CS246: Mining Massive Data Sets

Assignment number: 2

Fill in and include this cover sheet with each of your assignments. It is an honor code violation to write down the wrong time. Assignments and code are due at 5:00 PM on Scoryst and SNAP respectively. Failure to include the coversheet with you assignment will be penalized by 2 points. Each student will have a total of *two* free late periods. *One late period expires at the start of each class.* (Assignments are due on Thursdays, which means the first late period expires on the following Tuesday at 5:00 PM.) 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. (If an assignment is due to Thursday then we will not accept it after the following Thursday.)

Your name:Erli Zhou
Email:erlizhou@stanford.edu
SUNet ID:erlizhou

Collaborators: Jinfeng Huang, Ling-Ling Zhang

I acknowledge and accept the Honor Code.

*(Signed)*Erli Zhou

Answer to Question 1a

For S_I, the element in row i and column j is the cosine similarity of item i and item j.
cossimi(item i, item j) = $\frac{R_{:,i}R_{:,j}}{||R_{:,i}||||R_{:,j}||}$
Also, the norm of item i = $\sqrt{\Sigma_{k=1}^{m} R_{ki}^2} = \sqrt{\Sigma_{k=1}^{m} R_{ki}}$
The sum is equal to number of users who like the item, which is equivalent to Q_{ii}, thus cossimi(item i, item j) = $\frac{R_{:,i}R_{:,j}}{\sqrt{Q_{ii}Q_{jj}}}$
Because the matrix Q_{ii} and Q_{jj} are diagonal, cossimi(item i, item j) = $Q_{ii}^{-1/2}(R_{:,i}R_{:,j})Q_{jj}^{-1/2}$
So $S_I = Q^{-1/2}(RR_t)Q^{1/2}$

For S_U, the element in row i and column j is the cosine similarity of item i and item j.
cossimi(item i, item j) = $\frac{R_{:,j}R_{:,i}}{||R_{:,j}||||R_{:,i}||}$
Also, the norm of item i = $\sqrt{\Sigma_{k=1}^{m} R_{ki}^2} = \sqrt{\Sigma_{k=1}^{m} R_{ki}}$
The sum is equal to number of users who like the item, which is equivalent to P_{ii}, thus cossimi(item i, item j) = $\frac{R_{:,j}R_{:,i}}{\sqrt{P_{jj}P_{ii}}}$
Because the matrix P_{ii} and P_{jj} are diagonal, cossimi(item i, item j) = $R_{jj}^{-1/2}(R_{:,j}R_{:,i})P_{ii}^{-1/2}$
So $S_U = P^{-1/2}(R_tR)P^{1/2}$

Answer to Question 1b

For user-user collaborative recommendation,
$r_{u,s} = \Sigma_{x \in U} cos(x,u) R_{x,s} = \Sigma_{x=1}^{m} cos(x,u) R_{x,s} = \Sigma_{x=1}^{m} (S_U)_(u,x) R_{x,s} = (S_U R)_{us}$
So $\Gamma = S_U R = P^{-1/2}(R_t R) P^{1/2} R$

For item-item collaborative recommendation,
$r_{u,s} = \Sigma_{x \in I} R_{u,x} cos(x,s) = \Sigma_{x=1}^{n} R_{u,x} cos(x,s) = \Sigma_{x=1}^{n} R_{u,x} (S_I)_(x,s) = (R S_I)_{us}$
So $\Gamma = R S_I = R Q^{-1/2}(R R_t) Q^{1/2}$

