Homework Assignment 3

(100 points)

Assigned Date: Tuesday, March 24, 2015

Due Date:

Intermediate report: 4:00PM Tuesday, March 31, 2015
Final Submission: ~~3:00PM Thursday, April 9, 2015~~

3:00PM Tuesday, April 14, 2015

Educational Goal

Perform **comparative studies** of two representative unsupervised learning algorithms to cluster data.

Requirements

Data downloaded at www.cs.umb.edu/~ding/classes/438_697/homework/hwk3.

Steps

1. Download homework 3 data at www.cs.umb.edu/~ding/classes/438_697/homework/hwk3

    a. Note the matrices are named in a similar style as the files we used for homework 2. Each file contains a matrix. The first character is the matrix ID. The second number is the number of clusters. The third number is the percentage of corrupted cells.
    b. (new!) The ground truth matrices, different from those you used in homework 2, have decimal values. In order to calculate the consensus matrix from a ground truth matrix, you should
        a. Scale each row of the ground truth matrix into unit vectors, you should find that the matrix consists of only ## distinct rows, where ## is the number of clusters.

b. Multiple the normalized ground truth matrix with its transpose to produce its consensus matrix like what you have done in homework 2.

2. You are required to run K-Means and Non-negative matrix factorization (NMF) on each input matrix. Both clustering algorithms are among the most popular clustering algorithms used in industry. Select whatever K (number of clusters) you think is appropriate.

3. You may use any accurate existing implementation of K-Means and NMF in this homework in any language of your choice. The R code for NMF is as follows.

The NMF code in R is:

```
install.packages('NMF')
library(NMF)
result = nmf(your_matrix, your_rank)
clustering_matrix = basis(result)
```

3. Generate a Consensus Matrix for each clustering result you have obtained and calculate classification accuracy for each matrix type at different noise levels.

4. Generate a plot of Noise Vs Accuracy with the 'plot.R' file provided in Homework 2. You should also report accuracy values.

5. Provide a brief write up of what algorithm, platform, and parameters where used order to reproduce results. **Provide a single runnable script for all the steps**.

Submission Requirements

1. **Intermediate report**: Prepare PPT slides (can be more than one slide; saved as a PDF file) to discuss
   a. Your **experiment plan** and your **first set of experimental results**

   b. Any “abnormal” or “strange” results indicate that you will need to change your experiment configuration
   c. Your **new research plan** for the final report

   Name the PDF file with your firstname_lastname.pdf and submit the PDF file to UMassOnline.
2. **Final report**: Prepare three PPT slides (saved as a PDF file) to compare the two clustering algorithms:
   a. Which clustering algorithm performs better; why?
   b. Why your experiment results are accurate and unbiased? What did you do to produce such reliable results?
   c. What is your recommendation for two clustering algorithms based on your empirical study? When should K-Means be used and when should NMF be used?
   d. Explain your journey of learning including initial plan, preliminary results, refined final plan, and final better results.
   e. Turn in the **paper copy** of the PPT report, a readme file on how to run your code, and source code in class. Zip all the files into one firstname_lastname.zip and submit the file to UMassOnline.
3. No hard copies or soft copies results in 0 points.