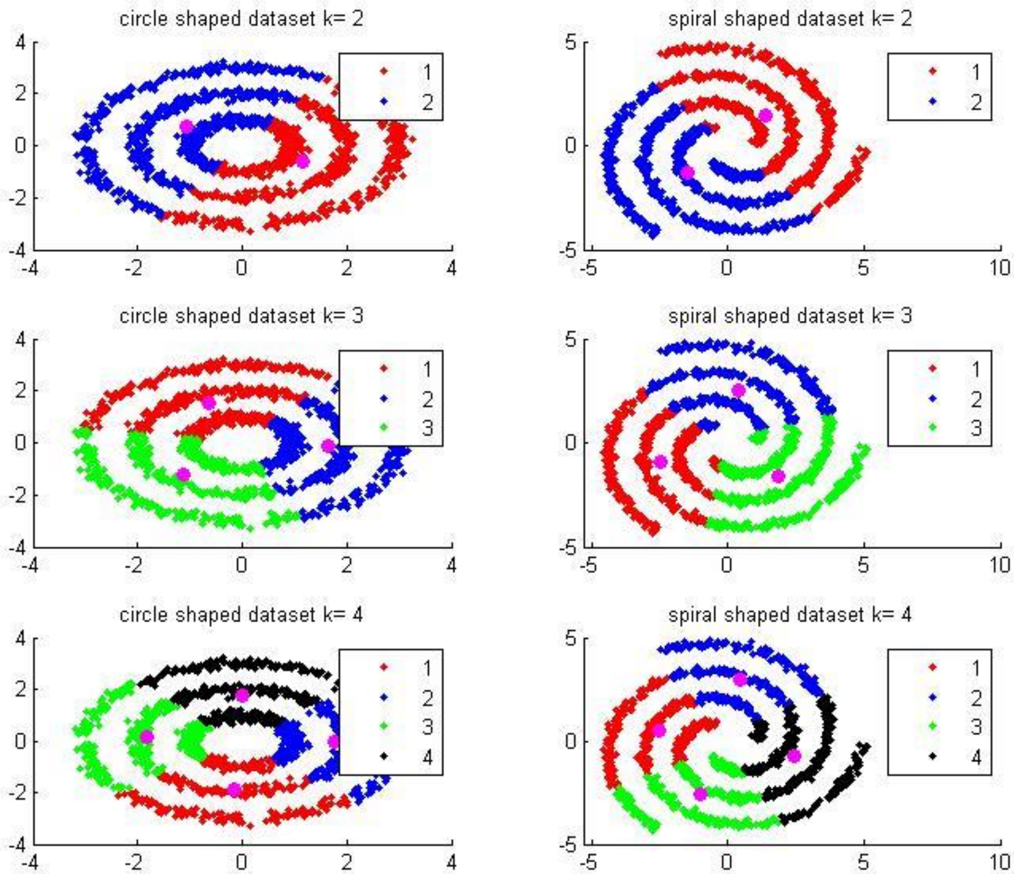
Athar Roshandelpoor MATLAB #4
22 April, 2016 U46981822

Problrm1

a)

i)



As we can see k_means here do not work properly for D1 and D2

ii) Overall within-cluster sums of points of points_to_cluster_centroid (Euclidean) l2 squared distances for each cluster.

For K=2,

	Data D1	Data D2
Cluster 1	2.38×10^3	5.047×10^3
Cluster 2	2.17×10^3	5.012×10^3

For K=3,

	Data D1	Data D2
Cluster 1	1.01×10^3	2.042×10^3
Cluster 2	967.79	1.814×10^3
Cluster 3	911.93	2.103×10^3

