CS246: Mining Massive Data Sets Problem Set 2

Name: Dat Nguyen
Date: 06/05/2019

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

1 Singular Value Decomposition and Principal Component Analysis (20 pts)

(a) MM^T and M^TM are symmetric because $(MM^T)^T = MM^T$ and $(M^TM)^T = M^TM$. They are square because M has size $p \times q$ and M^T has size $q \times p$ so MM^T has size $p \times p$ and M^TM has size $q \times q$.
They are real because both M and M^T are real.

(b) Let the nonzero eigenvalues of MM^T be λ and the corresponding eigenvector be v, we have

$$MM^Tv = \lambda v$$

Multiply both sides by M^T to the left we get

$$M^TMM^Tv = M^T\lambda v$$
$$(M^TM)(M^Tv) = \lambda(M^Tv)$$

Therefore we can see that λ is also the eigenvalue of M^TM with the eigenvector M^Tv. Assume that for every eigenvector v of MM^T, M^Tv is also eigenvector of MM^T. By that assumption, M^TM^Tv is also eigenvector of MM^T and so on. Because the set of eigenvector is finite, the assumption does not hold. Therefore the set of eigenvectors of MM^T and M^TM are not the same.

(c) Because M^TM is symmetric, square and real, by eigenvalue decomposition we can write M^TM as

