additional way as you see fit.

Some suggestions but not limited to the NMF tools to use:

1. R, see "A flexible R package for nonnegative matrix factorization", http://www.biomedcentral.com/1471-2105/11/367
2. MATLAB, MATLAB with the NMF library is installed at the Web Lab. All of you can access the lab using the door code provided at UMassOnline. You may use your Windows user names and passwords to access the computers at the Web Lab.
3. We have a High Performance Computing Cluster exclusively set up for our class. We will discuss about it in class on Thursday April 2, 2015.

Submission Requirements

1. One submission per team.
2. Submit all the scripts you used for this project.
3. Prepare 5 PPT slides in PDF file to illustrate Figure 1 and Figure 4 for 5 tumor types (one tumor type/slide).
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#_ph1. For example, team 1 of lead John Smith should name their file as Smith_John_team1_ph1.zip.
6. No hard copies or soft copies results in 0 points.