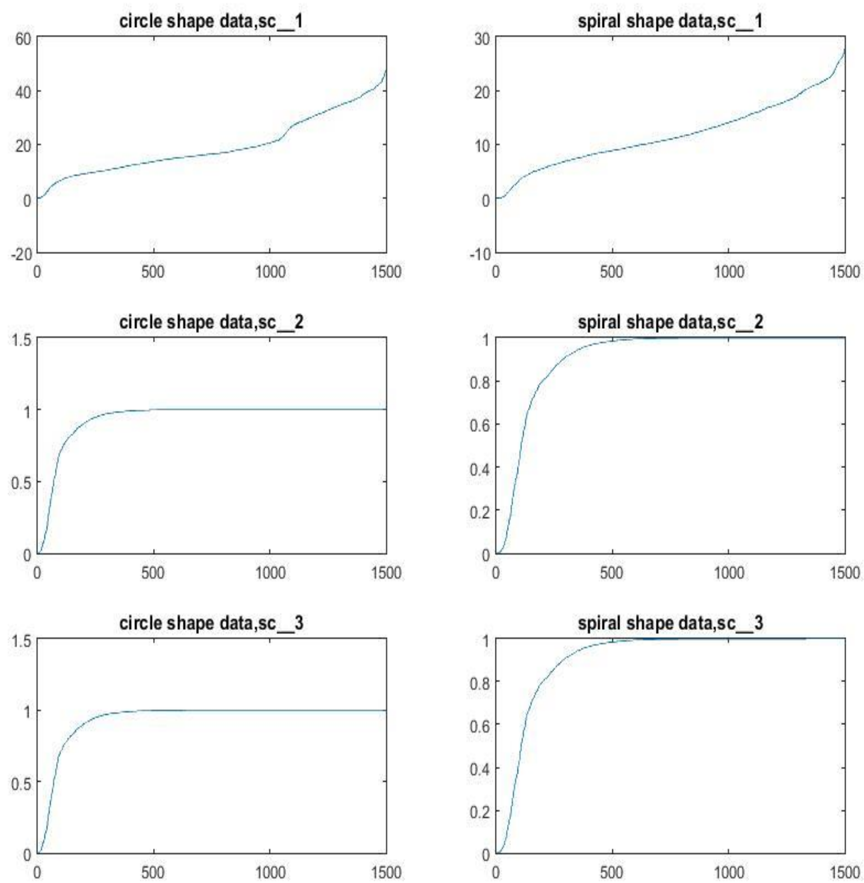
For K=4,

	Data D1	Data D2
Cluster 1	552.66	1.329×10^3
Cluster 2	489.81	1.116×10^3
Cluster 3	495.83	987.27
Cluster 4	581.12	979.74

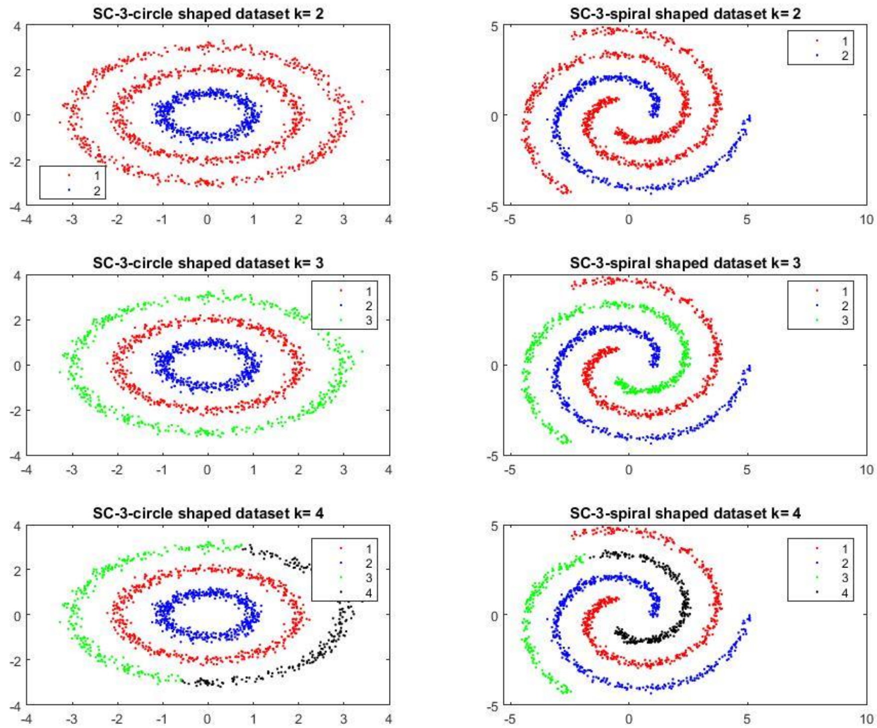
As we can see in each case (D1 and D2) when the number of cluster increasing, the average between all clusters of point to cluster centroid squared distance, decreases.

b) i) Eigenvalues of L, L_{rw} and L_{sym} for D1 and D2.

Note: **y_axis** for all figures in picture below is eigenvalues



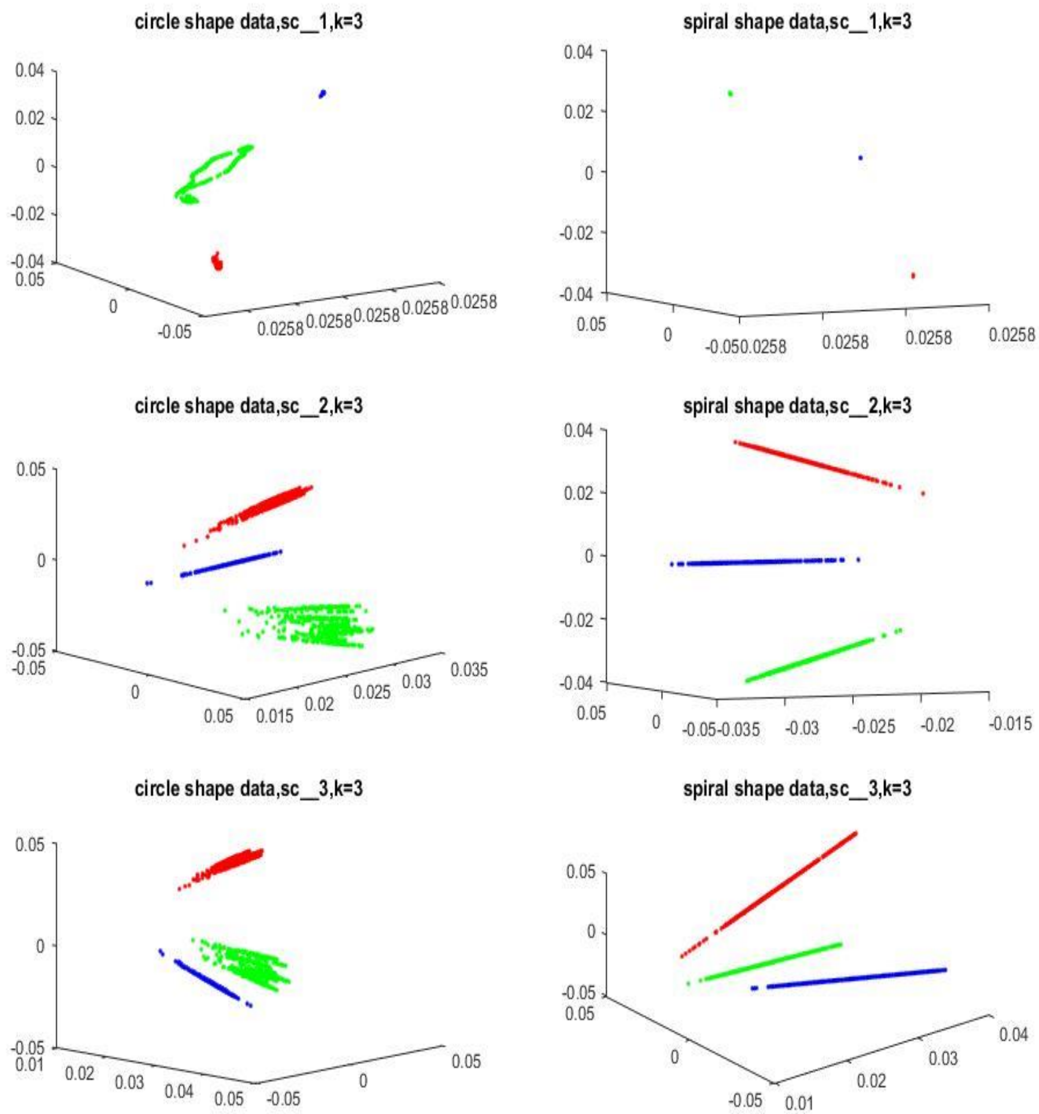
ii)



As it is shown, spectral clustering with L_{sym} works well on case K=3 for both D1 and D2.

So spectral clustering here is better than k_means that we apply in part a.

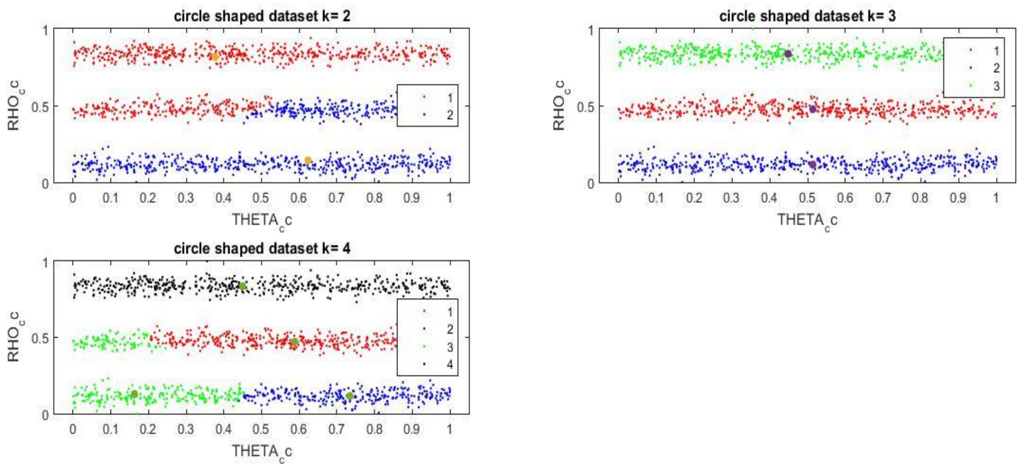
iii)



In spectral clustering, we compute the eigenvectors of L matrix (in a way that we define L), and put the as a column in V matrix. Then each row of V matrix is corresponding to a Data so we cluster the new representation of our dataset. Here as we can see. In new representation data are perfectly separated in three clusters as we can see them in different color. This is why spectral clustering is working well in our data when we set K=3.

c)

i)



As it is shown in the picture, when we transfer data to polar coordinate, data will be in 3 separate groups. So when we apply kmeans clustering to that with k=3, we get each circle in one cluster

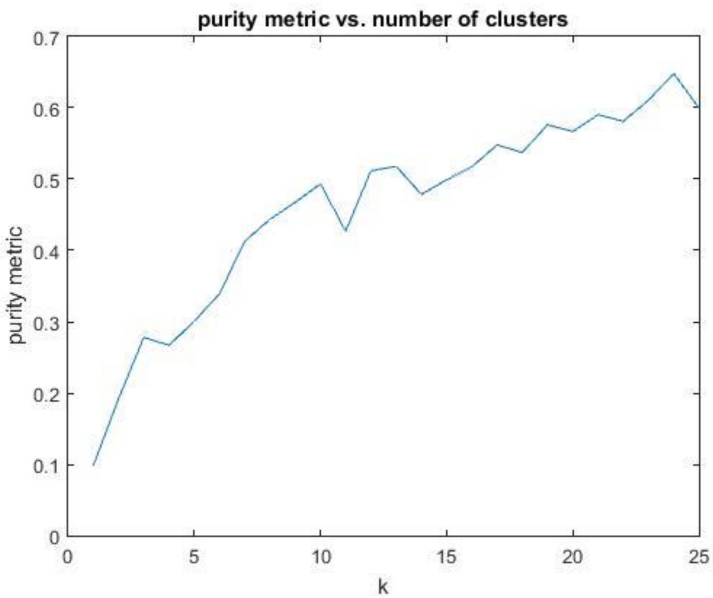
ii) Overall within-cluster sums of points_to_cluster_centroid (Euclidean) l2 squared distances.

	1	2	3	4
K=2	81.85	87.26		
K=3	38.43	43.29	41.25	
K=4	19.28	7.15	16.69	41.25

For k=2 and k=3, clusters are kind of same, so the WCSS numbers of two clusters is close to each other. However, this is not the case for k=4. As it is shown in the picture, cluster 2(blue one) is the smallest and have the less WCSS number.

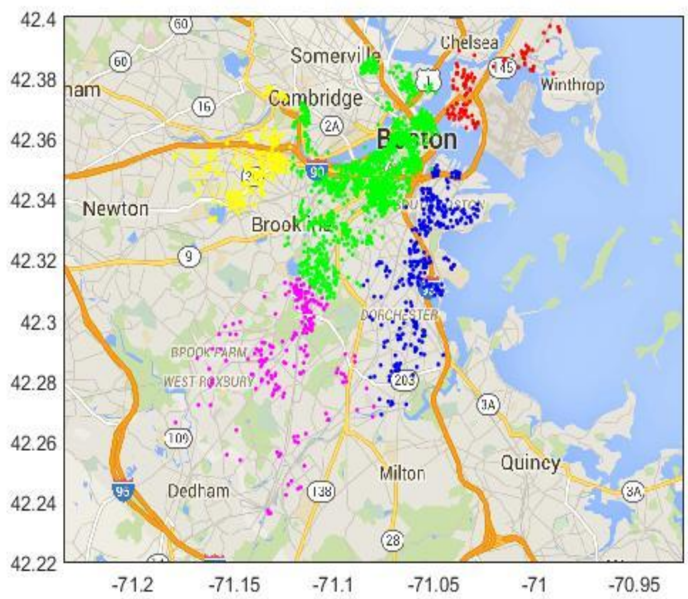
Problem 2

a)



As it is shown in the picture, by increasing number of clusters in spectral clustering (in this problem), the purity number increases. It means spectral algorithm will separate 25 neighborhood in Boston, so it works well.

b)



As it is shown, by setting k=5, we divide Boston to five districts which are represented in picture with five different colors. We have 25 neighborhoods, so according to their distances with each other, near neighborhoods will be in same cluster.

Matlab code

athar_matlab4_1a

```
%%
clear all
clc
close all

[data_circle,label_circle]=sample_circle(3,[500;500;500] );
[data_spiral,label_spiral]=sample_spiral(3,[500;500;500] );


k=1;
l=2;
rng(2)
dist_s_2=zeros(4,4);
dist_c_2=zeros(4,4);

for i=2:4
    center_c=zeros(1500,2);
    center_s=zeros(1500,2);
    [idx1,C1]= kmeans(data_circle,i,'Distance','sqeuclidean',...
        'Replicates',20);

    [idx2,C2]= kmeans(data_spiral,i,'Distance','sqeuclidean',...
        'Replicates',20);

    subplot(3,2,k)
    gscatter(data_circle(:,1),data_circle(:,2),idx1,'rbgk')
    hold on
    scatter(C1(:,1),C1(:,2),'filled','m')
    title(['circle shaped dataset k= ' num2str(i)])
    k=k+2;
    subplot(3,2,l)
    gscatter(data_spiral(:,1),data_spiral(:,2),idx2,'rbgk')
    hold on
    scatter(C2(:,1),C2(:,2),'filled','m')
    title(['spiral shaped dataset k= ' num2str(i)])
    l=l+2;

    for j=1:1500

        center_c(j,:)=C1(idx1(j),:);
        center_s(j,:)=C2(idx2(j),:);
    end
    dist_c=(data_circle-center_c).^2;
    dist_s=(data_spiral-center_s).^2;
    for kk=1:i
        label1=idx1==kk;
         label2=idx2==kk;

    dist_c_2(kk,i-1)=sum(sum(dist_c(label1,:),2));

    dist_s_2(kk,i-1)=sum(sum(dist_s(label2,:),2));
    end
end
```

athar_matlab4_1b

```
%%
clear all
clc
close all

[data_circle,label_circle]=sample_circle(3,[500;500;500] );
[data_spiral,label_spiral]=sample_spiral(3,[500;500;500] );

for j=1:1500
    % W_c(:,j)=-0.5*(1/0.2^2)*sum((data_circle-repmat(data_circle(j,:),1500,1)
).^2,2);
    %
    % W_s(:,j)=-0.5*(1/0.2^2)*sum((data_spiral-repmat(data_spiral(j,:),1500,1)
).^2,2);


    Xm_c = bsxfun(@minus,data_circle,data_circle(j,:));
    W_c(:,j) = dot(Xm_c,Xm_c,2);

    Xm_s = bsxfun(@minus,data_spiral,data_spiral(j,:));
    W_s(:,j) = dot(Xm_s,Xm_s,2);



end
W_c=exp(W_c./(-2*0.2^2));
W_s=exp(W_s./(-2*0.2^2));

D_c=diag(sum(W_c,2));
D_s=diag(sum(W_s,2));


%% Part i
L1_c=D_c-W_c;
L1_s=D_s-W_s;
[V1_c,D1_c] = eig(L1_c);

[~,a_c]=size(V1_c);
norm_V1_c=(sum(V1_c.^2,2)).^0.5;
V1_c_norm=V1_c./repmat(norm_V1_c,1,a_c,1);

[V1_s,D1_s] = eig(L1_s);
[~,I] = sort(diag(D1_s));
V1_s=V1_s(:,I);
[~,a_s]=size(V1_s);
norm_V1_s=(sum(V1_s.^2,2)).^0.5;
V1_s_norm=V1_s./repmat(norm_V1_s,1,a_s,1);

L2_c=D_c^(-1)*(L1_c);
L2_s=D_s^(-1)*(L1_s);
[V2_c,D2_c] = eig(L2_c);
[V2_s,D2_s] = eig(L2_s);
[~,I] = sort(diag(D2_s));
V2_s=V2_s(:,I);

[~,a2_c]=size(V2_c);
norm_V2_c=(sum(V2_c.^2,2)).^0.5;
V2_c_norm=V2_c./repmat(norm_V2_c,1,a2_c,1);
```

```
[~,a2_s]=size(V2_s);
norm_V2_s=(sum(V2_s.^2,2)).^0.5;
V2_s_norm=V2_s./repmat(norm_V2_s,1,a2_s,1);


L3_c=D_c^(-0.5)*(L1_c)*D_c^(-0.5);
L3_s=D_s^(-0.5)*(L1_s)*D_s^(-0.5);
[V3_c,D3_c] = eig(L3_c);

[~,a3]=size(V3_c);
norm_V3_c=(sum(V3_c.^2,2)).^0.5;
V3_c_norm=V3_c./repmat(norm_V3_c,1,a3,1);
[V3_s,D3_s] = eig(L3_s);
[~,I] = sort(diag(D3_s));
V3_s=V3_s(:,I);
[~,a3_s]=size(V3_s);
norm_V3_s=(sum(V3_s.^2,2)).^0.5;
V3_s_norm=V3_s./repmat(norm_V3_s,1,a3_s,1);
figure(1)

subplot(3,2,1)
plot(1:1500,sort(diag(D1_c)))
title('circle shape data,sc__1')
subplot(3,2,2)
plot(1:1500,sort(diag(D1_s)))
title('spiral shape data,sc__1')

subplot(3,2,3)
plot(1:1500,sort(diag(D2_c)))
title('circle shape data,sc__2')
subplot(3,2,4)
plot(1:1500,sort(diag(D2_s)))
title('spiral shape data,sc__2')

subplot(3,2,5)
plot(1:1500,sort(diag(D3_c)))
title('circle shape data,sc__3')
subplot(3,2,6)
plot(1:1500,sort(diag(D3_s)))
title('spiral shape data,sc__3')
%% Part ii
k=1;
l=2;
figure(2)

for i=2:4
    rng(2)
    [idx1,~]= kmeans(V3_c_norm(:,1:i),i);
    rng(2)
    [idx2,~]= kmeans(V3_s_norm(:,1:i),i);

    subplot(3,2,k)
    gscatter(data_circle(:,1),data_circle(:,2),idx1,'rbgk')
    hold on

    title(['SC-3-circle shaped dataset k= ' num2str(i)])
    k=k+2;
    subplot(3,2,l)
    gscatter(data_spiral(:,1),data_spiral(:,2),idx2,'rbgk')
    hold on

    title(['SC-3-spiral shaped dataset k= ' num2str(i)])
    l=l+2;
```

```
end

%%
i=3;

[idx1,~]= kmeans(V1_c_norm(:,1:i),i);

[idx1_s,~]= kmeans(V1_s_norm(:,1:i),i);

[idx2,~]= kmeans(V2_c_norm(:,1:i),i);

[idx2_s,~]= kmeans(V2_s_norm(:,1:i),i);
[idx3,~]= kmeans(V3_c_norm(:,1:i),i);

[idx3_s,~]= kmeans(V3_s_norm(:,1:i),i);

figure(3)

subplot(3,2,1)
hold on
plot3(V1_c_norm(idx1==1,1),V1_c_norm(idx1==1,2),V1_c_norm(idx1==1,3),'r.')
plot3(V1_c_norm(idx1==2,1),V1_c_norm(idx1==2,2),V1_c_norm(idx1==2,3),'b.')
plot3(V1_c_norm(idx1==3,1),V1_c_norm(idx1==3,2),V1_c_norm(idx1==3,3),'g.')
title('circle shape data,sc__1,k=3')
hold off

subplot(3,2,2)
hold on
plot3(V1_s_norm(idx1==1,1),V1_s_norm(idx1==1,2),V1_s_norm(idx1==1,3),'r.')
plot3(V1_s_norm(idx1==2,1),V1_s_norm(idx1==2,2),V1_s_norm(idx1==2,3),'b.')
plot3(V1_s_norm(idx1==3,1),V1_s_norm(idx1==3,2),V1_s_norm(idx1==3,3),'g.')
title('spiral shape data,sc__1,k=3')
hold off

subplot(3,2,3)
hold on
plot3(V2_c_norm(idx1==1,1),V2_c_norm(idx1==1,2),V2_c_norm(idx1==1,3),'r.')
plot3(V2_c_norm(idx1==2,1),V2_c_norm(idx1==2,2),V2_c_norm(idx1==2,3),'b.')
plot3(V2_c_norm(idx1==3,1),V2_c_norm(idx1==3,2),V2_c_norm(idx1==3,3),'g.')
title('circle shape data,sc__2,k=3')
hold off

subplot(3,2,4)
hold on
plot3(V2_s_norm(idx1==1,1),V2_s_norm(idx1==1,2),V2_s_norm(idx1==1,3),'r.')
plot3(V2_s_norm(idx1==2,1),V2_s_norm(idx1==2,2),V2_s_norm(idx1==2,3),'b.')
plot3(V2_s_norm(idx1==3,1),V2_s_norm(idx1==3,2),V2_s_norm(idx1==3,3),'g.')
title('spiral shape data,sc__2,k=3')
hold off

subplot(3,2,5)
hold on
plot3(V3_c_norm(idx1==1,1),V3_c_norm(idx1==1,2),V3_c_norm(idx1==1,3),'r.')
plot3(V3_c_norm(idx1==2,1),V3_c_norm(idx1==2,2),V3_c_norm(idx1==2,3),'b.')
plot3(V3_c_norm(idx1==3,1),V3_c_norm(idx1==3,2),V3_c_norm(idx1==3,3),'g.')
title('circle shape data,sc__3,k=3')
hold off

subplot(3,2,6)
hold on
plot3(V3_s_norm(idx1==1,1),V3_s_norm(idx1==1,2),V3_s_norm(idx1==1,3),'r.')
plot3(V3_s_norm(idx1==2,1),V3_s_norm(idx1==2,2),V3_s_norm(idx1==2,3),'b.')
```

```
plot3(V3_s_norm(idx1==3,1),V3_s_norm(idx1==3,2),V3_s_norm(idx1==3,3),'g.')
title('spiral shape data,sc__3,k=3')
hold off
```

athar_matlab4_1c

```
%%
clear all
clc
close all

[data_circle,label_circle]=sample_circle(3,[500;500;500] );

[THETA_c,RHO_c] = cart2pol(data_circle(:,1),data_circle(:,2));
k=1;
rng(2)
%linear scaling
THETA_cc=(THETA_c-min(THETA_c).*ones(length(THETA_c),1))./(max(THETA_c)-min(THETA_c));
RHO_cc=(RHO_c-min(RHO_c).*ones(length(RHO_c),1))./(max(RHO_c)-min(RHO_c));

dist_c_2=zeros(4,4);
sum1=zeros(4,3)
Data=[THETA_cc,RHO_cc];
for i=2:4
    center_c=zeros(1500,2);
    rng(2)
    [idx1,C1,sum1(1:i,i-1)]= kmeans(Data,i,'Distance','cityblock',...
        'Replicate',20);



    subplot(3,2,k)
    gscatter(THETA_cc,RHO_cc,idx1,'rbgk')
    hold on
    scatter(C1(:,1),C1(:,2),'filled')
    title(['circle shaped dataset k= ' num2str(i)])
    k=k+1;
    %     C1(:,1)=min(THETA_c)+C1(:,1)*(max(THETA_c)-min(THETA_c));
    %     C1(:,2)=min(RHO_c)+C1(:,2)*(max(RHO_c)-min(RHO_c));
    %     [C1(:,1),C1(:,2)]=pol2cart(C1(:,1),C1(:,2));
    %
    %     for j=1:1500
    %
    %         center_c(j,:)=C1(idx1(j),:);
    %
    %     end
    %
    %     for j=1:1500
    %
    %         center_c(j,:)=C1(idx1(j),:);
    %
    %     end
    %     dist_c=(data_circle-center_c).^2;
    %
    %     for kk=1:i
    %         label1=idx1==kk;
    %
    %
    %     dist_c_2(kk,i-1)=sum(sum(dist_c(label1,:),2));
    %
    %
    %      end
    %    for j=1:1500
```

```
    %
    %         center_c(j,:)=C1(idx1(j),:);
    %
    %     end
    %      dist_c=([THETA_c,RHO_c]-center_c).^2;
    % %     [dist_cc(:,1),dist_cc(:,2)]=pol2cart(dist_c(:,1),dist_c(:,2));
    % %     dist_cc=dist_cc.^2;
    %     %     dist_c_2(i)=sum(sum(dist_c,2));
    %     for kk=1:i
    %         label1=idx1==kk;
    %         dist_c_2(kk,i-1)=sum(sum(dist_c(label1,:),2));
    %     end
    %
end
```

athar_matlab4_2a_b

```
%%
clear
clc
load('BostonListing.mat');
data=[latitude, longitude];

for j=1:2558
    % W(:,j)=-0.5*(1/0.1^2)*sum((data-repmat(data(j,:),2558,1) ).^2,2);
    Xm = bsxfun(@minus,data,data(j,:));
    W(:,j) = dot(Xm,Xm,2);

end
W=exp(W./(-2*0.1^2));
D=diag(sum(W,2));
L1=D-W;
L3=D^(-0.5)*(L1)*D^(-0.5);
L3 = 1/2*(L3+L3');
[V3,D3] = eig(L3);
[~,I] = sort(diag(D3));
V3=V3(:,I);
[~,a3]=size(V3);
norm_V3=(sum(V3.^2,2)).^0.5;
V3_norm=V3./repmat(norm_V3,1,a3);

%preparing the ground _truth classes
A=unique(neighbourhood);
AA=zeros(2558,1);
for k=1:length(A)
    tf = strcmp(neighbourhood,A(k));
    AA(tf)=k*ones(sum(tf),1);
end


for k=1:25
    rng(2)
    [idx1,~]= kmeans(V3_norm(:,1:k),k);

    for i=1:k
        for j=1:25
            label=idx1==i;
            n(i,j)=sum(AA(label)==j);

        end

    end
    nn=max(n,[],2);
```

```
    Purity_metric(k)=sum(nn)/2558;
end
figure(1)
plot(1:25,Purity_metric)
title('purity metric vs. number of clusters')
xlabel('k')
ylabel('purity metric')

%% Part B
figure(2)
k=5;

rng(2)
[idx2,~]= kmeans(V3_norm(:,1:k),k);


plot(longitude(idx2==1),latitude(idx2==1),'.r');
hold on;
plot(longitude(idx2==2),latitude(idx2==2),'.b');
plot(longitude(idx2==3),latitude(idx2==3),'.y');
plot(longitude(idx2==4),latitude(idx2==4),'.g');
plot(longitude(idx2==5),latitude(idx2==5),'.m');
plot_google_map;
hold off;
```