Answer to Question 1c

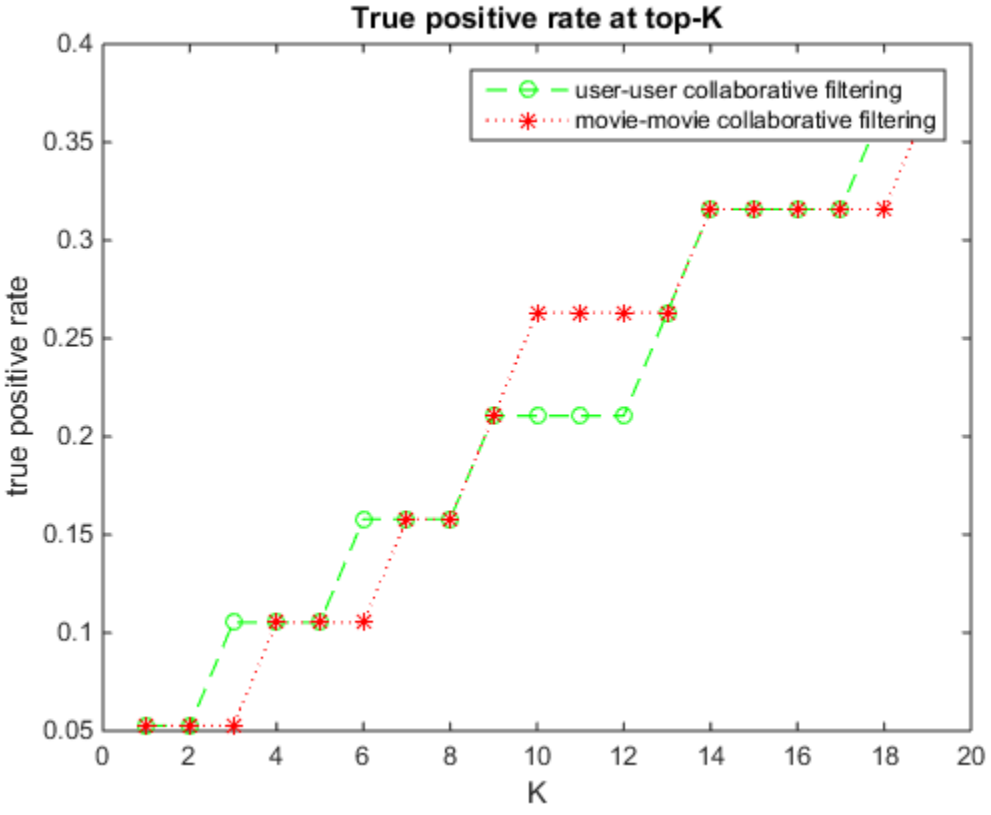
If a user i likes item j, then $R_{i,j} = 1$, otherwise $R_{i,j} = 0$
$T_{ii} = (R * R^T)_{ii} = \Sigma_{k=1}^{n} R_{ik} R_{ki}$
In the bipartite graph, T_{ii} equals the number of items user i like, which means the degree of user i.

$T_{ij} = (R * R^T)_{ij} = \Sigma_{k=1}^{n} R_{ik} R_{kj}, i \neq j$
In the bipartite graph, T_{ij} equals the number of users that liked item i, which means the degree of item i.

Answer to Question 1d

The five TV shows that have the highest similarity scores for Alex for the user-user collaborative filtering:
"FOX 28 News at 10pm" 908.4801
"Family Guy" 861.176
"2009 NCAA Basketball Tournament" 827.6013
"NBC 4 at Eleven" 784.782
"Two and a Half Men" 757.6011

The five TV shows that have the highest similarity scores for Alex for the movie-movie collaborative filtering:
"FOX 28 News at 10pm" 31.3647
"Family Guy" 30.0011
"NBC 4 at Eleven" 29.3968
"2009 NCAA Basketball Tournament" 29.227
"Access Hollywood" 28.9713



The user-user collaborative filtering and movie-movie collaborative provide similar plots, with true positive rate at top-K differed by at most around 5 percent.

Answer to Question 2a

1.$C = A^T A = (USV^T)^T(USV^T) = V(US)^T(USV^T) = VS^TU^T(USV^T) = VS^2V^T$
$C^T = (VS^2V^T)^T = V(VS^2)^T = V(S^2)^TV^T = VS^TS^TV^T = VS^2V^T$
Since $C = C^T$, C is symmetric.

$K = AA^T = (USV^T)(USV^T)^T = (USV^T)V(US)^T = (USV^T)VS^TU^T = US^2U^T$
$K^T = (US^2U^T)^T = U(US^2)^T = U(S^2)^TU^T = US^TS^TU^T = US^2U^T$
Since $K = K^T$, K is symmetric.

The non-zero eigenvalues of C are S^2 and the eigenvectors are columns of V.

