COMS 4721 Machine Learning for Data Science

Name: Sung-Yen Liu

HW#4

UNI: sl3763

Problem1

1.

Code:

Hw4pr1.m

```
% hw4.1

mu1 = [0 0];
sigma1 = [1 0 ; 0 1];
mu2 = [3 0];
sigma2 = [1 0 ; 0 1];
mu3 = [0 3];
sigma3 = [1 0 ; 0 1];
p = [0.2 0.5 0.3];

sampleG = sampleDiscrete(500,p);
g1 = mvnrnd(mu1,sigma1,length(sampleG(sampleG==1)));
g2 = mvnrnd(mu2,sigma2,length(sampleG(sampleG==2)));
g3 = mvnrnd(mu3,sigma3,length(sampleG(sampleG==3)));

alldat = [g1;g2;g3]; % 500*2
errorRec = zeros(21,4);

for k = 2:5
        rp = randperm(500);
        initC = alldat(rp(1:k),:);
        nextC = zeros(size(initC,1),2);

        [nextCinfo, error] = calculateMinDistance(initC,alldat);
        errorRec(1,k-1) = sum(error);
        for ite = 1:20
                for ce = 1:k
                        temp = alldat(find(nextCinfo==ce),:);
                        nextC(ce,:) = mean(temp);
                end
                [nextCinfo, error] = calculateMinDistance(nextC,alldat);
                errorRec(ite+1,k-1) = sum(error);
        end
```

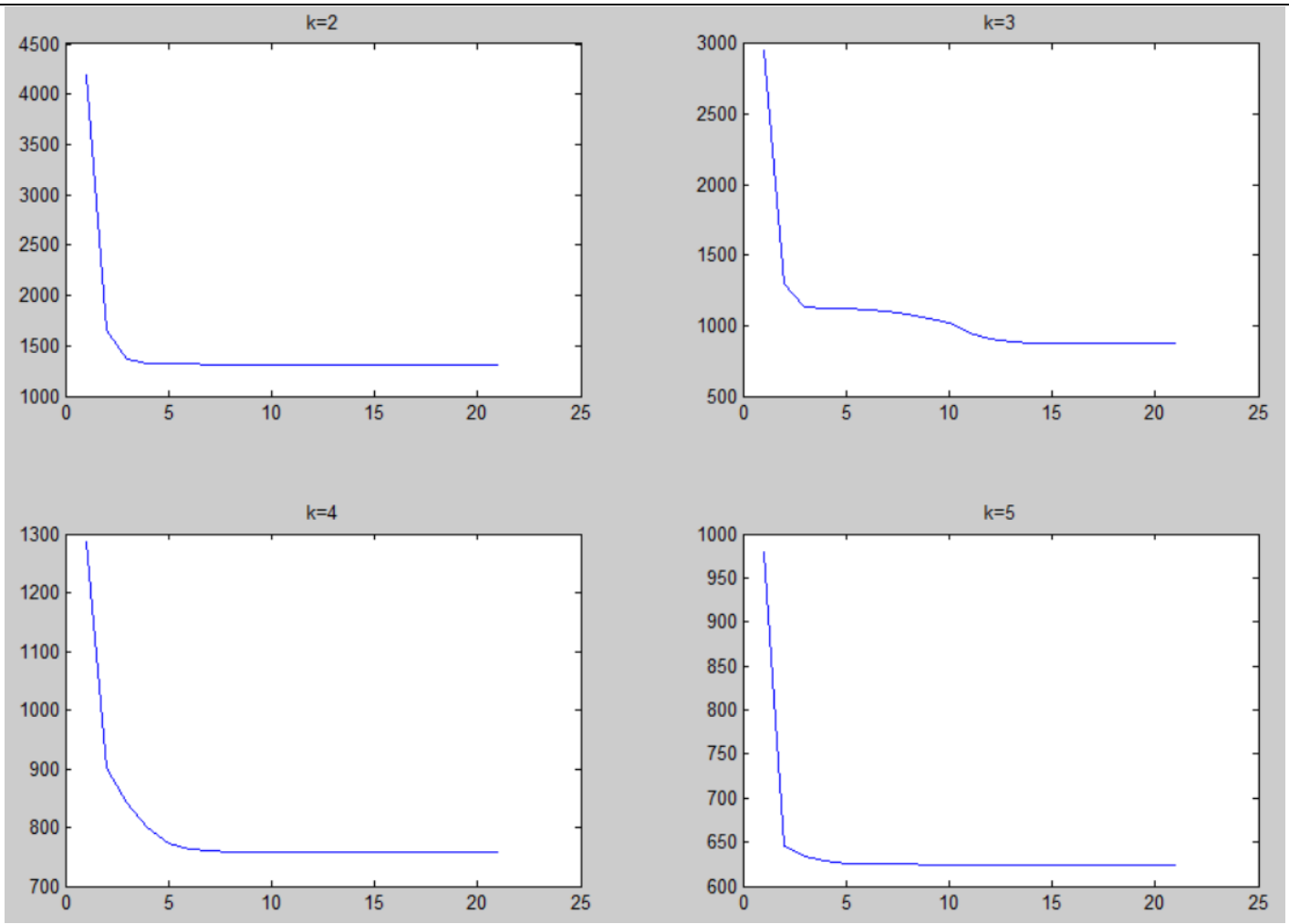
```
        if (k==3)
                group1 = alldat(find(nextCinfo==1),:);
                group2 = alldat(find(nextCinfo==2),:);
                group3 = alldat(find(nextCinfo==3),:);

                subplot(1,2,1),
                scatter(group1(:,1),group1(:,2),'b'), hold on,
                scatter(group2(:,1),group2(:,2),'k'), hold on,
                scatter(group3(:,1),group3(:,2),'m'),
                title(['k=3'])
        end


        if (k==5)
                group1 = alldat(find(nextCinfo==1),:);
                group2 = alldat(find(nextCinfo==2),:);
                group3 = alldat(find(nextCinfo==3),:);
                group4 = alldat(find(nextCinfo==4),:);
                group5 = alldat(find(nextCinfo==5),:);

                subplot(1,2,2),
                scatter(group1(:,1),group1(:,2),'b'), hold on,
                scatter(group2(:,1),group2(:,2),'c'), hold on,
                scatter(group3(:,1),group3(:,2),'g'), hold on,
                scatter(group4(:,1),group4(:,2),'k'), hold on,
                scatter(group5(:,1),group5(:,2),'m'),
                title(['k=5'])
        end

end

subplot(2,2,1), plot(errorRec(:,1)), title(['k=2']);
subplot(2,2,2), plot(errorRec(:,2)), title(['k=3']);
subplot(2,2,3), plot(errorRec(:,3)), title(['k=4']);
subplot(2,2,4), plot(errorRec(:,4)), title(['k=5']);
```

