CS 7641 CSE/ISYE 6740 Homework 2 Report

GTID:903070716
Liu Yujia

October 2014

Programming: Image compression [30 pts]

Task [20 pts]

Implement the algorithm and run on the toy dataset `data.mat`. You can find detailed description about the data in the `homework2.m` file. Observe the results and compare them with the provided true clusters each document belongs to. Report the evaluation (e.g. accuracy) of your implementation.

Hint: We already did the word counting for you, so the data file only contains a count matrix like the one shown above. For the toy dataset, set the number of clusters $n_c = 4$. You will need to initialize the parameters. Try several different random initial values for the probability of a word being W_j in topic c, μ_{jc}. Make sure you normalized it. Make sure that you should not use the true cluster information during your learning phase.

Solution:

After running the algorithm in row for 20 times, the result is listed as below. It consists of the accuracy and elapsed time of each run.

Accuracy	Elapsed Time
78.5000	0.3764
64.2500	0.3581
89.7500	0.3041
75.2500	0.3475
85.5000	0.3621
80.2500	0.3445
69.2500	0.3466
70.5000	0.3540
87.5000	0.3702
76.0000	0.3486
87.5000	0.3523
74.2500	0.3483
74.7500	0.3560
82.0000	0.3497
67.5000	0.3459
79.0000	0.3548
79.7500	0.3518
68.2500	0.3519
54.0000	0.3581
79.0000	0.2087

We know from the table that the maximum accuracy is 89.7500, the minimum is 54.0000 and the average

is 76.1375. The maximum elapsed time is 0.3764, the minimum is 0.2087 and the average is 0.3445. This accuracy performance is acceptable and the elapsed time is comparatively short. We can decrease the elapsed time by modifying the constraint of iterations(number of iterations) to a certain degree that the accuracy performance is not going to be affected.