Term Project Phase IV (the last phase): Compute the Cure

(100 points)

Assigned Date: Tuesday, April 28, 2015

Due Date:

Deadline: 3:00 PM Thursday May 21, 2015

Educational Goal

To learn the process of applying machine learning to a challenging real-world problem.

Requirements

At this phase, you will apply supervised learning and unsupervised learning to breast cancer data to identify subtypes in breast cancer data. There are four known breast cancer subtypes: Basal (Triple Negative) Breast Cancer, HER-2 Over-expression Breast Cancer, Luminal A and B Breast Cancers. You will need to check whether the cancer data we use in this project contain four different subtypes.

Unsupervised Learning. Use the stand NMF introduced by [Brunet 2004] and the Sparse NMF used by [Kim and Park 2007, 2008] to find the best K in the cancer data, respectively. You may then label the data to be used for Supervised Learning.

Discussion:

1. [Brunet 2005] uses Cophenetic Coefficient to select the best K and [Kim and Park 2007, 2008] Dispersion Coefficient to select the best K. In your opinion, which coefficient is better?
2. Design an artificial data set that is the same size of the breast cancer data. The dataset should be binary and sparse with 4 clusters. Some clusters should overlap with each other (soft clustering) and some clusters should not (hard clustering). Which NMF, standard or sparse, gives better classification result with respect to the ground truth?
3. Add different level of random noise to the artificial data with 5%, 20%, and 50% noise. Please note that the number of 1's and 0's in the dataset must be the same regardless of the noise level. Which NMF, standard or sparse, gives better classification results with respect to the ground truth?

Supervised Learning. Use logistic regression and ANN to classify the cancer data. You should use 10-fold cross validation to evaluate the classifiers and report the average prediction accuracy. Please discuss any method you have used to improve classification accuracy. Which classifier has better performance?

List of links:

The breast cancer data (cancerData.csv) is located at
http://www.cs.umb.edu/~ding/classes/438_697/homework/term_project/

Reference paper:

[Kim and Park 2008] Sparse Nonnegative Matrix Factorization for Clustering, CSE Technical Reports ; GT-CSE-08-01, Georgia Institute of Technology, http://www.cc.gatech.edu/~hpark/papers/GT-CSE-08-01.pdf

[Kim and Park 2007] Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, http://www.cs.wustl.edu/~zhang/teaching/cs517/Spring12/CourseProjects/Bioinformatics-2007-Kim-1495-502.pdf

[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

Submission Requirements

1. One submission per team.
2. Submit all the scripts you used for this project from Phase I to Phase IV.
3. Prepare PPT slides in PDF file of **30 minutes presentation** to report your results from Phase I to Phase IV. Name the PDF of your presentation slides as **teamleadlastname_firstname_team#.pdf**.
4. **The slides must demonstrate the journey of learning that includes all the experiences and lessons that you have learned in this project from Phase I to Phase IV.**
5. **Each team's presentation will be evaluated by the other teams in the class.**
6. 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.
7. Name your file with teamleadlastname_firstname_team#. For example, team 1 of lead John Smith should name their file as Smith_John_team1_ph3.zip.
8. No hard copies or soft copies results in 0 points.