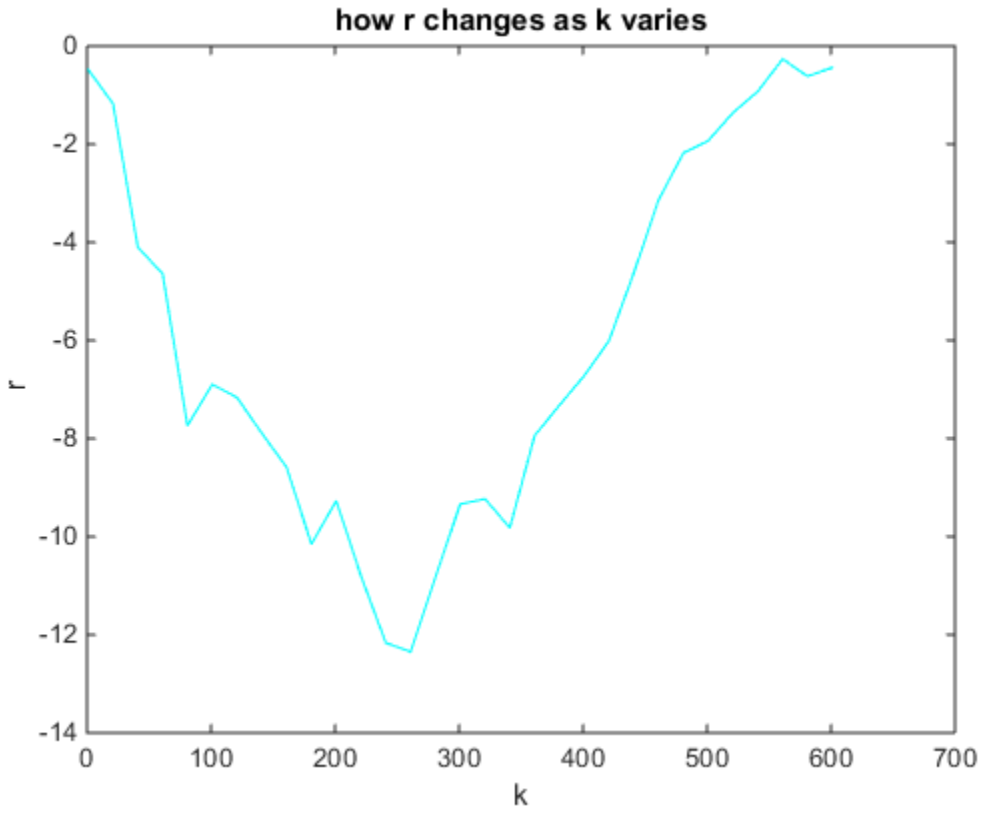
The non-zero eigenvalues of K are S^2 and the eigenvectors are columns of U.

2. Since m = 100 and n = 10,000, we should use K because it is only 100*100 compared to C which is 10,000*10,000

3. According to part 1, columns of V are the eigenvectors of C
From $A = USV^T, AV = US, U = AVS^{-1}$

4. According to part 1, columns of U are the eigenvectors of K
From $A = USV^T, U^TA = SV^T, V = (S^{-1}U^TA)^T = (U^TA)^TS^{-1} = A^TUS^{-1}$

Answer to Question 2b



Answer to Question 2c

1. Using the definition in the problem $||A||_F^2 = \Sigma_{i=1}^n ||a_i||_2^2, ||A - XW||_F^2 = \Sigma_{i=1}^n ||a_i - Xw_i||_2^2$
From the linear regression vector, $w_i = (X^T X)^{-1} X^T a_i$
So $W = (X^T X)^{-1} X^T A$

2. $W = (\tilde{U}^T \tilde{U})^{-1} \tilde{U}^T U S V^T = S V^T = \tilde{S} \tilde{V}^T$

3. $||A - XW||_F^2 = ||A - X(X^T X)^{-1} X^T A||_F^2 = ||A - X X^T A||_F^2$

4.

5.

Answer to Question 3a

$\Sigma_{x\in S}||x-z||^2 - \Sigma_{x\in S}||x-c(S)||^2 - |S|*||c(S)-z||^2 = \Sigma_{x\in S}(||x-z||^2 - ||x-c(S)||^2 - ||c(S)-z||^2) = \Sigma_{x\in S}[x^2+z^2-2xz-x^2-c(S)^2+2xc(S)-c(S)^2-z^2+2zc(S)] = \Sigma_{x\in S}[-2xz+2xc(S)+2zc(S)-2c(S)^2] = 2\Sigma_{x\in S}[x-c(S)]^T[c(S)-z]$.
Since $c(S) = \frac{\Sigma_{x\in S}x}{|S|}, \Sigma_{x\in S}[x-c(S)] = 0$, thus $2\Sigma_{x\in S}[x-c(S)]^T[c(S)-z] = 0$
So $\Sigma_{x\in S}||x-z||^2 - \Sigma_{x\in S}||x-c(S)||^2 = |S|*||c(S)-z||^2$

Answer to Question 3b

Suppose at the beginning of each iteration t,
$C^{(t-1)} = \{c_1^{(t-1)}, c_2^{(t-1)}, ..., c_k^{(t-1)}\}$ are k current clusters and $\phi^{(t-1)} = \Sigma_{x \in \mathcal{X}} min_{c \in C^{(t-1)}} ||x - c||^2$ is the cost function
For step 2 in the algorithm, we set the cluster $C_i^{(t)} = \{x \in \mathcal{X} | argmin_{1 \leq j \leq k} \{||x - c_j^{(t)}||\} = i\}$
So $\forall x \in \mathcal{X}, min_{1 \leq j \leq k} \{||x - c_j^{(t)}||\} \leq min_{1 \leq j \leq k} \{||x - c_j^{(t-1)}||\}$ since x either moves to a cluster whose centroid is closer to it than the previous centroid or remains at the previous cluster
For step 3 in the algorithm, we set $c_i^{(t)}$ to be the center of mass of $C_i^{(t)}, c_i^{(t)} = \frac{1}{|C_i|} \Sigma_{X \in C_i} X$

