CS550: Massive Data Mining and Learning Spring 2022

Problem Set 2

Due 11:59pm Monday, March 21, 2022

Submission instructions: These questions require thought but do not require long answers. Please be as concise as possible. You should submit your answers as a writeup in PDF format, for those questions that require coding, write your code for a question in a single source code file, and name the file as the question number (e.g., question_1.java or question_1.py), finally, put your PDF answer file and all the code files in a folder named as your Name and NetID (i.e., Firstname-Lastname-NetID.pdf), compress the folder as a zip file (e.g., Firstname-Lastname-NetID.zip), and submit the zip file via Canvas.

For the answer writeup PDF file, we have provided both a word template and a latex template for you, after you finished the writing, save the file as a PDF file, and submit both the original file (word or latex) and the PDF file.

Late Policy: The homework is due on 3/21 (Monday) at 11:59pm. We will release the solutions of the homework on Canvas on 3/25 (Friday) 11:59pm. If your homework is submitted to Canvas before 3/21 11:59pm, there will no late penalty. If you submit to Canvas after 3/21 11:59pm and before 3/25 11:59pm (i.e., before we release the solution), your score will be penalized by 0.9^k, where k is the number of days of late submission. For example, if you submitted on 3/24, and your original score is 80, then your final score will be 80* 0.9^3=58.32 for 24-21=3 days of late submission. If you submit to Canvas after 3/25 11:59pm (i.e., after we release the solution), then you will earn no score for the homework.

Honor Code: Students may discuss the homework problems with peers. 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 directly obtained from the web or others 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)_____KK____

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

Answer to Question 1(a)

The matrix MM^T, M^TM are symmetric wrt each other, square and real.

Test of Symmetry: A = A^T : $LHS = MM^T$; $RHS = \left(MM^T\right)^T$ => M^TM

So MM^T, M^TM are symmetric w.r.t. each other.

Square: Size of matrix M: $p\ x\ q$ and size of M^T: $q\ x\ p$. So size of MM^T: $(p\ x\ q)\ x\ (q\ x\ p) = p\ x\ p$ =>this is a square matrix. Similarly, size of M^TM : $(q\ x\ p)\ x\ (p\ x\ q) = q\ x\ q$ => this is a square matrix as well.

Real: Since $M,\ M^T$ are real, their multiplication will also contain real numbers.

Answer to Question 1(b)

Let v be eigen vectors and λ be eigenvalues of M^TM

So, $\left(M^TM - \lambda I\right)v = 0$ – eq(1)

Now we multiply with M on both sides of eq(1), we get

$M\left(M^TM - \lambda I\right)v = M\ x\ 0$

=> $\left(MM^TM - \lambda MI\right)v = 0$

=>$\left(MM^T - \lambda I\right)Mv = 0.$

We can see here that Mv is the eigenvector for MM^T and λ is the eigenvalue.

So M^TM and MM^Thave the same eigenvalues but different eigenvectors.

Answer to Question 1(c)

Eigenvalue decomposition of a real symmetric and square matrix B is given as:

$$B = Q\Lambda Q^T$$

Since we already proved that M^TM is a real, symmetric and square matrix, we can write it eigenvalue decomposition in a similar way.

$$M^TM = Q\Lambda Q^T \quad \text{– eq(2)}$$

Answer to Question 1(d)

Given $M = U\Sigma V^T$

$M^T = \left(U\Sigma V^T\right)^T$ => $V\Sigma^TU^T$

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

$= V\Sigma^T\Sigma V^T$ (because U is column orthonormal, $U^TU = I$)

We know Σ is a diagonal matrix so $\Sigma^T = \Sigma$

$= V\Sigma^2 V^T$

Finally, $M^T M = V\Sigma^2 V^T$ – eq(3)

Answer to Question 1(e)(a)

Result of

```
U,Sigma,VT = linalg.svd(M, full_matrices = False)
```

Is

U:
[[-0.27854301   0.5]
 [-0.27854301 -0.5]
 [-0.64993368   0.5]
 [-0.64993368 -0.5]]

$Sigma$ (Σ):
 [7.61577311 1.41421356]

V^T:
 [[-0.70710678 -0.70710678]
 [-0.70710678 0.70710678]]

Answer to Question 1(e)(b)

Evals and Evecs after rearrangement:

Evals:
 [58. 2.]

