CS303B: Lab 8–Clustering

Aim: To understand how to perform clustering using k-means and agglomerative clustering in Matlab.

Take the iris dataset as the example, and use the petal lengths and widths as predictors by ignoring the species information.

https://www.mathworks.com/help/stats/kmeans.html#bue6nb7-1

1. **Visualize the data, and clustering results**.

```
%Load Fisher's iris data set.

load fisheriris
X = meas(:,3:4);
figure;
plot(X(:,1),X(:,2),'k*','MarkerSize',5);
title 'Fisher''s Iris Data';
xlabel 'Petal Lengths (cm)';
ylabel 'Petal Widths (cm)';
```

Question: Looking at the plot, how many clusters can you observe intuitively from this data distribution?

(hints: the larger cluster seems to be split into a lower variance region and a higher variance region. This might indicate that the larger cluster is two, overlapping clusters.)

2. **Cluster the data using kmeans. Specify k = 3 clusters.**

```
rng(1); % For reproducibility
[idx,C] = kmeans(X,3);

x1 = min(X(:,1)):0.01:max(X(:,1));
x2 = min(X(:,2)):0.01:max(X(:,2));
[x1G,x2G] = meshgrid(x1,x2);
XGrid = [x1G(:),x2G(:)]; % Defines a fine grid on the plot

idx2Region = kmeans(XGrid,3,'MaxIter',1,'Start',C);
```

2.1. Visualize the clustering results

```
figure;
```

```
hold on;
%gscatter(XGrid(:,1),XGrid(:,2),idx2Region, [0,0.75,0.75;0.75,0,0.75;0.75,0.75,0],'..');
gscatter(X(:,1),X(:,2),idx, 'rgb','osd');
%plot(X(:,1),X(:,2),'k*','MarkerSize',5);
title 'Fisher''s Iris Data';
xlabel 'Petal Lengths (cm)';
ylabel 'Petal Widths (cm)';
%legend('Region 1','Region 2','Region 3','Data','Location','Best');
hold off;
```

Questions: Now you would be able to see that the data points in iris dataset are now partitioned into three clusters. Do those clusters match your intuition? And do those clusters match the ground truth species of the plant, i.e., data points within the same each species are roughly clustered in one local region?

2.2. Try different distance metric such as 'city block' to see if there is any difference in the results. 'cityblock' -- sum of absolute differences, a.k.a. L1 distance

2.3. Try different attribute features in the iris dataset, and redo clustering, e.g., X
X= meas(:,1:2);

3 Clustering using hierarchical clustering.

```
%Load the sample data.
load fisheriris

%Compute four clusters of the fisheriris data using single-linkage

Z = linkage(meas(:,3:4),'single','euclidean');  %create the linkage tree using single-
%link

c = cluster(Z,'maxclust',3);

% See how the cluster assignments correspond to the three species.

crosstab(c, species)

% Create a dendrogram plot of Z, and visualize it.

dendrogram(Z)
```

Questions: how is the result compared to k-means?