Cover Sheet

Assignment Submission Fill in and include this cover sheet with each of your assignments. Assignments are due at 11:59pm. All students (SCPD and non-SCPD) must submit their homeworks via GradeScope (`http://www.gradescope.com`). Students can typeset or scan their homeworks. Make sure that you answer each question on a separate page. Students also need to upload their code at `http://snap.stanford.edu/submit`. Put all the code for a single question into a single file and upload it. Please do not put any code in your GradeScope submissions.

Late Day Policy Each student will have a total of *two* free late periods. *One late period expires at the start of each class.* (Homeworks are usually due on Thursdays, which means the first late periods expires on the following Tuesday.) Once these late periods are exhausted, any assignments turned in late will be penalized 50% per late period. However, no assignment will be accepted more than *one* late period after its due date.

Honor Code We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently i.e., each student must understand the solution well enough in order to reconstruct it by him/herself. 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 (github/google/previous year solutions etc.) is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code very seriously and expect students to do the same.

Your name: MANOJ RAVI

Email: manoj1390@gmail.com **SUID:** manojr

CS246: Mining Massive Datasets
Homework 2

Answer to Question 1(a)

$$T = R \times R^T$$

$$= \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 1 & 3 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

T_{ii} => represents the no. of paths from u_i

T_{ij} => Similarly of the paths from u_i and u_j in terms of distination the paths finally reach to.

The diagonal elements of products transpose(R) and R AND R*transpose(R) contain diagonal elements of Q and P respectively which is also a reflection of the number of paths from ii and ui from the bipartite graph. The manipulation along with the division by square root terms of P and Q inverses will get us the representation of Si and Su

Answer to Question 1(b)

$$R * R^T = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 1 & 3 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

→ P-diagonal

$$S_I = R^T * R$$

$$= \begin{bmatrix} 3 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$Q = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

$$S_I = \begin{bmatrix} 3/\sqrt{3}\sqrt{3} & 1/\sqrt{3} & 2/\sqrt{3}\sqrt{3} \\ 1/\sqrt{3} & 3/\sqrt{3}\sqrt{3} & 1/\sqrt{3} \\ 2/\sqrt{3}\sqrt{3} & 1/\sqrt{3} & 3/\sqrt{3}\sqrt{3} \end{bmatrix}$$

$$Q^{-1} = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/3 \end{bmatrix}$$

$$= \frac{R^T * R}{} * ((Q^{-1})^{1/2}$$

$$(R^{-1})^{1/2} = \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/\sqrt{3} \end{bmatrix}$$

$$S_I = \begin{bmatrix} 1 & 1/\sqrt{3} & 2/3 \\ 1/\sqrt{3} & 1 & 1/\sqrt{3} \\ 2/3 & 1/\sqrt{3} & 1 \end{bmatrix}$$

$$(Q^{-1})^{1/2} * (Q^{-1})^{1/2} = \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/\sqrt{3} \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/\sqrt{3} \end{bmatrix}$$

$$= \begin{bmatrix} & & \end{bmatrix} \quad D_{ii} = \begin{bmatrix} 1 & 1/\sqrt{3} & 1/3 \\ 1/\sqrt{3} & 1 & 1/\sqrt{3} \\ 1/3 & 1/\sqrt{3} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix}$$

$$S_U = \begin{bmatrix} 1 & 1/\sqrt{2} & 2/\sqrt{2}\sqrt{3} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 & 1/\sqrt{3} & 0 \\ 2/\sqrt{2}\sqrt{3} & 1/\sqrt{3} & 1 & 1/\sqrt{3} \\ 1/\sqrt{2} & 0 & 1/\sqrt{3} & 1 \end{bmatrix}$$

$$= [(R * R^T) - P + I]^2 * (P^{-1})^{1/2}$$

$$= [(R^T * R) - Q + I] * (Q^{-1})^{1/2}$$

The recommendation matrix inherently has the product of R and Si(or)Su matrices per definition. Hence, the definition of a recommendation matrix has variations based upon the type of collaborative filtering and user-user /item-item similarities accordingly. The more similar these are the higher the chances of recommendation for that given item for a given user.

