COMS 474
Homework 9

Haadi Majeed

Spring 2022

Problem 1

[26 Points]
Suppose that we have a data set with four samples. You calculate the pair-wise distances as

	{1}	{2}	{3}	{4}
{1}		0.3	0.4	0.7
{2}	0.3		0.5	0.8
{3}	0.4	0.5		0.45
{4}	0.7	0.8	0.45	

$$dist(C_1, C_2) := \min_{a \epsilon C_1, b \epsilon C_2} dist(a, b)$$
$$dist(C_1, C_2) := \max_{a \epsilon C_1, b \epsilon C_2} dist(a, b)$$

A

Sketch the dendrogram that results from hierarchically clustering these four samples using complete linkage. Indicate on the plot the height (distance) at which each fusion occurs, as well as the samples corresponding to each leaf in the dendrogram.

beans

B

Repeat part (a) for single linkage clustering.

beans

C

Suppose that we cut the dendrogram obtained in (a) such that two clusters result. Which samples are in each cluster?

beans

D

Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which samples are in each cluster?

beans

E

Draw a dendrogram that is equivalent to the dendrogram in (a) with a different horizontal arrangement for the samples.

beans

Problem 2

A

Make a scatter-plot

beans

B

For n_clustersīrange(1,16), apply K-means clustering. Make a scatter plot for each. You only need to include 3 of them in your homework submission. Select the three pictures whose clusters you think look the best.

beans

C

For bottom-up hierarchical clustering (aka 'Agglomerative Clustering'), make a dendrogram using 'single' linkage.

beans

D

Manually select and report a distance threshold for single linkage. Look for regions in the dendrogram where there are few mergers (eg a big vertical gap in distance threshold between mergers). Use that to make a scatter plot of the data clustered based on that threshold.

beans

E

Repeat the previous two steps, using 'average' linkage.

beans

F

Repeat the previous two steps, using 'complete' linkage.

beans

G

In a few sentences for each clustering method (k-means, single linkage aggl., average linkage aggl., and complete linkage aggl.), comment on the clusters found by the methods and how they compared or differed. Which clustering method do you think resulted in the best clusters for this data set and why?

Beans