CZ4041/CE4041: Machine Learning

Week 10: Clustering

Nerd Joke Time

- Hollywood Executive: Give me an idea for a movie.
- Me: It's a movie about high-school girls trying to figure out which clique they belong to. They move from one clique to the next until they minimize their differences. It's called K-means Girls.

Question 1

- MIN or Single Link: Distance of two clusters is based on the two most closest points in the different clusters

	P1	P2	P3	P4	P5
P1	0.00	0.90	0.10	0.65	0.20
P2	0.90	0.00	0.70	0.60	0.50
P3	0.10	0.70	0.00	0.40	0.30
P4	0.65	0.60	0.40	0.00	0.80
P5	0.20	0.50	0.30	0.80	0.00

Distance matrix

Agglomerative Clustering Algorithm

- Basic algorithm:
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. **Until** only a single cluster remains

distance or similarity

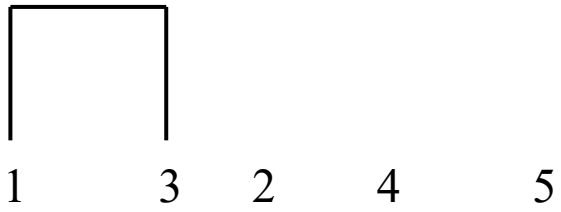
smallest distance or largest similarity

Question 1 (cont.)

Step 1: Merge the two closest clusters (smallest distance)

	P1	P2	P3	P4	P5
P1	0.00	0.90	0.10	0.65	0.20
P2	0.90	0.00	0.70	0.60	0.50
P3	0.10	0.70	0.00	0.40	0.30
P4	0.65	0.60	0.40	0.00	0.80
P5	0.20	0.50	0.30	0.80	0.00

Distance matrix

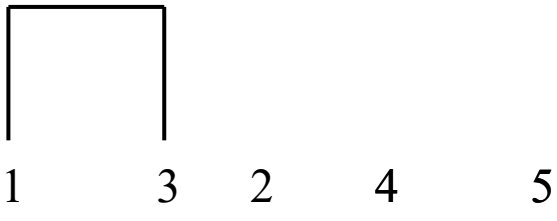


Question 1 (cont.)

- Step 2: Update proximity matrix based on MIN: proximity of two clusters is based on the two closest points in different clusters (smallest distance)

	P1∪P3	P2	P4	P5
P1∪P3	0.00	0.70	0.40	0.20
P2	0.70	0.00	0.60	0.50
P4	0.40	0.60	0.00	0.80
P5	0.20	0.50	0.80	0.00

Distance matrix

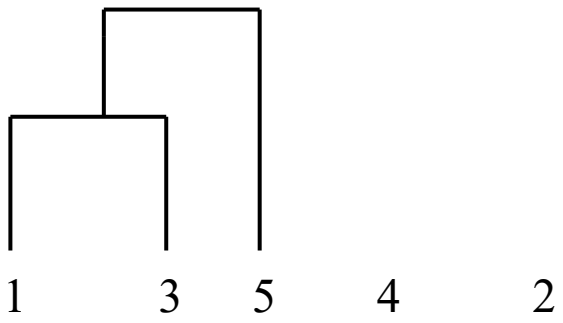


Question 1 (cont.)

Step 1: Merge the two closest clusters (smallest distance)

	P1∪P3	P2	P4	P5
P1∪P3	0.00	0.70	0.40	0.20
P2	0.70	0.00	0.60	0.50
P4	0.40	0.60	0.00	0.80
P5	0.20	0.50	0.80	0.00

Distance matrix

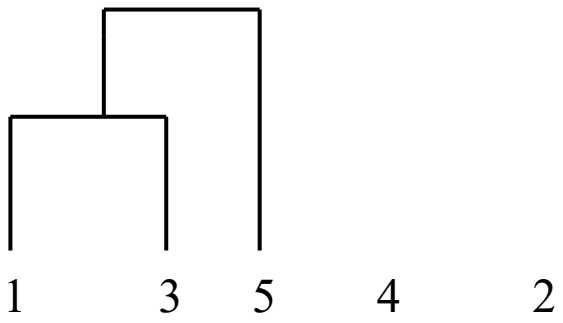


Question 1 (cont.)

- Step 2: Update proximity matrix based on MIN: proximity of two clusters is based on the two closest points in different clusters (smallest distance)

	P1∪P3∪P5	P2	P4
P1∪P3∪P5	0.00	0.50	0.40
P2	0.50	0.00	0.60
P4	0.40	0.60	0.00

Distance matrix

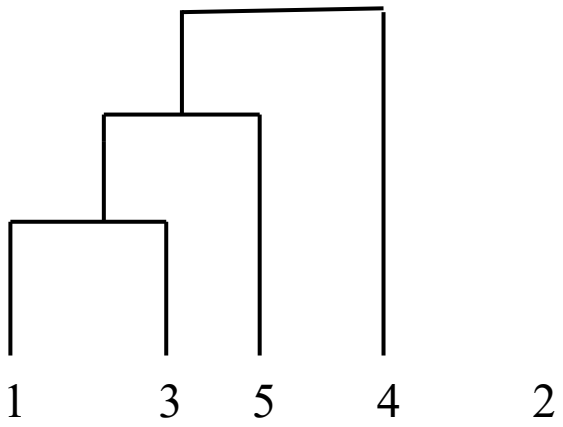


Question 1 (cont.)