Answer to Question 3c

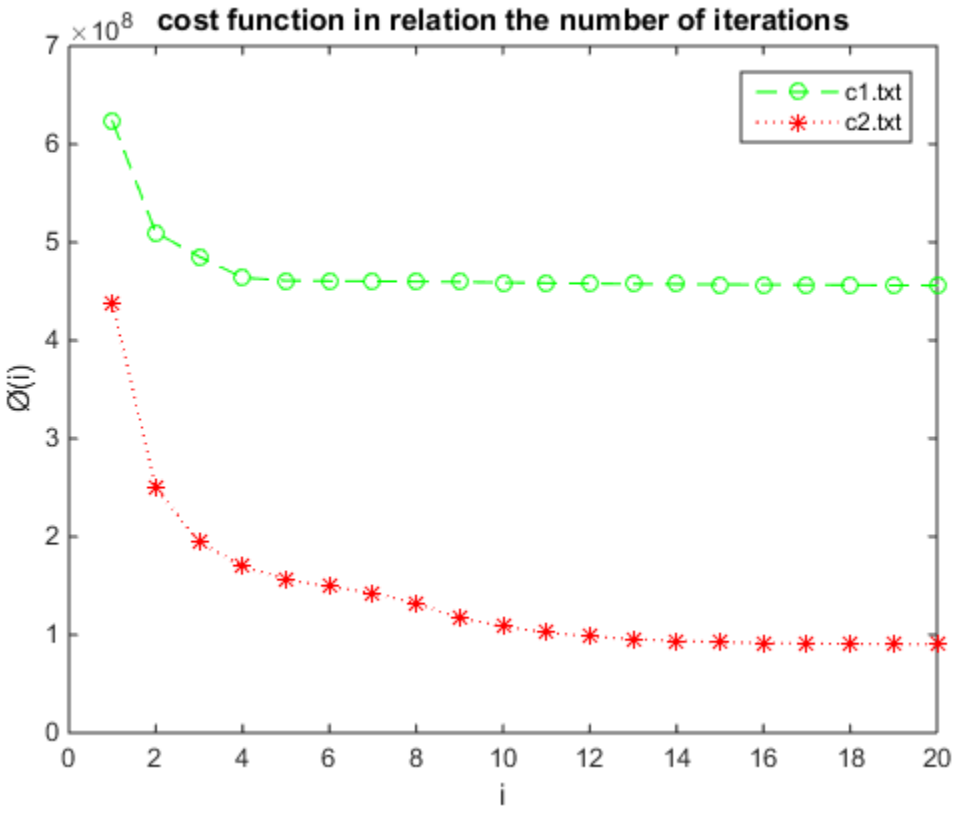
Question 3b proved that the cost function $\phi^{(t)}$ is either decreasing or remains equal, which fits the description of a monotonic decreasing sequence.
Besides, $\phi^{(t)}$ has a lower bound of zero, since the square of of any norm is always non-negative.
Since a bounded monotonic decreasing sequence is convergent, ϕ converges to a finite value.

Since $\binom{n}{k}$, the total number of possible selections of the clusters, is finite, there are finite number of states.
The algorithm is constantly improving the cost function, so based on Markov chain theory it doesn't revisit a state.
As a result, it will reach a state where the centroids converge.

Answer to Question 3d

Suppose we have 4 points $A(-2,0)B(-2,10)C(r,0)D(r,10), r > 1$
An optimal cluster should lead to $c_1(-2,5)$ and $c_2(r,5)$, with $\phi = 25 * 4 = 100$
If we set $(r/2,0)$ and $(r/2,10)$ as the initialization points though, A and C belongs to C_1 while B and D belongs to C_2. So $c_1(r/2-1,0)$ and $c_2(r/2-1,10)$, the cluster no longer changes. $\phi = (r/2+1)^2 * 4$, which definitely has an unbounded larger cost than the optimal clustering. Given an arbitrary number $r > 1$, the converged cost with bad initialization is at least r times larger than the cost of the optimal clustering.

Answer to Question 4b



Answer to Question 4c

No, random initialization using c1.txt is not better than c2.txt in terms of cost $\phi(i)$, a bad random initialization can lead to a lower cost bound which is much higher than the optimal cluster, as shown in question 3d. In this aspect, choosing centroids which are as far as possible reduces the chance of bad initialization happening since the odds that the cost function already converges are extremely low.

After 10 iterations, cost decreased by 26.40 percent in c1.txt and by 75.26 percent in c2.txt

Answer to Question 4d

5 best tags, in descending order, are instillation, methoxyfenozide, sodium-sensitive, post-infectious and teleological.