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

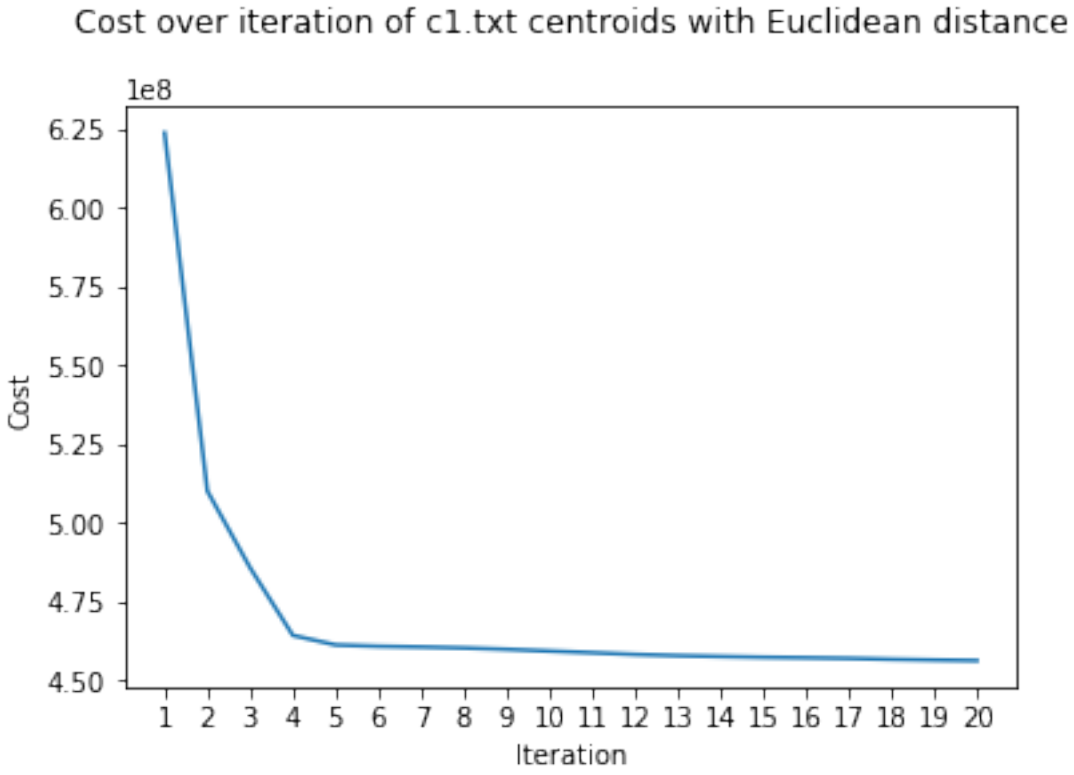
(d) We have

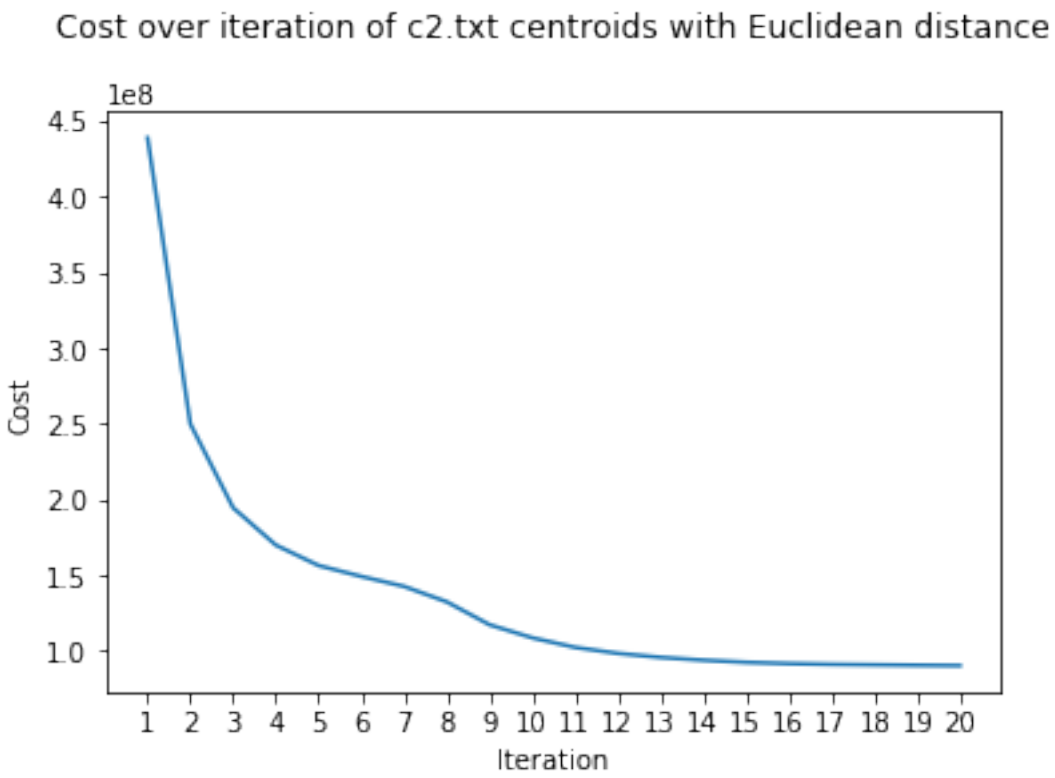
$$\begin{aligned} M^TM &= (U\Sigma V^T)^TU\Sigma V^T \\ &= V\Sigma U^TU\Sigma V^T \\ &= V\Sigma^2V^T \end{aligned}$$

(e)
- The columns of V produced by SVD are equal to some scalars time the corresponding columns of the matrix of eigenvectors $Evecs$. This is because from part d we see that $M^TM = V\Sigma^2V^T$ and $M = U\Sigma V^T$. In addition, if we multiply an eigenvector with a scalar we also get an eigenvector with the same eigenvalue.
- The eigenvalues of M^TM are the squares of singular values of M. The reason is that from part d we have $M^TM = V\Sigma^2V^T$ and $M = U\Sigma V^T$. Also, Σ is a diagonal matrix.

2 k-means on Spark (20 points)

