CS550: Massive Data Mining and Learning
Homework 2

Twisha Gaurang Naik (tn268)

Due 11:59pm Monday, March 23, 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:

I acknowledge and accept the Honor Code.

(Signed) Twisha Gaurang Naik (tn268)

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, square and real.
Explanations for each of these are as follows:

- **Symmetric**: Matrices are symmetric if $A = A^T$
  $(MM^T)^T = (M^T)^TM^T = MM^T$
  Similarly, $(M^TM)^T = M^T(M^T)^T = M^TM$
- **Square**: Here, dim of M = p x q and dim of M^T = q x p
  Dim of MM^T = (p x q) x (q x p) = p x p
  Dim of M^TM = (q x p) x (p x q) = q x q
- **Real**: Given that matrix M is real, the product of M and its transpose will also be real.

Answer to Question 1(b)

Let x be the eigenvector and λ be the corresponding eigenvalue of matrix MM^T.

$$\begin{aligned} MM^T(x) &= \lambda(x) \\ \implies M^T(MM^T)(x) &= M^T\lambda(x) \\ \implies M^TM(M^Tx) &= \lambda(M^Tx) \end{aligned}$$

Eigenvalue of $M^TM = \lambda$, but the eigenvector = M^Te.
Thus, MM^T and M^TM have the **same eigenvalues** but **different eigenvectors**.

Answer to Question 1(c)

A real, symmetric and square matrix B, can be decomposed in the following way:
$B = Q \wedge Q^T$
From question 1a, the matrix M^TM also has these properties.

Thus, $\boxed{M^TM = Q \wedge Q^T}$

Answer to Question 1(d)

$$\begin{aligned} M &= U\Sigma V^T \\ \therefore M^T M &= (U\Sigma V^T)^T (U\Sigma V^T) \\ &= (V\Sigma^T U^T)(U\Sigma V^T) \\ &= V\Sigma^T (U^T U)\Sigma V^T \\ &= V\Sigma^T I \Sigma V^T \end{aligned}$$

As Σ is a diagonal matrix, $\Sigma^T = \Sigma$. Thus, we get the following equation:

$$\boxed{M^T M = V\Sigma^2 V^T}$$

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)

The eigenvalues and vectors sorted in the decresing order of eigenvalues are as follows:
$Evals = \begin{bmatrix} 58. & 2. \end{bmatrix}$

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

Answer to Question 1(e)(c)

V and Evecs (after reordering based on the eigenvalues) are the same as it can be observed from the values from part (a) and (b).

Answer to Question 1(e)(d)

Singular values of M^TM are square of the singular values of M.

Theoretical Proof: From question 1d, $M = U\Sigma V^T$ and $M^TM = V\Sigma^2V^T$. Observing these equations, it is clear that singular values get squared for M^TM.
Empirical Proof: From the values in parts (a) and (b),
$58 = (7.61577311)^2$
$2 = (1.41421356)^2$.

Answer to Question 2(a)

Definition: **w(r)** = The sum of the components of r.

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

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

$$r' = Mr$$
$$\implies r_i^{'} = \sum_{j=1}^{n} M_{ij} r_j$$

Now, taking the sum of all the elements of r',

$$w(r^{'}) = \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} r_j$$
$$\implies w(r^{'}) = \sum_{j=1}^{n}(\sum_{i=1}^{n} M_{ij}) r_j$$
$$\implies w(r^{'}) = \sum_{j=1}^{n} r_j \qquad (\because \sum_{i=1}^{n} M_{ij} = 1 \;\; ;\forall j)$$
$$\implies w(r^{'}) = w(r)$$

Answer to Question 2(b)

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

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

Taking summation of all terms of r',

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

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

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

$$\implies w(r^{'}) = \beta \sum_{j=1}^{n} r_j + (1-\beta) \qquad (\because \sum_{i=1}^{n} M_{ij} = 1; \quad \forall j)$$

$$\implies w(r^{'}) = \beta w(r) + (1-\beta) \qquad (\because \sum_{i=1}^{n} r_j = w(r))$$

Let, $w(r^{'}) = w(r) = w$,

$$w = \beta w + (1-\beta)$$
$$\therefore (1-\beta) w = (1-\beta)$$
$$\implies w = 1$$

The condition is: $w(r^{'}) = w(r) = 1$.

Answer to Question 2(c)(a)

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

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

Splitting the term $\sum_{j \in dead} \frac{r_j}{n}$ into $\beta \cdot \sum_{j \in dead} \frac{r_j}{n}$ and $(1-\beta) \cdot \sum_{j \in dead} \frac{r_j}{n}$ in the equation,

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

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

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

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

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

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

Intuitively, the last term added in the rank makes up for the score lost due to dead ends by distributing it equally to all the nodes.

Answer to Question 2(c)(b)

Given, w(r) = 1.
From 2(c)(a),

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

Summing all the terms of $r^{'}$,

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

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

We know, $\sum_{i=1}^{n} M_{ij} = 1; \quad \forall j \in live.$
Also, $\sum_{i=1}^{n} M_{ij} = 0; \quad \forall j \in dead$ (as there are no outgoing links).

