

CS550: Massive Data Mining and Learning  
Problem Set 2  
Due 11:59pm Monday, March 23, 2019

Spring 2020

Only one late period is allowed for this homework (11:59pm Tuesday 3/24)

### Submission Instructions

**Assignment Submission:** Include a signed agreement to the Honor Code with this assignment. Assignments are due at 11:59pm. All students must submit their homework via Sakai. Students can typeset or scan their homework. Students also need to include their code in the final submission zip file. Put all the code for a single question into a single file.

**Late Day Policy:** Each student will have a total of **two** free late days, and for each homework only one late day can be used. If a late day is used, the due date is 11:59pm on the next day.

**Honor Code:** Students may discuss and work on homework problems in groups. This is encouraged. However, each student must write down their solutions independently to show they understand the solution well enough in order to reconstruct it by themselves. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code seriously and expect students to do the same.

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

[https://sites.cs.ucsb.edu/~veronika/MAE/parallelkmeansmapreduce\\_zhao.pdf](https://sites.cs.ucsb.edu/~veronika/MAE/parallelkmeansmapreduce_zhao.pdf)  
[https://www.tutorialspoint.com/jfreechart/jfreechart\\_line\\_chart.htm](https://www.tutorialspoint.com/jfreechart/jfreechart_line_chart.htm)

I acknowledge and accept the Honor Code.

(Signed) \_\_\_\_\_ Fatima AlSaadeh \_\_\_\_\_

If you are not printing this document out, please type your initials above.

Answer to Question 1(a)

- a)  $MM^T$ ,  $M^TM$  symmetric, square and real.
- $M_{p \times q}$  P: data points  
q: dimension
- \*  $(MM^T)^T = M^T(M^T)^T$  matrix is symmetric  
if  $A^T = A$
  - $= M^T M$
  - $(M^T M)^T = M^T M^T$  both are symmetric
  - \* if  $M$  has dimensions  $p \times q$  then  $M^T = q \times p$ , therefore  $MM^T_{p \times q \times q \times p} = p \times p$  which is square
  - same for  $M^TM \Rightarrow M^T_{q \times p} \times M_{q \times p} \times M^T_{p \times q} = p \times q$  which is square
  - \* if  $MM^T$ ,  $M^TM$  are symmetric then it will have real eigenvalues and therefore its real.

Answer to Question 1(b)

- b) eigen values of  $MM^T$ ,  $M^TM$ , are the eigen vectors same?
- \* since  $(MM^T)^T = (M^TM)$
  - the matrix and its transpose have same eigen values:
- $$\det |MM^T - \lambda I| = \det |(MM^T)^T - \lambda I|$$
- $$\det |MM^T - \lambda I| = \det |(MM^T)^T|$$
- $$\Leftrightarrow \det |(MM^T - \lambda I)^T| = \det |(MM^T)^T - \lambda I| \Rightarrow I^T = I$$
- $$\det |(MM^T)^T - \lambda I| = \det |MM^T - \lambda I| \rightarrow \text{same eigenvalues.}$$
- \*  $MM^T$  and  $M^TM$  don't have the same eigen vectors.

Answer to Question 1(c)

$$c) \text{ If } B_{d \times d} = Q \Lambda Q^T \\ (M^T M)_{q \times q} = Q \Lambda Q^T$$

Answer to Question 1(d)

$$d) M = U \Sigma V^T \quad U^T U = I \\ M^T M = (U^T \Sigma^T V) U \Sigma V^T \\ = V (\Sigma^T \Sigma) V^T \quad \Sigma = \Sigma^T \\ = V \Sigma^2 V^T$$

Answer to Question 1(e)(a)

$$U = [ \begin{matrix} -0.27854301 & 0.5 \\ -0.27854301 & -0.5 \\ -0.64993368 & 0.5 \\ -0.64993368 & -0.5 \end{matrix} ]$$

$$\begin{aligned} \text{Sigma} &= [7.61577311 \ 1.41421356] \\ V^T &= [ \begin{matrix} -0.70710678 & -0.70710678 \\ -0.70710678 & 0.70710678 \end{matrix} ] \end{aligned}$$

Answer to Question 1(e)(b)

Sorted Evals=

$$[58. \ 2.]$$

Sorted Evecs=

$$[ \begin{matrix} 0.70710678 & -0.70710678 \\ 0.70710678 & 0.70710678 \end{matrix} ]$$

Answer to Question 1(e)(c)

$V = [$

$[-0.70710678 \ -0.70710678]$

$[-0.70710678 \ 0.70710678]$

$]$

In  $V$  the columns are the eigenvectors

first column is the eigen vector for eigen value 58 multiplied by -1

second column is the eigen vector for eigen value 2

Answer to Question 1(e)(d)

sort\_Evals = Sigma^2 = [58. 2.]