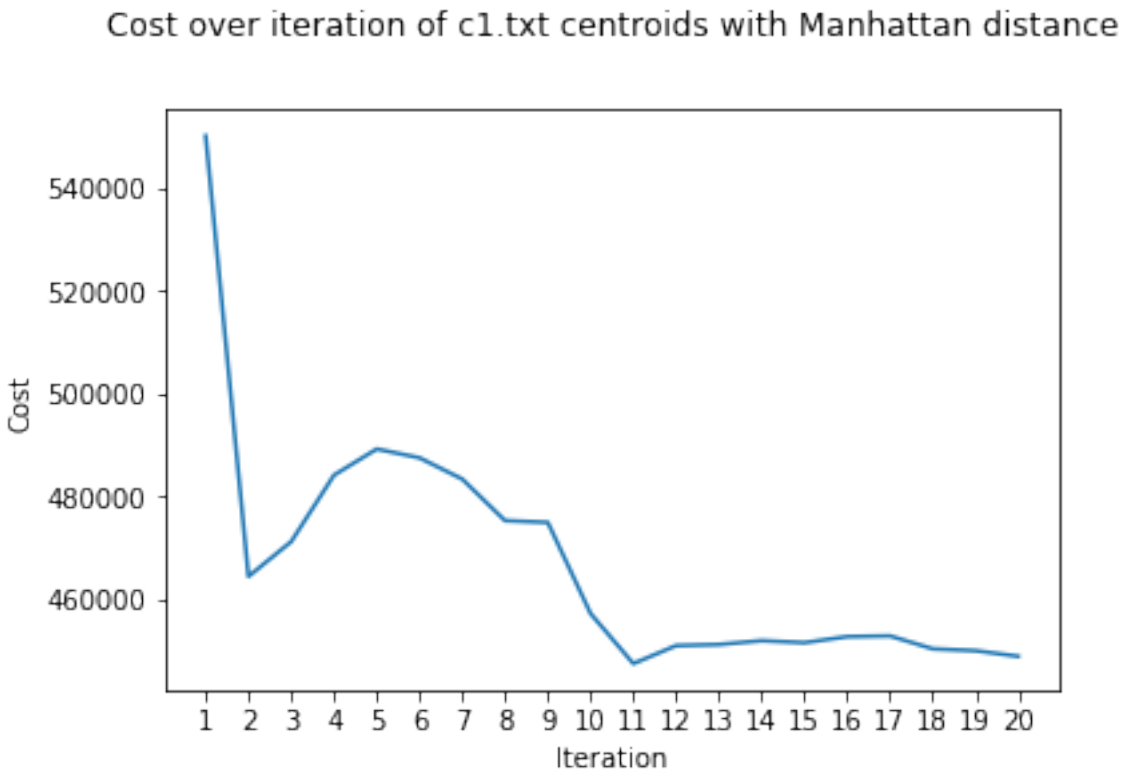
(a) 1. Plots of cost vs. iteration

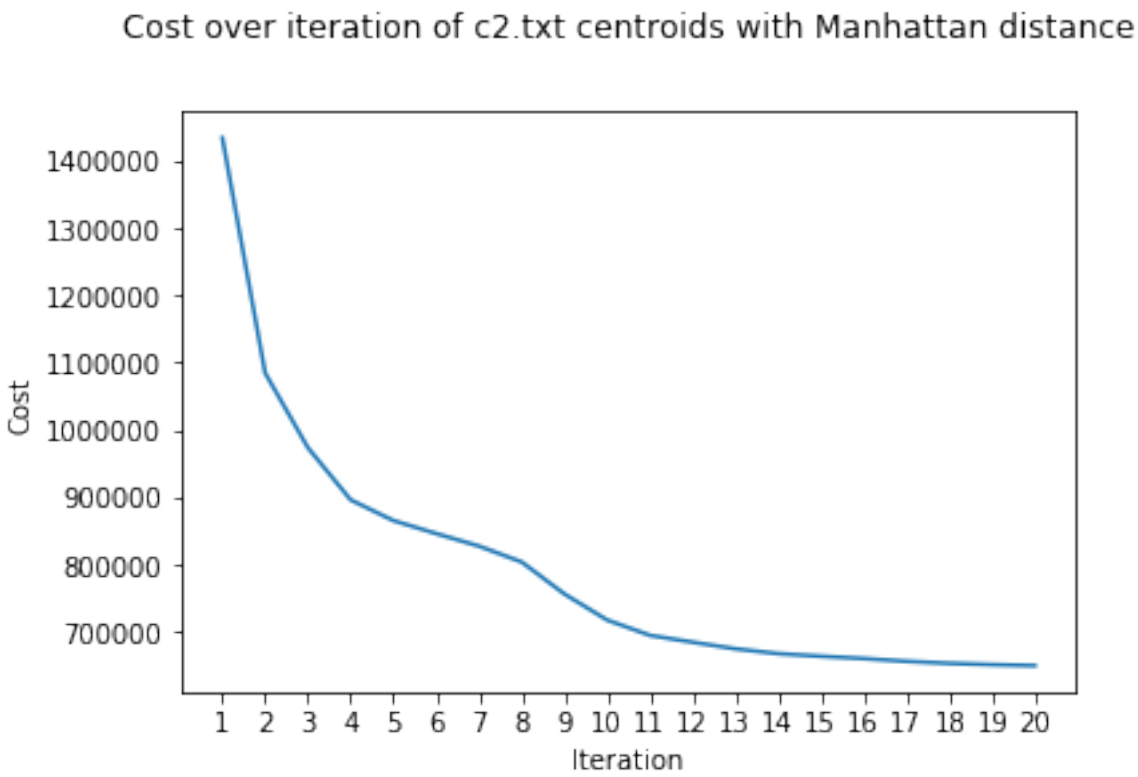




2. The percentage changes in cost after 10 iterations when cluster centroids are initialized using **c1.txt** and **c2.txt** with Euclidean distance are 26.4% and 75.3% respectively. The random initialization of k-means using **c2.txt** is better than initialization using **c1.txt** in term of cost $\phi(i)$. The reason is that when the initial centroids are as far apart as possible, the algoritm are less likely to get stuck into local minima, since we expect that the correct cluster centroids are also far apart from each other.

(b) 1. Plots of cost vs. iteration


Cost over iteration of c1.txt centroids with Manhattan distance
Cost
540000
520000
500000
480000
460000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Iteration


Cost over iteration of c2.txt centroids with Manhattan distance
Cost
1400000
1300000
1200000
1100000
1000000
900000
800000
700000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Iteration

2. The percentage changes in cost after 10 iterations when cluster centroids are initialized using **c1.txt** and **c2.txt** with Manhattan distance are 16.9% and 50.0% respectively. The random initialization of k-means using **c1.txt** is better than initialization using **c2.txt** in term of cost $\psi(i)$. The reason is that the way the centroids in **c2.txt** are initialized is by maximum Euclidean distance from each other, but the algoritm uses Manhattan distance so the distance metrics do not match.

3 Latent Features for Recommendations (35 points)

(a) Derivative of the error E with respect to R_{iu}

$$\epsilon_{iu} = 2(R_{iu} - q_i \cdot p_u^T)$$

Derivative of E with respect to q_i

$$\frac{\partial E}{\partial q_i} = \Big(\sum_{\{u|(i,u)\in \text{ratings}\}} -2(R_{iu} - q_i \cdot p_u^T)p_u \Big) + 2\lambda q_i$$