Answer to Question 1(c)

$$R = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

4×3

User-user

$$\mathcal{K}_{uu} = \begin{bmatrix} 1+\frac{1}{\sqrt{2}}(1)+\frac{\sqrt{2}}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} & 1+\frac{\sqrt{2}}{\sqrt{3}}+\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}}+1+\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}}+\frac{1}{\sqrt{3}} \\ \frac{\sqrt{2}}{\sqrt{3}}+\frac{1}{\sqrt{3}}+1 & 1 & \frac{\sqrt{2}}{\sqrt{3}}+1+\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}}+\frac{1}{\sqrt{3}} & 1/\sqrt{3} & \frac{1}{\sqrt{2}}+\frac{1}{\sqrt{3}}+1 \end{bmatrix}$$

$$= \begin{bmatrix} 2.525 & 0.816 & 2.52 \\ 2.228 & 0.577 & 1.28 \\ 2.39 & 1 & 2.39 \\ 1.28 & 0.577 & 2.28 \end{bmatrix}$$

item-item

$$\mathcal{K}_{ii} = \begin{bmatrix} 1+\frac{2}{\sqrt{3}} & \frac{1}{\sqrt{3}}+\frac{1}{\sqrt{3}} & \frac{2}{3}+1 \\ 1 & 1/\sqrt{3} & 2/3 \\ 1+1/\sqrt{3}+2/3 & \frac{1}{\sqrt{3}}+1+\frac{1}{\sqrt{3}} & \frac{2}{3}+\frac{1}{\sqrt{3}}+1 \\ 2/3 & 1/\sqrt{3} & 1 \end{bmatrix} = \begin{bmatrix} 2.155 & 1.15 & 1.66 \\ 1 & 0.577 & 0.66 \\ 2.237 & 2.15 & 2.237 \\ 0.66 & 0.577 & 1 \end{bmatrix}$$

$$P_u = S_I * R * = [(R^T - R - Q) + \varepsilon]^{-1/2} * R$$

$$P_I = R * S_u * = [[R \cdot (I) - P + \varepsilon]]^{-1/2}$$

Answer to Question 1(d)

Top 5 user-user collaborative filtering :

Fox 28 News at 10pm

Family Guy

2009 NCAA Basketball Tournament

NBC 4 at Eleven

Two and a Half Men

Top 5 item-item :

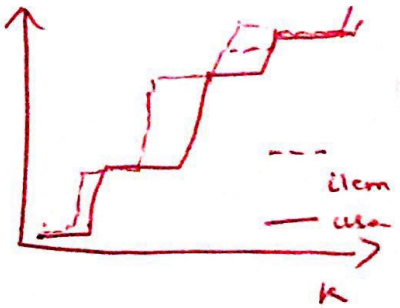
Fox 28 News at 10pm

Family Guy

2009 NCAA Basketball Tournament

NBC 4 at Elem

Access Hollywood



(Graph attached)

Answer to Question 2(a)

M^TM & MM^T is symmetric, square and real

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

Illy, $(M^TM)^T = (M)^T(M^T)^T = M^TM$

if $M \to p \times q$

$M^T \to q \times p$

$$\xrightarrow{\text{dim}} (MM^T) \Rightarrow p \times q \cdot q \times p$$

$$\Rightarrow p \times p$$

(or $q \times q$ in the case of M^TM)

if M is real MMT will also be real.

Answer to Question 2(b)

$\{v \mid Av = \lambda v\}$

if Eigen value of A is λ then $Av = \lambda v$ → nontrivial subspace
($A_{n \times n}$) ($v \rightarrow$ non zero vector)

determinant $|\lambda I - A|$ is called characteristic polynomial

Zeros of this polynomial are eigen values of A.

if M is non-singular, MM^T & M^TM are similar

$$A = M^TM = M^{-1}(MM^T)M$$

For a square matrix A of size $n \times n$

$$\det(A - zI) = (-1)^n (z - \lambda_1)(z - \lambda_2) \cdots (z - \lambda_n)$$

($\lambda_1, \lambda_2 \ldots$ eigen values)