calculateMinDistance.m

```
function [x y] = calculateMinDistance(center,datapoints)
```

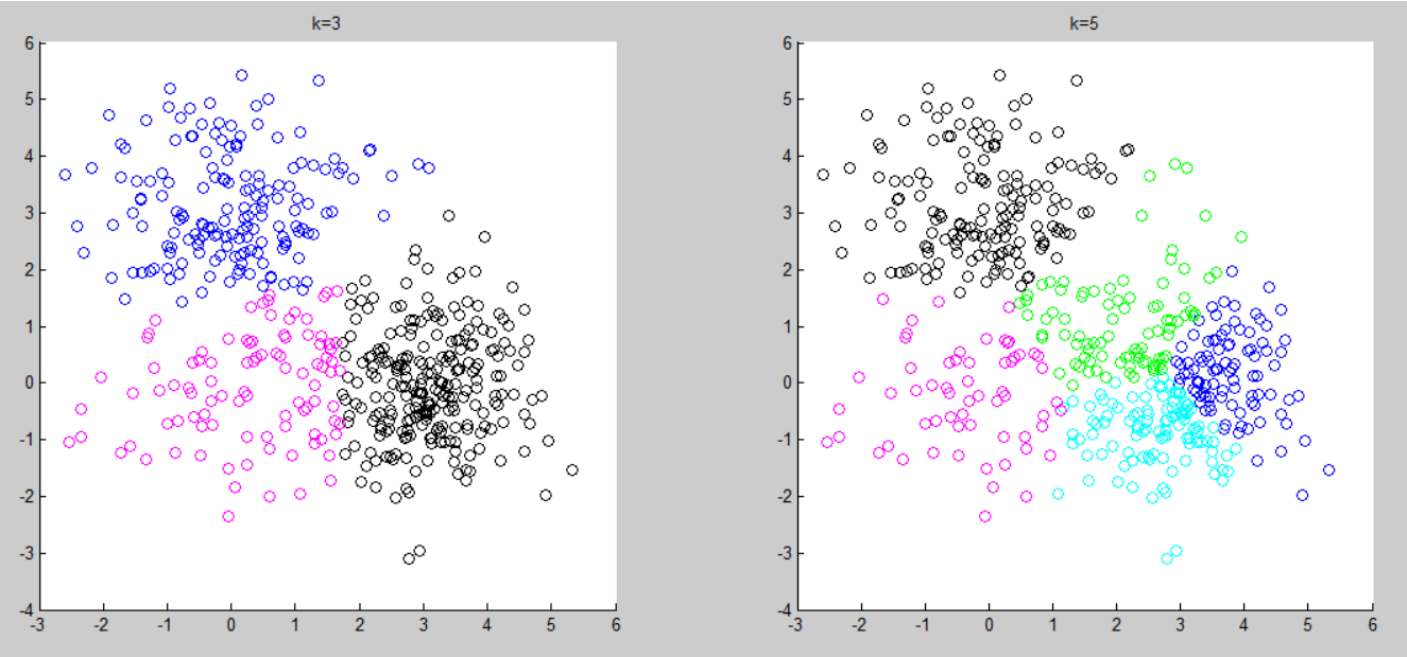
```
% center: m*n - m center points with n dimension
% datapoint: p*n - p points with n dimension
% x: new category information
% y: min dist information
        x = zeros(size(datapoints,1),1);
        y = zeros(size(datapoints,1),1);

        for i = 1:size(datapoints,1)
                data = datapoints(i,:);
                repdata = repmat(data,[size(center,1) 1]); % m*n
                diff = center - repdata; % m*n
                dist = sum(diff.*diff,2); % euclidean distance
                y(i) = min(dist);
                x(i) = find(dist==min(dist));
        end
end
```