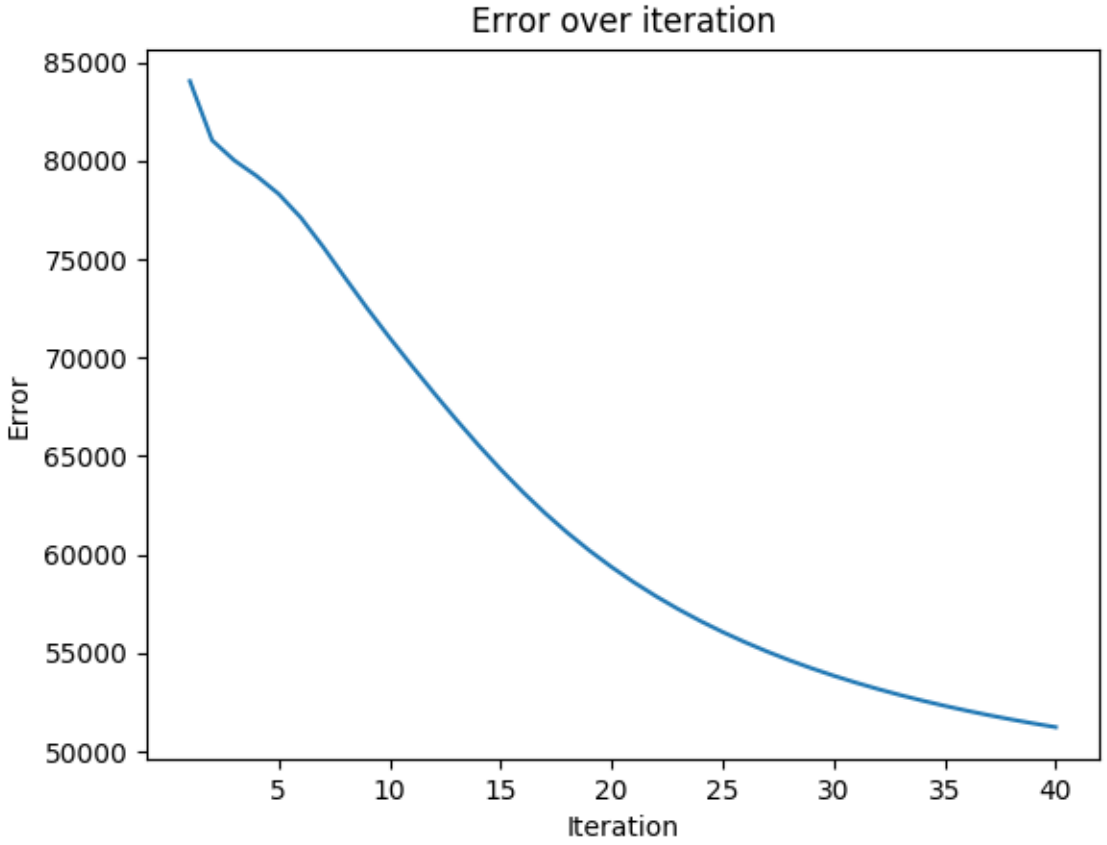
Derivative of E with respect to p_u

$$\frac{\partial E}{\partial p_u} = \Big(\sum_{\{i|(i,u)\in \text{ratings}\}} -2(R_{iu} - q_i \cdot p_u^T)q_i \Big) + 2\lambda p_u$$

Update equations

$$q_i = q_i - \eta \frac{\partial E}{\partial q_i}$$
$$p_u = p_u - \eta \frac{\partial E}{\partial p_u}$$

(b) The value of η is 0.009.



4 Recommendation Systems (25 points)

(a) T_{ii} indicates the degrees of user node i. $T_{ij}(i \neq j)$ indicates the number of item nodes to which there are paths from both user node i and user node j.

(b) Let $M = R^T R$ and R_i be column i^{th} of R. We can see that entry j in column M_i of M indicates $R_j^T R_i$. Let $Q_{r,:}^{-1/2}$ and $Q_{:,c}^{-1/2}$ be row r and column c of matrix $Q^{-1/2}$ respectively, then item at row r and column c of S_I is

$$
\begin{aligned}
& Q_{r,:}^{-1/2} \begin{bmatrix} M_1 & \dots & M_n \end{bmatrix} Q_{:,c}^{-1/2} \\
=& \begin{bmatrix} Q_{r,:}^{-1/2} M_1 & \dots & Q_{r,:}^{-1/2} M_n \end{bmatrix} Q_{:,c}^{-1/2} \\
=& \begin{bmatrix} \frac{R_r^T R_1}{||R_r||} & \dots & \frac{R_r^T R_n}{||R_r||} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1/||R_c|| \\ \vdots \\ 0 \end{bmatrix} \\
=& \frac{R_r^T R_c}{||R_r||||R_c||} \\
=& \text{cos-sim}(R_r, R_c)
\end{aligned}
$$

Therefore we conclude that the item at row r and column c of S_I is the cosine similarity of item r and item c.
Similarly, the user similarity matrix S_U can be defined as

$$S_U = P^{-1/2} R R^T P^{-1/2}$$

The derivation is in the same manner as the derivation for S_I shown above.

(c) Consider the element at row i and column j of $\Gamma = RQ^{-1/2}R^T RQ^{-1/2} = RS_I$

$$
\begin{aligned}
\Gamma_{ij} &= \sum_k R_{ik} * S_{I_{kj}} \\
&= \sum_k R_{ik} * \text{cos-sim}(\text{item } k, \text{item } j) \\
&= r_{i,j}
\end{aligned}
$$

Similary, for the user-user case we have $\Gamma = R^T P^{-1/2} R R^T P^{-1/2}$. Consider the element at row i and column j of *Gamma*

$$
\begin{aligned}
\Gamma_{ij} &= \sum_k R_{ik}^T * S_{U_{kj}} \\
&= \sum_k R_{ik}^T * \text{cos-sim}(\text{user } k, \text{user } j) \\
&= r_{i,j}
\end{aligned}
$$

Where $r_{i,j}$ is the calculated rating rated for item i by user k.

(d)
- The shows recommended by item-item collaborative filtering are "FOX 28 News at 10pm", "Family Guy", "NBC 4 at Eleven", "2009 NCAA Basketball Tournament", and "Access Hollywood". The corresponding similarity scores are 31.36, 30.00, 29.40, 29.23, and 28.97.

- The shows recommended by user-user collaborative filtering are "FOX 28 News at 10pm", "'Family Guy", "2009 NCAA Basketball Tournament", "NBC 4 at Eleven", and "Two and a Half Men". The corresponding similarity scores are 908.48, 861.18, 827.60, 784.78, and 757.60.