Term Project Phase II: Compute the Cure

(100 points)

Assigned Date: Tuesday, April 14, 2015

Due Date:

Deadline: 3:00 PM Thursday April 23, 2015

Educational Goal

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

Requirements

At this phase, you will be using unsupervised and supervised 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

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

1. Each team should work with your partner team to do cross validation and revise your report slides for Phase I. Consistent results should be reported by the team as well as your partner team.

2. Report the classification accuracies (percentage of correct classifications).
   a. Some tumor types do not have identifiable patterns, and therefore may be indistinguishable from others using the attributes.
   b. Train and test the classifier on the entire data set.
   c. Use **logistic regression**

3. Choose a subset of the data set (subset of table row/columns) with well-defined selection criteria and report the best classification accuracy you can do at phase II.

4. Run NMF on the subset you selected in Step 3. Produce Figure 1 and Figure 4 used in [Brunet 2004] with the K you selected.

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 Phase I and Phase II. Name the PDF of your presentation slides as **teamleadlastname_firstname_team#_ph2.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#_ph2. For example, team 1 of lead John Smith should name their file as Smith_John_team1_ph2.zip.
6. No hard copies or soft copies results in 0 points.