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

Answer to Question 3(a)

The top 5 nodes with highest page rank scores are:

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

Answer to Question 3(b)

The bottom 5 nodes with lowest page rank scores are:

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

Answer to Question 4(a)

Using the Euclidean distance as the distance measure, the cost function $\phi(i)$ is computed for every iteration i. Plot of the cost function $\phi(i)$ as a function of the number of iterations.

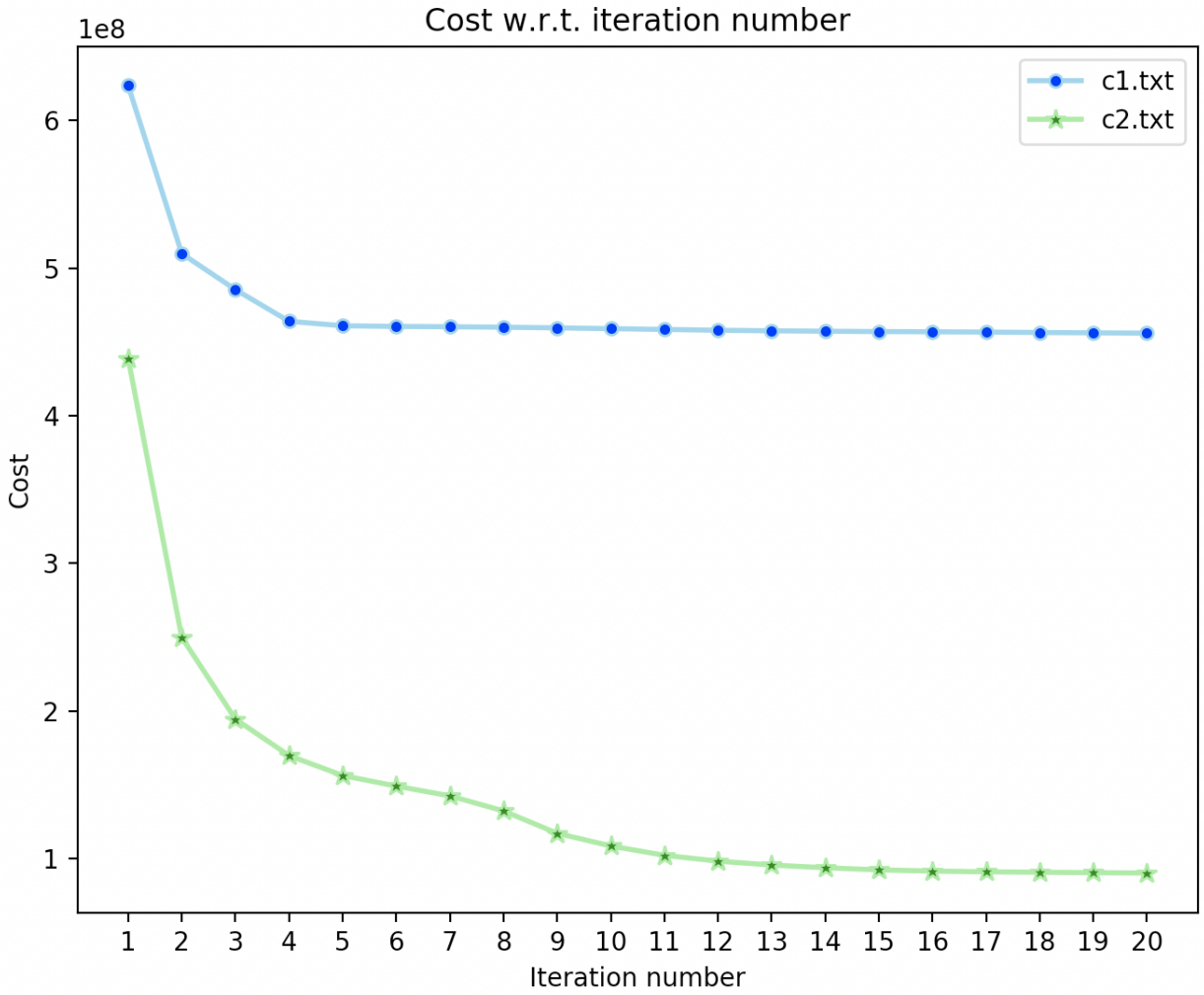

Figure 1: Cost vs iteration for two different centroid initialization

Answer to Question 4(b)

Percentage change in cost after 10 iterations of the k-Means algorithm when the cluster centroids are initialized using:

1. **c1.txt**
   Iteration 1 cost = $6.236603453064234 \times 10^8$
   Iteration 10 cost = $4.590211033422901 \times 10^8$
   Percentage change = 26.39886%

2. **c2.txt**
   Iteration 1 cost = $4.38747790027918 \times 10^8$
   Iteration 10 cost = $1.0854737717857017 \times 10^8$
   Percentage change = 75.25973%

Initialization using c2.txt is much better than that of c1.txt. The centroids chosen in c1.txt are quite random. In fact two of the centroids are the same. Whereas in c2.txt, the centroids are spread out in the given vector space. If observed carefully, value of at least one of the 58 dimension is quite high.
As the centroids are spread far apart in c2.txt, it proves to be a better initialization.