CS550: Massive Data Mining and Learning
Homework 2

Due 11:59pm Monday, March 23, 2020

Only one late period is allowed for this homework

Submitted by:

Name: Prakruti Joshi

NetID: phj15

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:

I acknowledge and accept the Honor Code.

(Signed) Prakruti Joshi

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

Answer to Question 1(a)

Yes, the matrices MM^T and M^TM are symmetric, real and square.

- **Symmetric**: A matrix is symmetric if the transpose of the matrix X^T is the same as the original matrix X.

$$(MM^T)^T = (M^T)^T M^T = MM^T$$
$$(M^TM)^T = M^T(M^T)^T = M^TM$$

  Thus, matrices MM^T and M^TM are symmetric.

- **Square**: M has dimensions are $(p \times q)$. Thus, the dimensions of MM^T are $(p \times p)$. The dimensions of M^TM are $(q \times q)$. Thus MM^T and M^TM are square matrices.

- **Real**: Since M is a real matrix, M^T is also a real matrix. Thus, the multiplication of M and M^T is also a real matrix.

Answer to Question 1(b)

Let the eigenvector of MM^T be x and the eigenvalue be λ. Thus,

$$MM^T(x) = \lambda x$$

Multiplying both sides by M^T,

$$M^TMM^T(x) = M^T\lambda x$$
$$M^TM(M^Tx) = \lambda(M^Tx)$$

Thus, the eigenvector of M^TM is (M^Tx) and the eigenvalue is λ, which is same as eigenvalue of MM^T.

Answer to Question 1(c)

Since M^TM is a real, symmetric and square matrix; its eigenvalue decomposition can be written in terms of Q and Λ as follows:

$$M^TM = Q\Lambda Q^T$$

Answer to Question 1(d)

$$M = U\Sigma V^T$$
$$M^TM = (U\Sigma V^T)^T(U\Sigma V^T)$$
$$M^TM = (V^T)^T(\Sigma)^T U^T U\Sigma V^T$$

Now, since U is column orthonormal matrix, $U^TU = I$.

$$M^TM = (V^T)^T(\Sigma)^T\Sigma V^T$$

Also, since Σ is a diagonal matrix, $(\Sigma)^T\Sigma$ is simply Σ^2. Therefore,

$$M^TM = V\Sigma^2V^T$$

Answer to Question 1(e)(a)

$$U = \begin{bmatrix} -0.27854301 & 0.5 \\ -0.27854301 & -0.5 \\ -0.64993368 & 0.5 \\ -0.64993368 & -0.5 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 7.61577311 & 1.41421356 \end{bmatrix}$$

$$V^T = \begin{bmatrix} 0.70710678 & 0.70710678 \\ -0.70710678 & 0.70710678 \end{bmatrix}$$

Answer to Question 1(e)(b)

$$evals = \begin{bmatrix} 2. & 58. \end{bmatrix}$$

$$evecs = \begin{bmatrix} -0.70710678 & 0.70710678 \\ 0.70710678 & 0.70710678 \end{bmatrix}$$

After reordering:

$$evals = \begin{bmatrix} 58. & 2. \end{bmatrix}$$

$$evecs = \begin{bmatrix} 0.70710678 & -0.70710678 \\ 0.70710678 & 0.70710678 \end{bmatrix}$$

Answer to Question 1(e)(c)

V matrix produced by the SVD is equal to the matrix of eigenvectors after reordering the column values according to the order of the singular values.

Answer to Question 1(e)(d)

The singular values of M are the square roots of the eigenvalues of $M^T M$.

Answer to Question 2(a)

$$w(r) = \sum_{j=1}^{n} r_j$$

Since the web has no dead ends, $\sum_{i=1}^{n} M_{ij} = 1$ for each j. That is the sum of the each column is 1. Now,

$$\begin{aligned}
r_i^{'} &= \sum_{j=1}^{n} M_{ij} r_j \\
\implies w(r^{'}) &= \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j \\
\implies w(r^{'}) &= \sum_{j=1}^{n} (\sum_{i=1}^{n} M_{ij}) r_j \\
\implies w(r^{'}) &= \sum_{j=1}^{n} r_j \qquad (\because \sum_{i=1}^{n} M_{ij} = 1 \quad \forall j) \\
\implies w(r^{'}) &= w(r)
\end{aligned}$$

Hence proved.