Answer to Question 2(a)

Question 2:  $M_{n \times n}$   $m_{ij} \xrightarrow{O: \text{no link from } i \rightarrow j}$   
 $\xrightarrow{k: \text{k# of arcs out of } j}$   
 $\hookrightarrow$  column  $j$  is all 0's if node  $j$  deadened  
 $\hookrightarrow$  column  $j$  has  $\frac{1}{k}$  #arcs and others are 0

$r = [r_1, r_2, \dots, r_n]^T \rightarrow$  PageRank vector  
 $r_i \rightarrow$  estimate PageRank of node  $i$

$w(r) = \sum_{i=1}^n r_i$

Next iteration:  $r' = Mr$   $r'_i = \sum_{j=1}^n M_{ij} r_j$

(a) If no deadends Prove  $w(r') = w(r)$

$w(r') = \sum_{i=1}^n \left( \sum_{j=1}^n M_{ij} r_j \right)$

$w(r') = w(r) ? \quad \sum_{i=1}^n \left( \sum_{j=1}^n m_{ij} r_j \right) = \sum_{i=1}^n r_i$

\* If no deadends  $\frac{1}{k} + k = 1 = \sum_{i=1}^n M_{ij}$

$\sum_{j=1}^n r_j = \sum_{i=1}^n r_i \quad \checkmark$

Answer to Question 2(b)

(b) no descendants,  $0 < \beta < 1$

$$r_i^* = \beta \sum_{j=1}^n m_{ij} r_j + (1-\beta)/n$$

when  $w(r^*) = w(r)$

$$\begin{aligned} w(r^*) &= \sum_{j=1}^n \beta \sum_{i=1}^n m_{ij} r_j + \sum_{j=1}^n (1-\beta) \\ &= \sum_{j=1}^n \beta (\sum_{i=1}^n m_{ij}) r_j + \sum_{j=1}^n (1-\beta) \\ &= \sum_{j=1}^n \beta w(r) + (1-\beta) \frac{n}{n} = \beta w(r) + 1 - \beta \\ \text{if } w(r) &= \beta w(r) + 1 - \beta \\ w(r)(1-\beta) &= 1 - \beta \\ w(r) &= 1 \rightarrow \text{if } w(r) = 1 \text{ then } w(r) = w(r^*) \end{aligned}$$

Answer to Question 2(c)(a)

- ①  $r_i^* = (1-\beta)r_j$  if  $j$  alive  
 2.  $r_i^*$  if  $j$  dead.

$$r_i^* = \beta \sum_{j=1}^n M_{ij} r_j + \frac{(1-\beta)}{n} + \frac{\beta}{n} \sum_{j \in \text{dead}} r_j$$

Answer to Question 2(c)(b)

$$\begin{aligned} w(r^*) &= \sum_{i=1}^n \left( \beta \sum_{j=1}^n M_{ij} r_j + \frac{(1-\beta)}{n} + \frac{\beta}{n} \sum_{j \in \text{dead}} r_j \right) \\ &= \sum_{j=1}^n \beta r_j + (1-\beta) + \beta \sum_{j \in \text{dead}} r_j \quad \sum_{j \in \text{dead}} r_j = \sum_{j=1}^n r_j - \sum_{j \in \text{alive}} r_j \\ &= \beta \sum_{j \in \text{alive}} r_j + (1-\beta) + \beta \sum_{j \in \text{dead}} r_j \\ &= \beta \sum_{j=1}^n r_j + (1-\beta) = \beta w(r) + 1 - \beta \quad \text{if } w(r) = 1 \text{ --- given} \end{aligned}$$