K=2, 3, 4, 5



2.

K=3, 5



Problem2

1.

Code:

Hw4pr2.m

```
% hw 4.2
load('movies_matlab_matrix.mat');


%%
%{
fromuser = zeros(943,1682);
for i = 1:943
        user_movie = user(i).movie_id;
        user_rating = user(i).rating;
        fromuser(i,user_movie) = user_rating;
end

frommovie = zeros(943,1682);
for i = 1:1682
        movie_user = movie(i).user_id;
        movie_rating = movie(i).rating;
        frommovie(movie_user,i) = movie_rating;
```

```
end
%}

%%
inim = zeros(1,20);
iniv = 10*eye(20);

userd = mvnrnd(inim,iniv,943);
movied = mvnrnd(inim,iniv,1682);
cst = 2.5;
rec = zeros(943,20,10);

error = zeros(100,1);
logli = zeros(100,1);

%tic

for ite = 1:100
        lg1 = 0; lg2 = 0; lg3 = 0;

        for i = 1:943
                rated = find(fromuser(i,:)~=0); % rated_number*1
                vj = movied(rated,:); % rated_number*20
                former = inv(cst+trace(vj*vj')); %1*1
                temp = repmat((fromuser(i,rated)'),[1 20]); %repmat(rated_number*1,[1
20])
                latter = sum(temp.*vj); %
sum((rated_number*20).X(rated_number*20))=>1*20
                userd(i,:) = former*latter;

                lg2 = lg2+5*(norm(userd(i,:))^2);
        end

        for j = 1:1682
                rated = find(fromuser(:,j)~=0); % rated_number*1
                ui = userd(rated,:); % rated_number*20;
                former = inv(cst+trace(ui*ui')); % 1*1
                temp = repmat(fromuser(rated,j),[1 20]); %repmat(rated_number*1,[1 20])
= rated_number*20
```

```
                latter = sum(temp.*ui); %
sum((rated_number*20).X(rated_number*20))=>1*20
                movied(j,:) = former*latter;

                lg3 = lg3+5*(norm(movied(j,:))^2);
        end

        diff = zeros(943,1682);
        multi = userd*movied'; % 943*1682
        for i = 1:943
                for j = 1:1682
                        if (fromuser(i,j)~=0)
                                diff(i,j) = multi(i,j)-fromuser(i,j);
                        end
                end
        end
        lg1 = 2*(norm(diff)^2);
        logli(ite) = -(lg1+lg2+lg3);

        errortmp = 0;
        for k = 1:5000
                diff = round(userd(ratings_test(k,1),:)* movied(ratings_test(k,2),:)') -
ratings_test(k,3);
                errortmp = errortmp + diff^2;
        end
        error(ite) = sqrt(errortmp/5000);
end

%toc
```