Answer to Question 2(b)

The web has no dead pages. Therefore, $\sum_{i=1}^{n} M_{ij} = 1$ for each j. The teleportation probability is $1 - \beta$. Thus,

$$
\begin{aligned}
r_i^{'} &= \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{(1-\beta)}{n} \\
\therefore w(r^{'}) &= \sum_{i=1}^{n} (\beta \sum_{j=1}^{n} M_{ij} r_j + \frac{(1-\beta)}{n}) \\
\implies w(r^{'}) &= \beta \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j + \sum_{i=1}^{n} \frac{(1-\beta)}{n} \\
\implies w(r^{'}) &= \beta \sum_{j=1}^{n} (\sum_{i=1}^{n} M_{ij}) r_j + \frac{n(1-\beta)}{n} \\
\implies w(r^{'}) &= \beta \sum_{j=1}^{n} r_j + (1-\beta) \qquad (\because \sum_{i=1}^{n} M_{ij} = 1 \quad \forall j) \\
\implies w(r^{'}) &= \beta w(r) + (1-\beta) \qquad (\because \sum_{i=1}^{n} r_j = w(r))
\end{aligned}
$$

Therefore, for $w(r^{'}) = w(r)$,

$$
\begin{aligned}
w(r) &= \beta w(r) + (1-\beta) \\
\therefore (1-\beta) w(r) &= (1-\beta) \\
\implies w(r) &= 1
\end{aligned}
$$

Thus, the condition for $w(r^{'}) = w(r)$ is w(r) = 1.

Answer to Question 2(c)(a)

The equation for $r_i^{'}$ in terms of β, M, and r:

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

Now adding and subtracting $\frac{\sum_{j \in dead} \beta r_j}{n}$ on the left side of the equation, we get:

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

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

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

$$\implies r_i^{'} = \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{1-\beta}{n} \sum_{j=1}^{n} r_j + \frac{\sum_{j \in dead} \beta r_j}{n}$$

$$\implies r_i^{'} = \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{1-\beta}{n} w(r) + \frac{\sum_{j \in dead} \beta r_j}{n}$$

$$\implies r_i^{'} = \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{1-\beta}{n}(1) + \frac{\sum_{j \in dead} \beta r_j}{n} \qquad (\because w(r) = 1)$$

$$\boxed{\therefore r_i^{'} = \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{1-\beta}{n} + \frac{\beta \sum_{j \in dead} r_j}{n}}$$

Answer to Question 2(c)(b)

To prove: $w(r^{'}) = 1$ given that w(r) = 1
Solution: From 2(c)(a),

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

Therefore,

$$w(r^{'}) = \sum_{i=1}^{n}(\beta \sum_{j=1}^{n} M_{ij} r_j) + \sum_{i=1}^{n} \frac{1-\beta}{n} + \sum_{i=1}^{n} \frac{\sum_{j \in dead} \beta r_j}{n}$$

$$\implies w(r^{'}) = \beta \sum_{j=1}^{n}(\sum_{i=1}^{n} M_{ij}) r_j + \frac{n(1-\beta)}{n} + n.\frac{\sum_{j \in dead} \beta r_j}{n}$$

Now, $\sum_{i=1}^{n} M_{ij} = 1 \ \ \forall j \in live$. And $\sum_{i=1}^{n} M_{ij} = 0 \ \ \forall j \in dead$.

$$\begin{aligned}
&\implies w(r^{'}) = \beta \sum_{j \in live} (1) r_j + (1-\beta) + \beta \sum_{j \in dead} r_j \\
&\implies w(r^{'}) = \beta (\sum_{j \in live} r_j + \sum_{j \in dead} r_j) + (1-\beta) \\
&\implies w(r^{'}) = \beta \sum_{j=1}^{n} r_j + (1-\beta) \\
&\implies w(r^{'}) = \beta w(r) + (1-\beta) \\
&\implies w(r^{'}) = \beta + (1-\beta) \qquad (\because w(r) = 1) \\
&\implies w(r^{'}) = 1
\end{aligned}$$

Hence proved.