then  $\beta + 1 - \beta = 1$

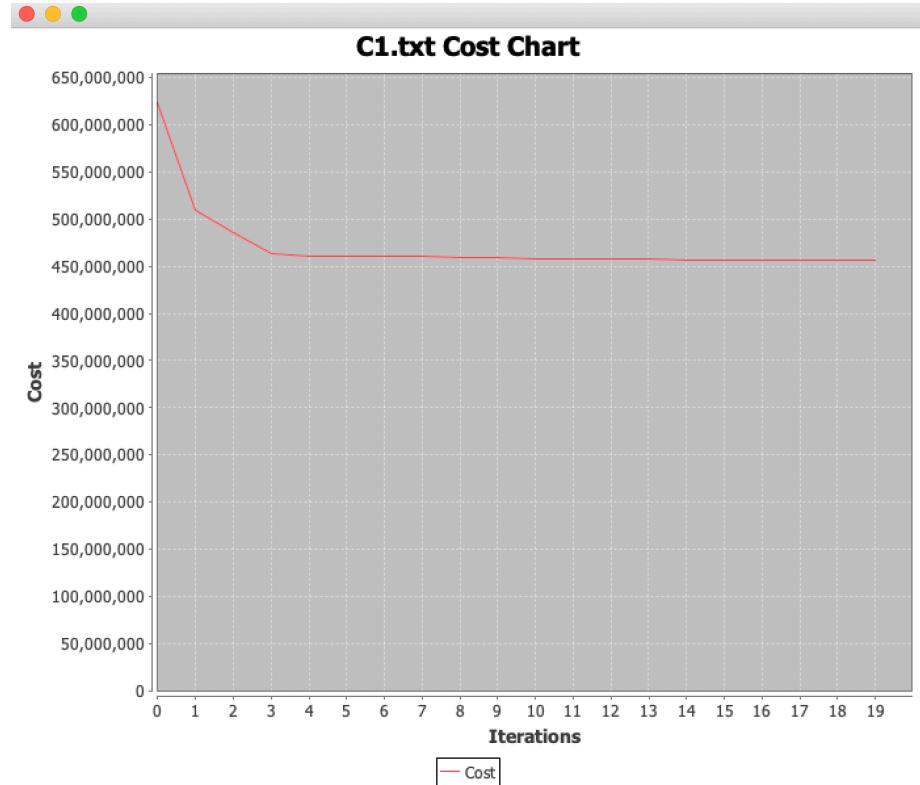
Answer to Question 3(a)

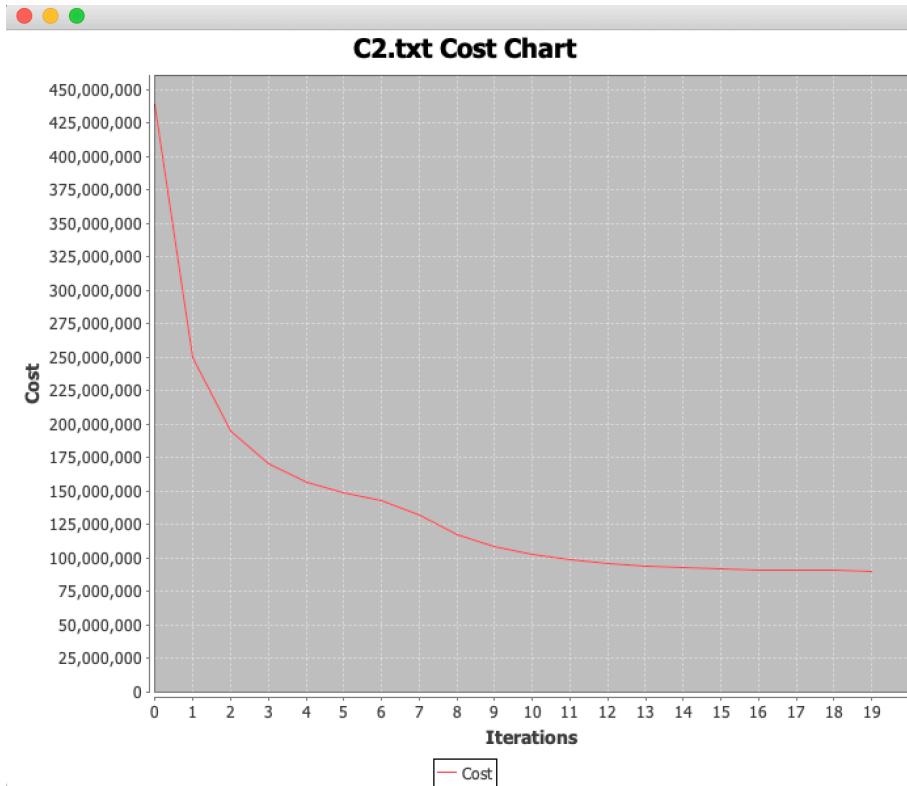
|    |                      |
|----|----------------------|
| 53 | 0.0357312022326716   |
| 14 | 0.034170906972591376 |
| 40 | 0.0336300871897439   |
| 1  | 0.030005979479788617 |
| 27 | 0.02972014420140539  |

Answer to Question 3(b)

|    |                       |
|----|-----------------------|
| 85 | 0.0034096940774028216 |
| 59 | 0.003669860660127284  |
| 81 | 0.0036953517493609916 |
| 37 | 0.003808204291611451  |
| 89 | 0.003922466019802268  |

Answer to Question 4(a)





Answer to Question 4(b)

c1.txt:

percentage change in cost after 10 iterations: 26.398863292044183%

c2.txt :

percentage change in cost after 10 iterations: 75.25973243724756%

Random initialization isn't better as picking the centroids which are as far apart as possible is a better technique because it'll give us a better insight on where the points should be clustered resulting a minimum cost.