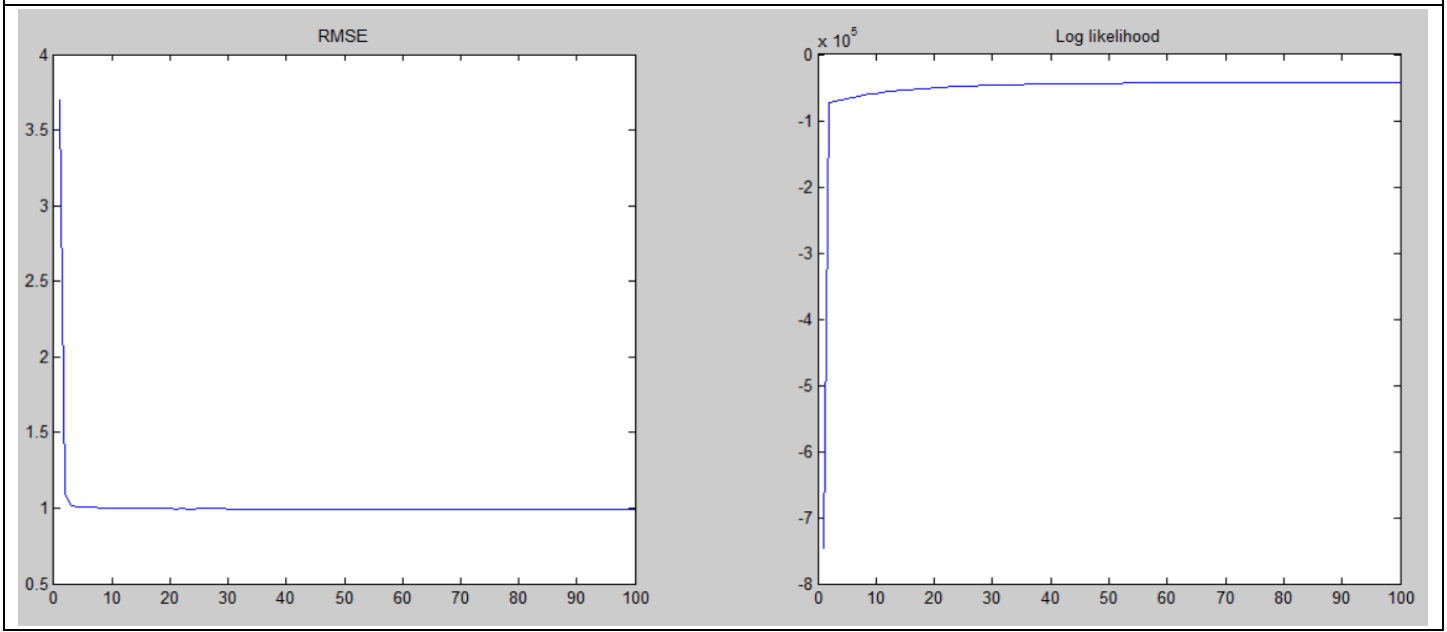
Hw4pr2p2.m

```
% hw4pr2 p2

load('hw4exp1.mat');
load('movies_matlab/movie_ratings.mat');

subplot(1,2,1), plot(error), title(['RMSE']);
subplot(1,2,2), plot(logli), title(['Log likelihood']);

selected_mv = [1101, 1106, 1121];
```

```
closest = zeros(3,6);
mvname = cell(3,6);
dist = zeros(3,6);



for i = 1:3
        mv = movied(selected_mv(i),:); % 1*20
        repmv = repmat(mv,[1682 1]); % 1682*20
        substract = movied-repmv;
        euc = sqrt(sum(substract.^2,2));
        [y idx] = sort(euc);
        closest(i,:) = idx(1:6);
        dist(i,:) = y(1:6);
        mvname(i,:) = movie_names(idx(1:6));
end



errorRec = zeros(21);

rp = randperm(943);
initC = userd(rp(1:30),:);
nextC = zeros(size(initC,1),20);
[nextCinfo, error] = calculateMinDistance(initC,userd);
errorRec(1) = sum(error);

k = 30;
for ite = 1:20
        for ce = 1:k
                temp = userd(find(nextCinfo==ce),:);
                nextC(ce,:) = mean(temp);
        end
        [nextCinfo, error] = calculateMinDistance(nextC,userd);
        errorRec(ite+1) = sum(error);
end

cent = [5 10 15 20 25];%1:30;
farnear = zeros(20,30);
farnear2 = zeros(20,30);
largest = zeros(10,length(cent));
```

```
mn = cell(10,length(cent));
test = zeros(1682,length(cent));
test2 = zeros(1682,length(cent));

for i = 1:length(cent)
        curcent = nextC(cent(i),:);
        prod = zeros(1,1682);
        for j = 1:1682
                prod(j) = dot(curcent,movied(j,:));%curcent*movied';
        end
        [rank idy] = sort(prod,'descend');
        largest(:,i) = rank(1:10);
        mn(:,i) = movie_names(idy(1:10));

        test(:,i) = rank;
        test2(:,i) = idy;
        farnear(1:10,i) = rank(1:10);
        farnear(11:20,i) = rank(1673:1682);
        farnear2(1:10,i) = idy(1:10);
        farnear2(11:20,i) = idy(1673:1682);
end
```