Evecs:
 [[0.70710678 -0.70710678]
 [0.70710678 0.70710678]]

Answer to Question 1(e)(c)

$V$is equivalent to the matrix of eigenvectors(Evecs) if we reorder the columns as per the ordering of the singular values.

Answer to Question 1(e)(d)

From the results of 1(c) and 1(d) i.e., eq(2) and eq(3), we can equate both the results and get:

$Q\Lambda Q^T = V\Sigma^2 V^T$

From this we can see that Q=V and $\Lambda = \Sigma^2$
We know

Σ = [7.61577311 1.41421356]

Σ^2 = [58. 2.] = Λ

Answer to Question 2(a)

Given there are no dead ends, we need to show $w(r') = w(r)$

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

It is given that $r'_i = \sum_{j=1}^{n} M_{ij} r_j$

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

We know that summation over i for M_{ij} is summing the column vectors of M, which is equal to 1 because there are no dead-ends.

So $w(r') = \sum_{j=1}^{n} r_j$ which is equivalent to $\sum_{i=1}^{n} r'_i$

Hence proved.

Answer to Question 2(b)

Given $r'_i = \beta \sum_{j=1}^{n} M_{ij} r_j + (1 - \beta)/n$

We first calculate $w(r')$

=> $w(r') = \sum_{i=1}^{n} r'_i$

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

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

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

=>$w(r') = \beta w(r) + (1 - \beta)$

We need to find the conditions in which $w(r') = w(r)$

$w(r) = \beta w(r) + (1 - \beta)$

=> $w(r) = 1$

So, if $w(r) = 1$, $w(r)$ and $w(r')$ will be equal.

Answer to Question 2(c)(a)

We need to get equation of r'_i in terms of β, M, r

$$r'_i = \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{(1-\beta)}{n} \sum_{j \epsilon live} r_j + \frac{1}{n} \sum_{j \epsilon dead} r_j$$
$$= \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{(1-\beta)}{n} \sum_{j \epsilon live} r_j + \frac{(1-\beta)+\beta}{n} \sum_{j \epsilon dead} r_j$$
$$= \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{(1-\beta)}{n} \left(\sum_{j \epsilon live} r_j + \sum_{j \epsilon dead} r_j \right) + \frac{\beta}{n} \sum_{j \epsilon dead} r_j$$
$$= \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{(1-\beta)}{n} \sum_{j} r_j + \frac{\beta}{n} \sum_{j \epsilon dead} r_j$$
It is given that $w(r) = 1 = \sum_{j=1}^{n} r_j$

Finally,
$$r'_i = \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{(1-\beta)}{n} + \frac{\beta}{n} \sum_{j \epsilon dead} r_j \quad -$$
eq(4)

Answer to Question 2(c)(b)

Given the previous result, we need to show that $w(r')$ is also 1.

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

From eq(4) we get:
$$= \sum_{i=1}^{n} \left(\beta \sum_{j=1}^{n} M_{ij} r_j + \frac{(1-\beta)}{n} + \frac{\beta}{n} \sum_{j \epsilon dead} r_j \right)$$
$$= \beta \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j + \sum_{i=1}^{n} \frac{(1-\beta)}{n} + \frac{\beta}{n} \sum_{i=1}^{n} \sum_{j \epsilon dead} r_j$$
$$= \beta \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j + (1 - \beta) + \beta \sum_{j \epsilon dead} r_j$$