Also a square matrix has same characteristic polynomial as its transpose: $\det(A - zI) = \det((A - zI)^T) = \det(A^T - zI^T)$

$$= \det(A^T - zI)$$

Also similar matrices share same characteristic polynomial same eigenvalues

If P is non-singular,

$$\det(A - zI) = \det(P^{-1}) \det(A - zI) \det(P)$$

$$= \det(P^{-1}AP - zP^{-1}IP) = \det(P^{-1}AP - zI)$$

Hence, A and $P^{-1}AP$ have same characteristic polynomial Hence, same eigen values. Eigen vectors could be scalar multiple of one other and could be transformed into on other

Answer to Question 2(c)

$$M^T M = Q \lambda Q^T$$

Answer to Question 2(d)

$$M = U \Sigma V^T$$

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

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

Since $U^T U = I$

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

Answer to Question 2(e)

$$V = \begin{bmatrix} -0.2785 & 0.5 \\ -0.2785 & -0.5 \\ -0.6499 & 0.5 \\ -0.6499 & 0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 7.6157 & 0 \\ 0 & 1.414 \end{bmatrix}$$

$$V^T = \begin{bmatrix} -0.7071 & -0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \quad V =$$

$$Evals = [2., 58.]$$

$$Evec = \begin{bmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}$$

We can see that- Evec and V are same (and linearly transformable from one other)

Eigenvalues from PCA (MTM) are squares of singular values of SVD

Answer to Question 3(a)

Centroid based

C_1 $\quad \frac{0.68+0.72+0.79}{3}$

$= \frac{2.19}{3} = 0.73$

C_2 $\quad \frac{0.87+0.92+0.99}{3}$

$= 0.93$

C_3 $\quad \frac{1.01+1.22+1.45}{3}$

$= 1.23$

C_1 and C_2 will be merged

Closest member based

C_1 & C_2 dist. = 0.08

C_2 & C_3 dist = 0.02

C_2 and C_3 will be merged

Centroid based merging will learn clusters such that one of the clusters correspond to low risk of lung cancer and other represent high risk of lung cancer.

Answer to Question 3(b)

Centroid based

↓

$C_1 = 0.73$

$C_2 = \frac{0.87 + 0.92}{2}$

$= 0.90$

$C_3 = 1.23$

C_1 and C_2 will be merged

Closest Member based

C_1 & C_2 dist. = 0.08

C_2 & C_3 dist. = 0.09

C_1 & C_2 will be merged

The centroid based merging strategy is stable

worst point $P_6 = 0.99$ and existing set of cluster

Answer to Question 3(c)

I would prefer a stable merge strategy over a unstable one because a stable merge strategy would have the chances of being less affected (or not at all affected) by the 5% of miscalculated NSR in the dataset and its clustered sets will not be decisive entirely based upon the miscalculated points or not at all dependent on them. Hence, we could get & expect a more reliable merge strategy from a stable one.

Answer to Question 4(a)

1. Graph plot

2. $^1C_0$ (scaled c_1, initial) $= 6.236603453064107E8 -$

$^1C_{10} = 4.584906561919809E8$ (10th)

Percentage change after 10 iterations

using c_1

$$\frac{^1C_0 - {}^1C_{10}}{^1C_0} = \frac{16516989}{C_0} = 26.48\% \text{ decrease}$$

using c_2 (% change)

$^2C_0 = 4.387477900279162 5E8 -$

$^2C_{10} = 1.022372033179960 5E8$

$$\frac{^2C_0 - {}^2C_{10}}{^2C_0} = 76.69\% \text{ decrease}$$

When using Euclidean distance for k-means and calculating the cost function, far-away assignment of initial clusters has a better chance of starting with a relatively lower cost and also in having better algorithm efficiency in terms of percentage decrease of cost function. This is because of basic algorithm of k-means which tends to cluster points closer into a cluster and farther away in distance clusters. Starting with faraway centroids reflects the assumption of far-away clusters which can generally help in good cost reduction unless we have many outliers.