RMSE & Loglikelihood



2.

Selected movie index: 1101, 1106, 1121

Selected Movie	'Six Degrees of Separation (1993)'	'Newton Boys, The (1998)'	'Umbrellas of Cherbourg, The (Parapluies de Cherbourg, Les) (1964)'
Closest Movies 1~5	'Horseman on the Roof, The (Hussard sur le toit, Le) (1995)'	'Of Human Bondage (1934)'	'Star Kid (1997)'
	'My Left Foot (1989)'	'Head Above Water (1996)'	'Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)'
	'Brassed Off (1996)'	'Something to Talk About (1995)'	'Bananas (1971)'
	'Streetcar Named Desire, A (1951)'	'Tetsuo II: Body Hammer (1992)'	'Platoon (1986)'
	'Top Hat (1935)'	'Bean (1997)'	'Flirting With Disaster (1996)'

Distance	0.0013	0.0003	0.0034
	0.0014	0.0035	0.0038
	0.0015	0.0043	0.0052
	0.0023	0.0047	0.0053
	0.0029	0.0049	0.0054

3.

Chosen centroid: [5, 10, 15, 20, 25]

'Pather Panchali (1955)'	'Pather Panchali (1955)'	'Pather Panchali (1955)'	'Pather Panchali (1955)'	'Pather Panchali (1955)'
'Close Shave, A (1995)'	'Close Shave, A (1995)'	'Close Shave, A (1995)'	'Close Shave, A (1995)'	'Close Shave, A (1995)'
'Wrong Trousers, The (1993)'	'Wrong Trousers, The (1993)'	'Wrong Trousers, The (1993)'	'Wrong Trousers, The (1993)'	'Wrong Trousers, The (1993)'
'Casablanca (1942)'	'Casablanca (1942)'	'Casablanca (1942)'	'Casablanca (1942)'	'Casablanca (1942)'
'Schindler"s List (1993)'	'Schindler"s List (1993)'	'Schindler"s List (1993)'	'Schindler"s List (1993)'	'Schindler"s List (1993)'
'Wallace & Gromit: The Best of Aardman Animation (1996)'	'Wallace & Gromit: The Best of Aardman Animation (1996)'	'Wallace & Gromit: The Best of Aardman Animation (1996)'	'Wallace & Gromit: The Best of Aardman Animation (1996)'	'Wallace & Gromit: The Best of Aardman Animation (1996)'
'Shawshank	'Shawshank	'Shawshank	'Shawshank	'Shawshank

Redemption, The (1994)'	Redemption, The (1994)'	Redemption, The (1994)'	Redemption, The (1994)'	Redemption, The (1994)'
'Rear Window (1954)'	'Rear Window (1954)'	'Rear Window (1954)'	'Rear Window (1954)'	'Rear Window (1954)'
'12 Angry Men (1957)'	'12 Angry Men (1957)'	'12 Angry Men (1957)'	'12 Angry Men (1957)'	'12 Angry Men (1957)'
'Star Wars (1977)'	'Star Wars (1977)'	'Star Wars (1977)'	'Star Wars (1977)'	'Star Wars (1977)'

Dot product value:				
3.3590	4.3395	4.1992	4.6688	6.0182
3.3554	4.3348	4.1946	4.6637	6.0117
3.3427	4.3184	4.1787	4.6460	5.9889
3.3361	4.3099	4.1702	4.6368	5.9771
3.3252	4.2958	4.1568	4.6217	5.9575
3.3061	4.2712	4.1330	4.5952	5.9234
3.2988	4.2617	4.1239	4.5850	5.9103
3.2539	4.2037	4.0677	4.5226	5.8198
3.2484	4.1965	4.0608	4.5149	5.8199
3.2434	4.1902	4.0547	4.5081	5.8111