Homework Assignment 2

(100 points)

Assigned Date: Tuesday, February 24, 2015

Due Date:

Intermediate report: 4:00PM Tuesday, March 3, 2015
Final Submission: 3:00PM Tuesday, March 10, 2015

Educational Goal

Apply an unsupervised learning algorithm to cluster data.

Requirements

Data downloaded at www.cs.umb.edu/~henryzlo/CS438_assignment2.zip.

Steps

1. Download project data at www.cs.umb.edu/~henryzlo/CS438_assignment2.zip

   a. Note the matrices folder. Each file in this folder contains a matrix. The first character is the matrix ID. The second number is the number of classes. The third number is the percentage of corrupted cells.

2. Run an unsupervised learning algorithm of your choice on each input matrix. For example, k-means, PCA, autoencoders, RBMs, mixture models, other clustering techniques or NMF (you can even write your own!). Select whatever K (number of clusters) you think is appropriate.

   a. The output MUST be an N x K *clustering matrix*, where N is the number of rows in the input matrix, and K is the number of clusters or factors. For example, in the case of mixture models, each row corresponds to a vector of likelihoods.

b. The rows should be one-hot (only one cell is a 1, all others are 0). In the case of non-binary clustering matrices, set the highest element in each row to 1, and all other elements to 0.

3. Multiply the clustering matrix with its transpose to get the *consensus matrix*.

4. Get the ground truth of the input matrix. For example, if your current matrix is 'B_5_0.32.csv', then the ground truth is 'B_5_0.csv'. Note that each matrix ID and class number has its own ground truth which is the one with 0 percentage of corrupted cells. Multiply the ground truth matrix by its transpose to get the *ground truth consensus matrix*.

5. Compare the consensus matrix to the ground truth matrix. Calculate the *accuracy*, which is the number of matches between the two, divided by the total number of cells.

6. Write all results into a csv file, in the exact same format as 'example_results.csv'.

7. Generate a plot of Noise Vs Accuracy with the 'plot.R' file. Note that you will need to install R. Read the source code for detailed instructions.

8. Provide a brief write up of what algorithm, platform, and parameters where used order to reproduce results. **Provide a single runnable script for steps 2-6**.

Submission Requirements

1. Intermediate report: Prepare One PPT slide (saved as a PDF file) to explain the clustering algorithm your will use and demonstrate how to use the plot.R file to produce a data plot.Save the PDF file with your firstname_lastname.pdf and submit the PDF file to UMassOnline.
2. Final report: Prepare two PPT slides (saved as a PDF file) to explain the clustering algorithm you use (the algorithm could be different from your intermediate report) and the plots produced by plot.R. Explain what you have learned from the plots. 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.