(Percentage change after 10 iterations)

Answer to Question 4(b)

1. Graph plotted 2.

C_1 - txt

$^1C_0 = 550117.142$

$^1C_{10} = 447494.3682$

$\rightarrow \frac{^1C_0 - ^1C_{10}}{^1C_0} = 18.655\%$ decrease

C_2 - txt

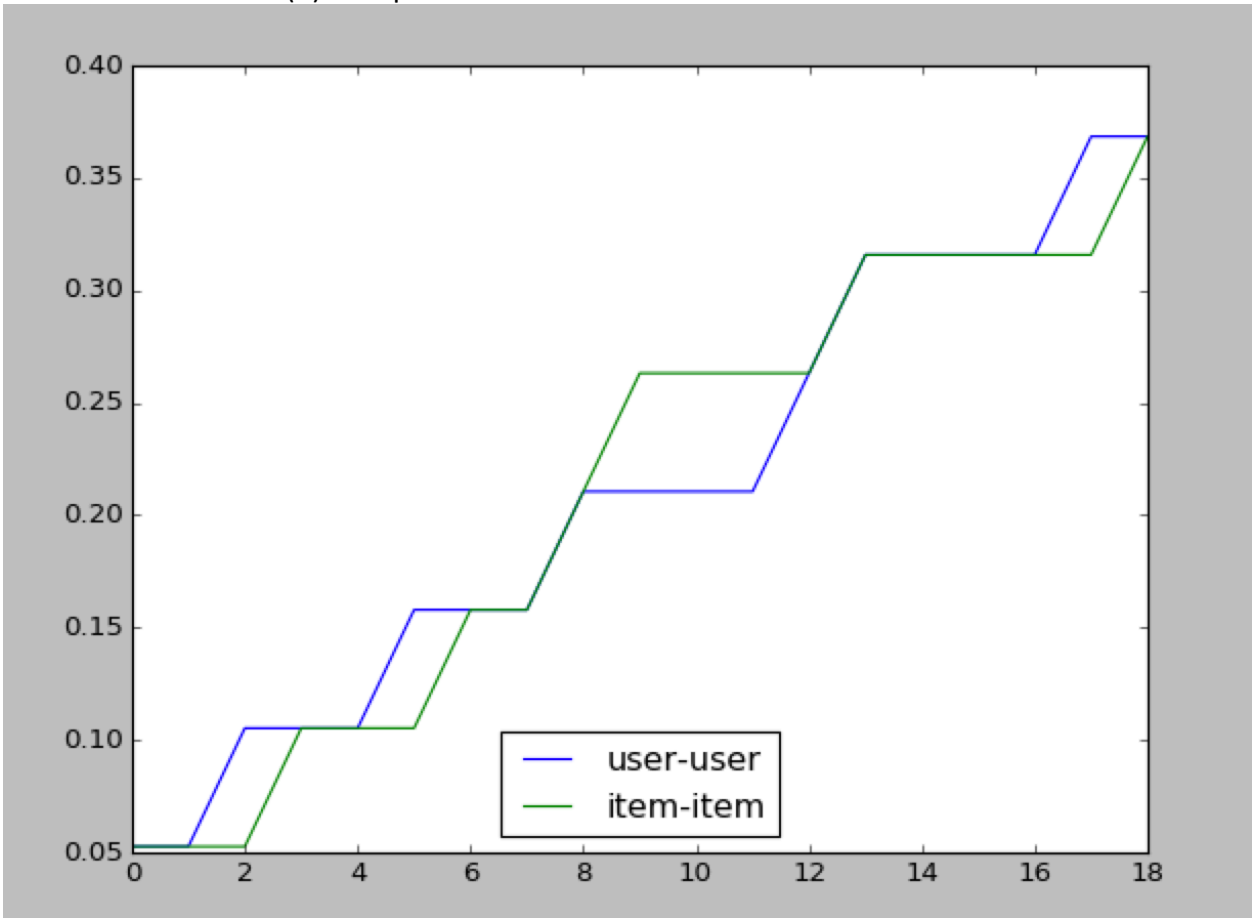
$^2C_0 = 1433739.31$

$^2C_{10} = 694587.92525$

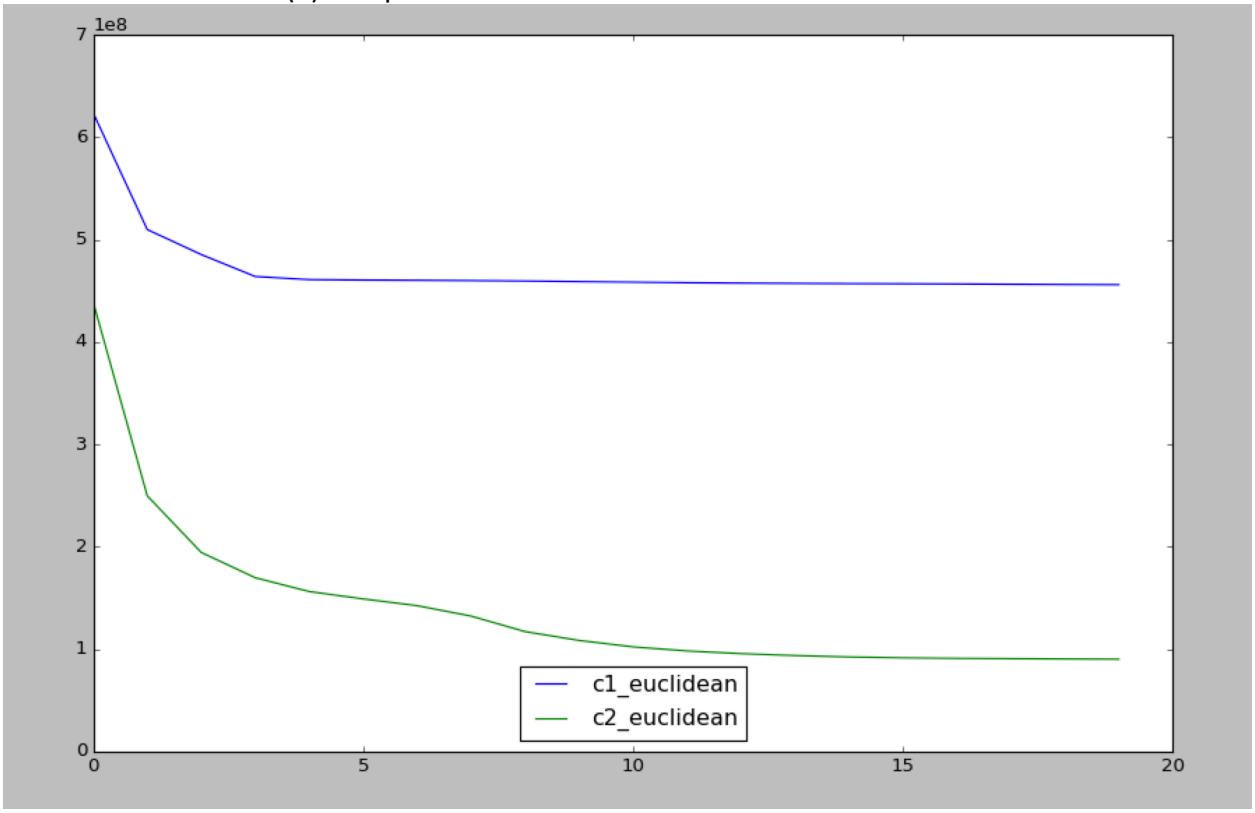
$\rightarrow \frac{^2C_0 - ^2C_{10}}{^2C_0} = 51.55\%$

Since the points in C_2 have been set up based upon farthest distance in L_2 space they need not be farthest ones according to Manhattan distance. So, initializing with lower cost seems more or equally likely with C_1 (random) initialization. But the algorithm effiency in terms of reducing the cost & C_2 still is better by starting with C_2 (like 4a). Because, it helps sustain the underlying concept of k-means which clusters close points together and farther away points in different clusters. Unless there is lot of outliers this should be an efficient start for more reduction of cost.

Answer to Question 1(d) - Graph:



Answer to Question 4(a)- Graph:



Answer to Question 4(b)- Graph: