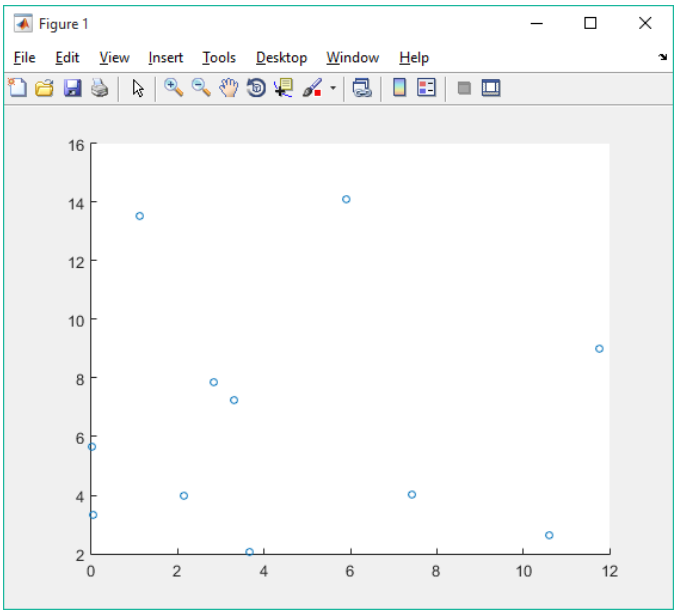
Satrya Budi Pratama
1301154428

Hierarchical Clustering

1. Load Data and Visualize
   - hierarchicalMain.m

```
%% Load the Data
hierarchical_1 = csvread('Clustering_DataSets/Hierarchical/Hierarchical_1.csv');
X = hierarchical_1;

%% Visualize the Data
scatter(X(:,1),X(:,2),20);
```

   - Output



2. 4 Function to measure dissimilarity
   i. Single-Link

```
function  distance = dSingleLink(X , Y)
    % this function is to calculate minimum distance between two cluster
     distance = min(X,Y);
end
```

   ii. Complete –Link

```
function distance = dCompleteLink( X , Y )
    % this function is to choose the max from 2 cluster
    distance = max( X, Y);
end
```

   iii. Group Average

Satrya Budi Pratama
1301154428

```
-Inside\Documents\Machine Learning\Assignment 4.2 Hierarchical Clustering\dAverage.m
1   function distance = dAverage( X , Y)
2       % this function is to calculate the average between to point
3       distance = (X + Y) / 2;
4   end
```

iv. Centroid-based

```
-Inside\Documents\Machine Learning\Assignment 4.2 Hierarchical Clustering\dCentroid.m
1   function distance = dCentroid( X,Y,N,R,i,j,v )
2   % this function is to calculate the distance using centroids approach
3
4    K = N(R(i))+N(R(j));
5    distance = (N(R(i)).*X+N(R(j)) .*Y - (N(R(i)).*N(R(j))*v) ./K) ./K;
6
7   end
```

3. Implement Hierarchical Clustering

Satrya Budi Pratama
1301154428

-Inside\Documents\Machine Learning\Assignment 4.2 Hierarchical Clustering\agglomerative.m

```
function result = agglomerative(Y, method)
% this function is hierarchical clustering implementation
% Y is distance from pdist function
% method is alogrithm approach to merge cluster

n = size(Y,2);       % length of Y (pdist)
m = ceil(sqrt(2*n)); % total of class
result = zeros(m-1,3);    % allocate the output matrix.

N = zeros(1,2*m-1); % N how many points are contained in each cluster.
N(1:m) = 1;
n = m; % since m is changing, we need to save m in n.
R = 1:n; % is a index of class

% Square the distances so updates are easier.
if any(strcmp(method,'centroid'))
   Y = Y .* Y; % only on centroids method
end

    % repeat until cluster = 1
    for s = 1:(n-1)
       [v, k] = min(Y); % search the smallest values and return the index and it

       i = floor(m+1/2-sqrt(m^2-m+1/4-2*(k-1))); % Search the row to merge from
       j = k - (i-1)*(m-i/2)+i; % Search the coloumn to merge from index

       result(s,:) = [R(i) R(j) v]; % get the class from index of i and j and al

       % Update Y.
       I1 = 1:(i-1);
       I2 = (i+1):(j-1);
       I3 = (j+1):m; % these are temp variables
       U = [I1 I2 I3]; % remaining point not yet clustered.
       I = [I1.*(m-(I1+1)/2)-m+i i*(m-(i+1)/2)-m+I2 i*(m-(i+1)/2)-m+I3]; %index
       J = [I1.*(m-(I1+1)/2)-m+j I2.*(m-(I2+1)/2)-m+j j*(m-(j+1)/2)-m+I3]; %inde

       switch method
       case 'single' % single linkage
         Y(I) = dSingleLink(Y(I),Y(J));
       case 'complete' % complete linkage
          Y(I) = dCompleteLink(Y(I),Y(J));
       case 'average' % weighted average linkage
          Y(I) = dAverage(Y(I), Y(J));
       case 'centroid' % centroid linkage
          Y(I) = dCentroid(Y(I),Y(J),N,R,i,j,v);
       end
       J = [J i*(m-(i+1)/2)-m+j];
       Y(J) = []; % no need for the cluster information about j.

       % update m, N, R
       m = m-1;
       N(n+s) = N(R(i)) + N(R(j));
       R(i) = n+s;
       R(j:(n-1))=R((j+1):n);
    end

if any(strcmp(method,'centroid'))
   result(:,3) = sqrt(result(:,3));
end

result(:,[1 2])=sort(result(:,[1 2]),2); % sort from dissmilarity
end
```

- hierarchicalMain.m

Satrya Budi Pratama
1301154428

```
%% do hierarchical clustering
% single-link method
Y = pdist(X);   %calculate euclidean distance
Z = agglomerative(Y,'single');
disp(Z);

% complete-link method
Y = pdist(X);   %calculate euclidean distance
Z = agglomerative(Y,'complete');
disp(Z);

% average group method
Y = pdist(X);   %calculate euclidean distance
Z = agglomerative(Y,'average');
disp(Z);

% centroid method
Y = pdist(X);   %calculate euclidean distance
Z = agglomerative(Y,'centroid');
disp(Z);
```

Output :

- Single-Link

```
Command Window
>> %% do hierarchical clustering
% single-link method
Y = pdist(X);   %calculate euclidean distance
Z = agglomerative(Y,'single');
disp(Z);
    8.0000   10.0000    0.7768
    7.0000   11.0000    2.1869
    9.0000   13.0000    2.3226
    3.0000   14.0000    2.4257
    1.0000    2.0000    3.4645
   12.0000   15.0000    3.4712
   16.0000   17.0000    4.2426
    5.0000    6.0000    4.8313
   18.0000   19.0000    5.9168
    4.0000   20.0000    6.4684
```

- Complete-Link

```
Command Window
>> % complete-link method
Y = pdist(X);   %calculate euclidean distance
Z = agglomerative(Y,'complete');
disp(Z);
    8.0000   10.0000    0.7768
    7.0000   11.0000    2.1869
    9.0000   13.0000    2.6952
    1.0000    2.0000    3.4645
    5.0000    6.0000    4.8313
    3.0000   14.0000    5.0876
   12.0000   17.0000    5.8343
    4.0000   15.0000    6.5989
   16.0000   18.0000   12.2626
   19.0000   20.0000   14.4433
```

- Average

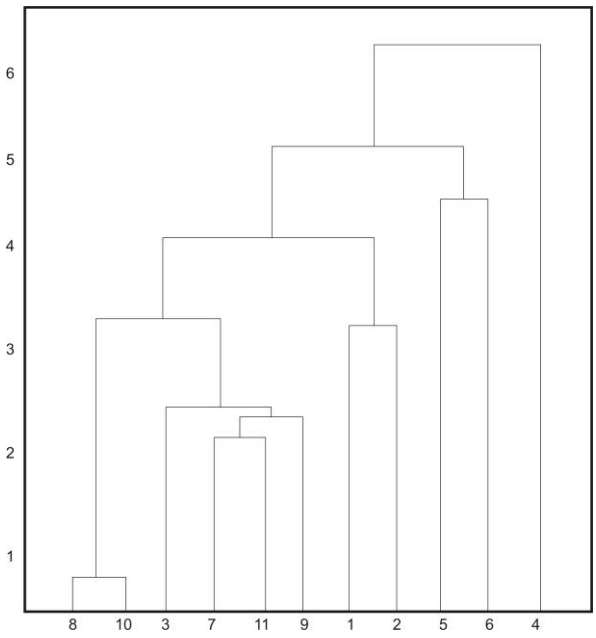
Satrya Budi Pratama
1301154428

```
Command Window

>> % average group method
Y = pdist(X);   %calculate euclidean distance
Z = agglomerative(Y,'average');
disp(Z);
    8.0000   10.0000    0.7768
    7.0000   11.0000    2.1869
    9.0000   13.0000    2.5089
    1.0000    2.0000    3.4645
   12.0000   14.0000    4.0415
    3.0000   16.0000    4.8028
    5.0000    6.0000    4.8313
    4.0000   15.0000    6.5337
   17.0000   19.0000    8.6695
   18.0000   20.0000   10.5188
```

- Centroid

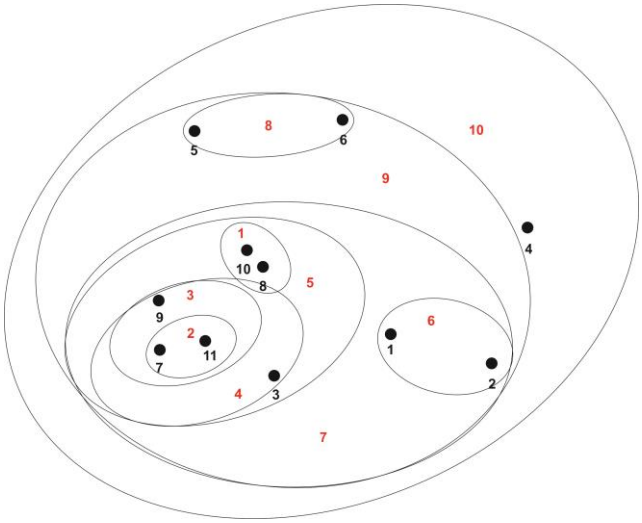
```
Command Window

>> % centroid method
Y = pdist(X);   %calculate euclidean distance
Z = agglomerative(Y,'centroid');
disp(Z);
    8.0000   10.0000    0.7768
    7.0000   11.0000    2.1869
    9.0000   13.0000    2.2658
    1.0000    2.0000    3.4645
    3.0000   14.0000    3.6727
   12.0000   16.0000    4.1225
    5.0000    6.0000    4.8313
    4.0000   15.0000    6.3002
   17.0000   19.0000    7.9334
   18.0000   20.0000    8.7954
```

4. Visualize
    - Single-Link
        - Dendrogram

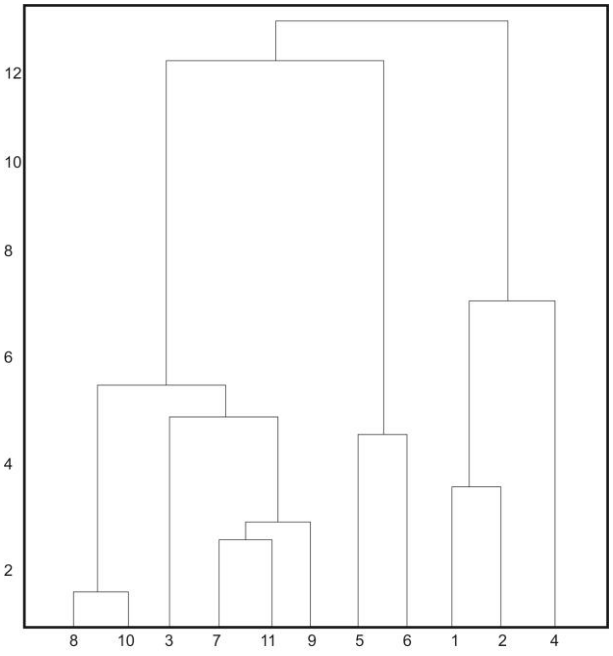


        - Nested Cluster

Satrya Budi Pratama
1301154428

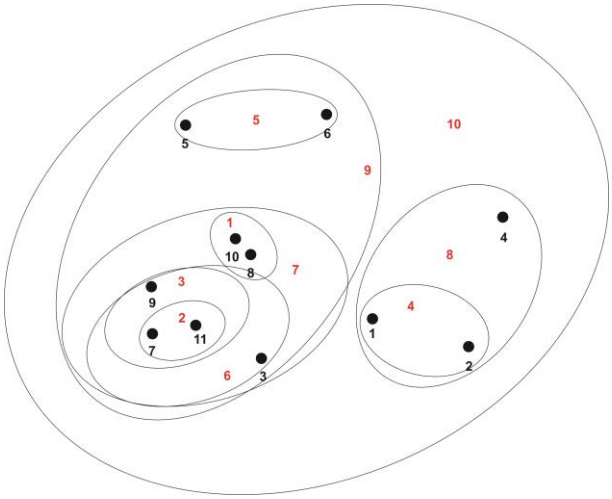


- Complete-Link
  - Dendrogram

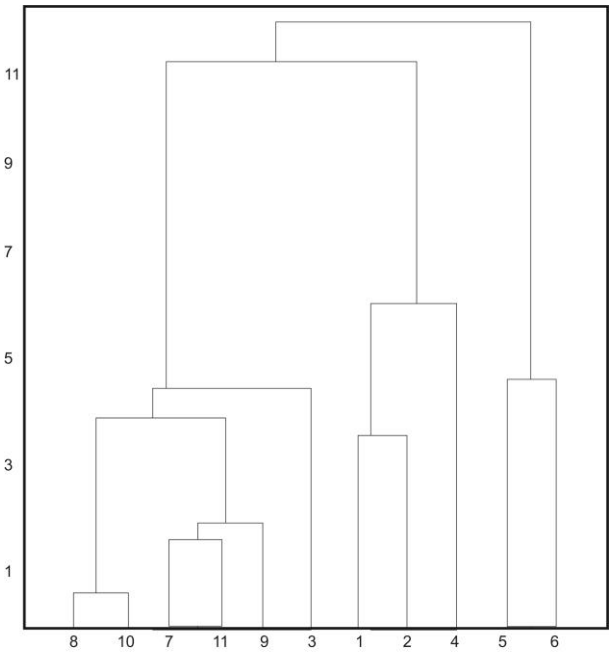


  - Nested Cluster

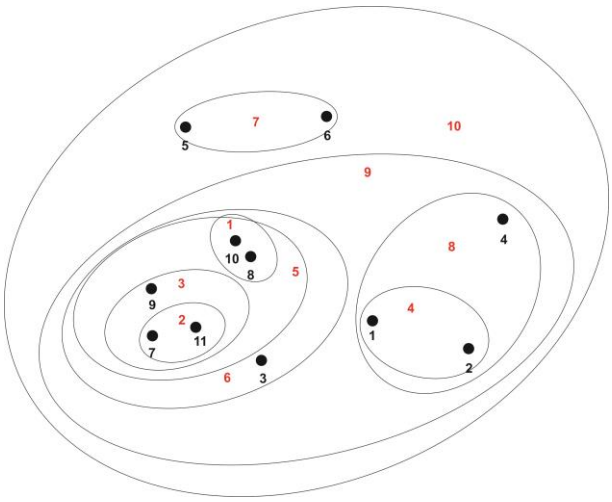
Satrya Budi Pratama
1301154428



- Average
  - Dendrogram

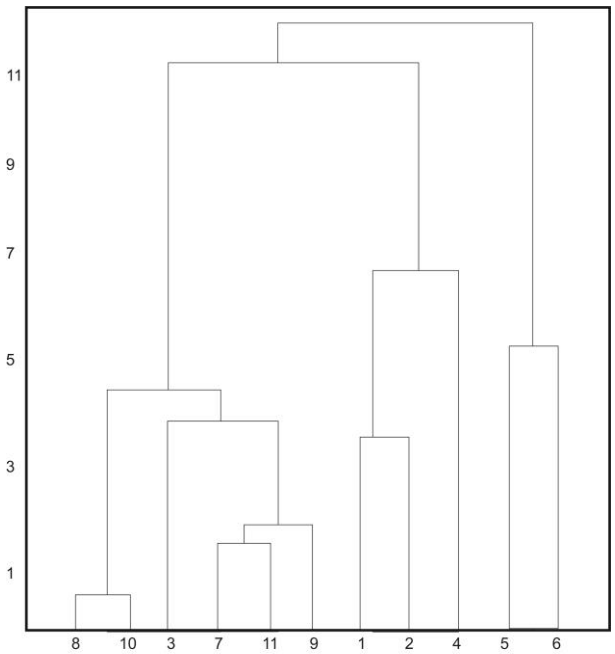


  - Nested Cluster

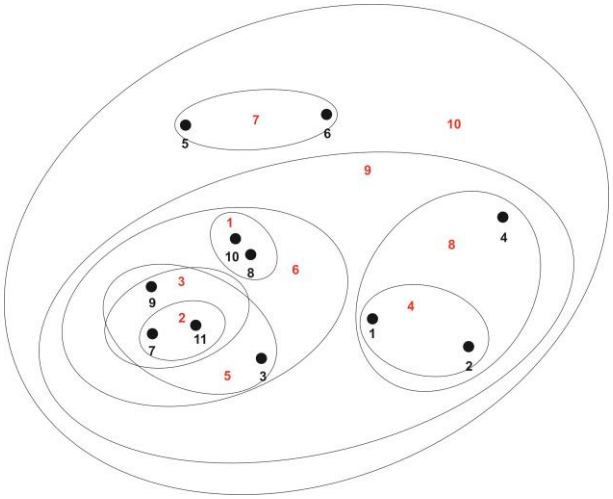


Satrya Budi Pratama
1301154428

- Centroid
  - Dendrogram



  - Nested Cluster



5. Compare Cluster

   Analisis :

   Untuk single-link list sangat cocok untuk cluster yang benar-benar terpisah. Tetapi jika saling berdekatan cendurung mengidentifikasi secara berurutan. Adapun untuk complete-link pengclusteran cendurung memilki diameter yang sama karena menghitung jarak terjauh dari cluster sehingga akan sensitive terhadap data percilan. Kemudian untuk average mempertimbangkan nilai kesluruhan anggota cluster sehingga lebih sentral untuk pengelompokkan cluster. Dan method centroid sama seperti average namun perhitungan melibatkan centroid pada cluster.