Answer to Question 3(a)

Nodes with Highest page rank scores and their ids: (Top 5)

Node ID	PageRank Score
53	0.037868613328747594
14	0.035866772133529436
1	0.03514138301760087
40	0.03383064398237689
27	0.03313019554724851

Answer to Question 3(b)

Nodes with Lowest page rank scores and their ids: (Bottom 5)

Node ID	PageRank Score
85	0.003234819143382019
59	0.003444256201194502
81	0.003580432413995564
37	0.003714283971941924
89	0.0038398576156450873

Answer to Question 4(a)

Cost v/s iteration plot for the two initialization strategies:

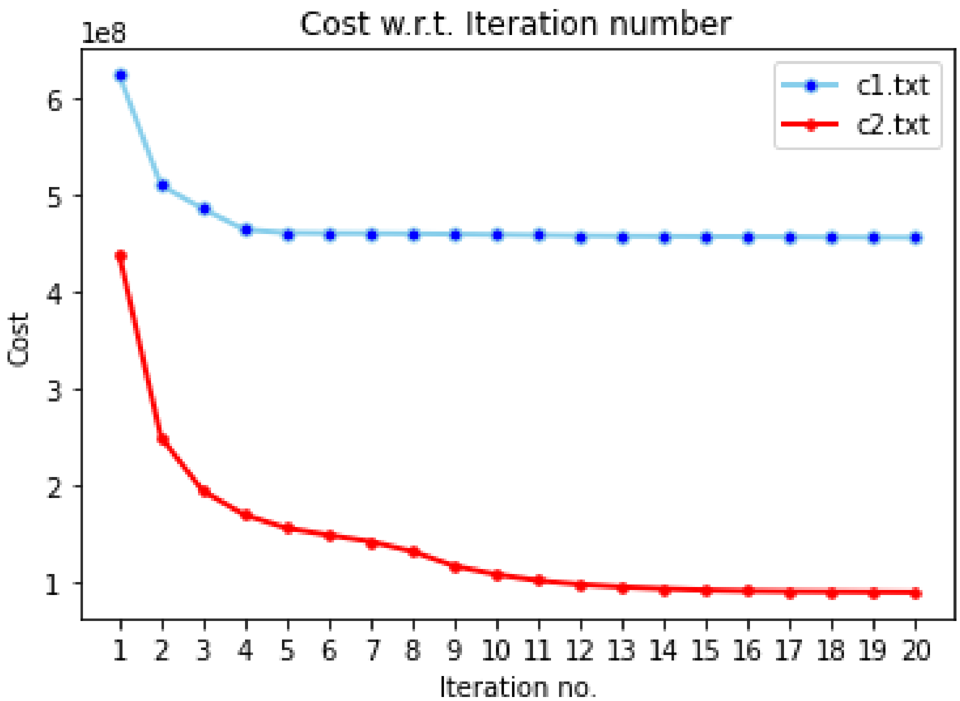

Figure 1:

Answer to Question 4(b)

Iteration number	c1	c2
1	6.2366×10^8	4.38747×10^8
10	4.59021×10^8	1.08547×10^8
20	4.559868×10^8	9.02164×10^7

Random initialization using c1 improves the cost around **26.40 %** in the first 10 iterations. Whereas, initialization using c2 improves the cost around **75.26 %**
Initialization using c2.txt is better than random initialization using c1.txt as can be seen of the decrease in the cost function over various iterations. The reasoning for this is that c2 initializes the centroids of the clusters as far as possible from each other. Thus, it leads to less overlap between the clusters when the points are assigned to the cluster. The inter-cluster distances are maximized and the intra-cluster distances are minimized. Thus the value of the cost function is reduced and it leads to better assignment of cluster centroids as compared to random initialization. Thus, true clusters are split less in this case which leads to overall better clusters.