We know that $\sum_{i=1}^{n} M_{ij} = 1 \ \forall \ j \epsilon \ live$ and $\sum_{i=1}^{n} M_{ij} = 0 \ \forall \ j \ \epsilon \ dead$

$$\Rightarrow \beta \sum_{j \epsilon live} (1) r_j + (1 - \beta) + \beta \sum_{j \epsilon dead} r_j$$
$$= \beta \left(\sum_{j \epsilon live} (1) r_j + \sum_{j \epsilon dead} r_j \right) + (1 - \beta)$$
$$= \beta \sum_{j=1}^{n} r_j + (1 - \beta)$$
$$= \beta w(r) + (1 - \beta)$$
$$= \beta + 1 - \beta$$
$$= 1$$

Hence Proved.

Answer to Question 3(a)

Top 5 nodes with highest page rank score: [53 14 1 40 27]

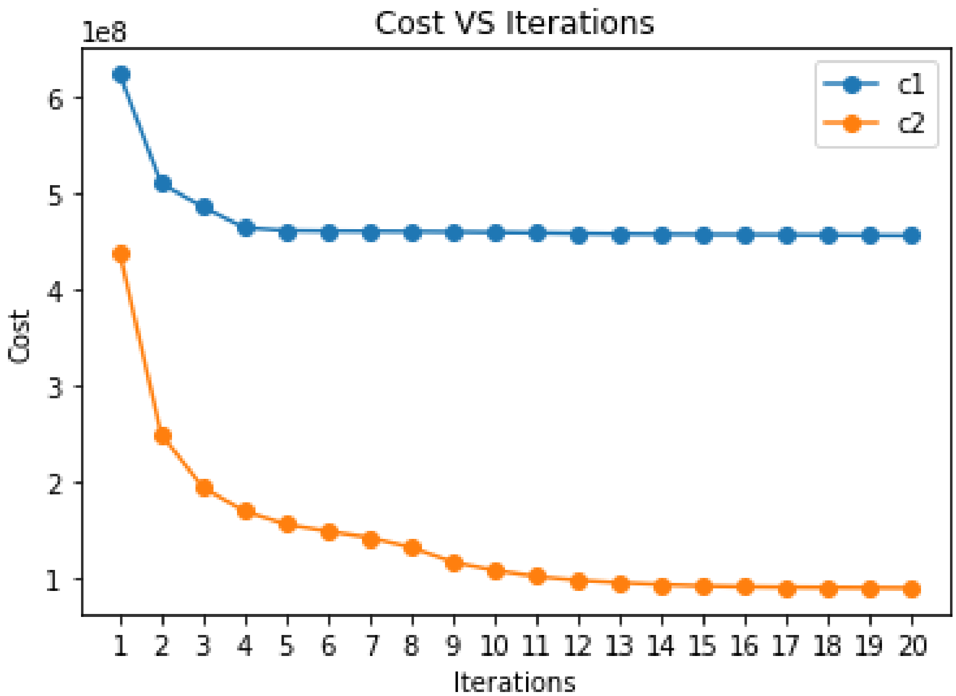
Node ID	Page rank
53	0.037868613328747594
14	0.03586677213352943
1	0.03514138301760087
40	0.03383064398237689
27	0.033130195547248505

Answer to Question 3(b)

Bottom 5 nodes with lowest page rank: [85 59 81 37 89]

Node ID	Page rank
85	0.0032348191433820193
59	0.003444256201194502
81	0.003580432413995564
37	0.003714283971941924
89	0.0038398576156450873

Answer to Question 4(a)



Answer to Question 4(b)

The percentage decrease in cost functions after 20 iterations are:

c1.txt = 26.8%
c2.txt = 79.4%

Clearly c2.txt has better clustering since the percentage decrease is profound. This is understandable because in c1.txt the cluster centers are chosen randomly, so K-means takes more iterations to reach the optimal cluster centers. Whereas c2.txt has clusters chosen as far away as possible so the optimal clustering is achieved very quickly and hence the rapid reduction in the value of cost function.