Text, letter

Description automatically generated

Hamamoto Yuichi

Table

Description automatically generated with medium confidence

A picture containing text, whiteboard

Description automatically generated



b. Repeat part (a) for single linkage clustering.

A picture containing text, whiteboard

Description automatically generated

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

1,2 and 3,4

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

4 and (3, (1,2)

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

A picture containing text, whiteboard

Description automatically generated

**Problem 2.** [20 points] In this problem you will run several clustering procedures and compare them on an example data set, HW9data.csv. Example Python code is posted as HW9-cluster-template.ipynb; you may need to make a few, minor changes.

1. Make a scatter-plot of the data, coloring each data point black.

Chart, scatter chart

Description automatically generated

1. For n\_clusters=range(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.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

1. For bottom-up hierarchical clustering (aka ‘Agglomerative Clustering’), make a dendo- gram using ‘single’ linkage.

Chart, histogram

Description automatically generated

1. Manually select and report a distance threshold for single linkage. Look for regions in the dendogram 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.

Chart, scatter chart

Description automatically generated

1. Repeat the previous two steps, using ‘average’ linkage.

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

1. Repeat the previous two steps, using ‘complete’ linkage.

Chart

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

1. 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?

Average one and complete one look fairly similar and nicely clustered whereas single one is not nicely spread out. I think the one using average linkage is better than others for this dataset. It is because compared to the rest of them, this one is nicely evenly spread out.