Step 1: Merge the two closest clusters (smallest distance)

	P1∪P3∪P5	P2	P4
P1∪P3∪P5	0.00	0.50	0.40
P2	0.50	0.00	0.60
P4	0.40	0.60	0.00

Distance matrix

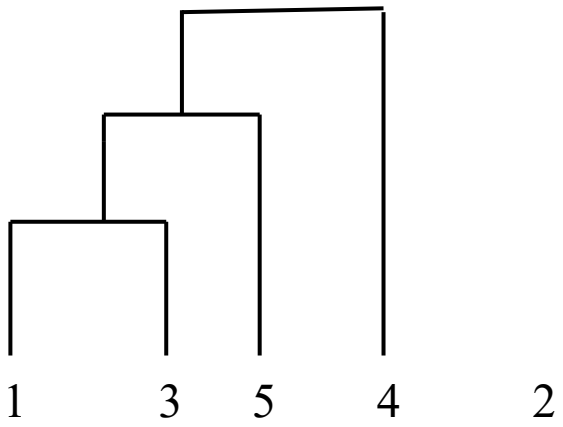


Question 1 (cont.)

- Step 2: Update proximity matrix based on MIN: proximity of two clusters is based on the two closest points in different clusters (smallest distance)

	P1∪P3∪P5∪P4	P2
P1∪P3∪P5∪P4	0.00	0.50
P2	0.50	0.00

Distance matrix

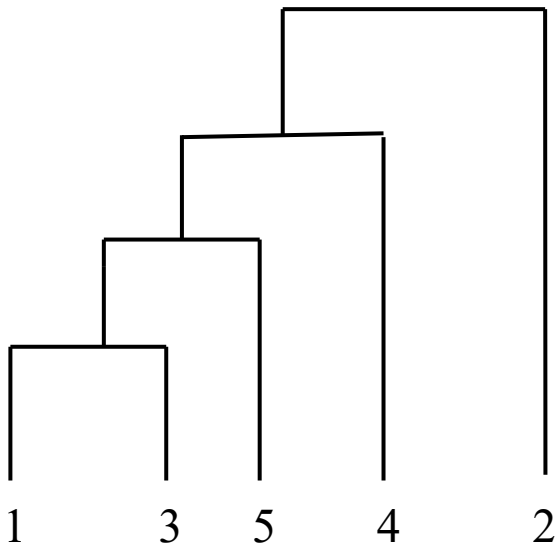


Question 1 (cont.)

Step 1: Merge the two closest clusters (smallest distance)

	P1∪P3∪P5∪P4	P2
P1∪P3∪P5∪P4	0.00	0.50
P2	0.50	0.00

Distance matrix

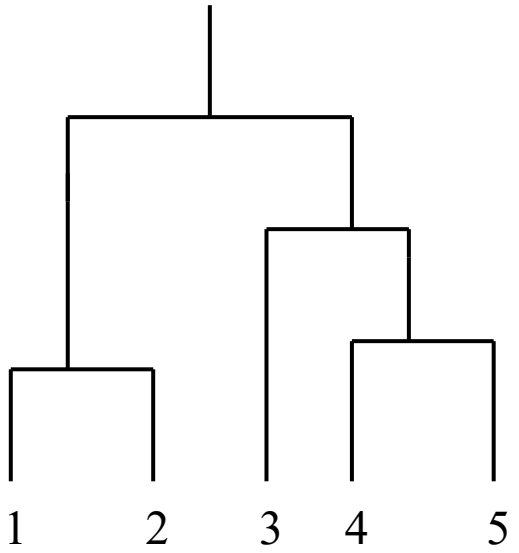


Question 2

- MAX or Complete Link: Similarity of two clusters is based on the two least similar points in the different clusters

	P1	P2	P3	P4	P5
P1	1.00	0.90	0.10	0.65	0.20
P2	0.90	1.00	0.70	0.60	0.50
P3	0.10	0.70	1.00	0.40	0.30
P4	0.65	0.60	0.40	1.00	0.80
P5	0.20	0.50	0.30	0.80	1.00

Similarity matrix



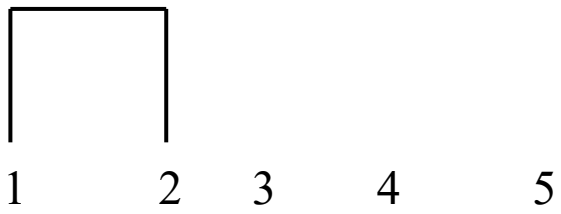
Tutorial

Question 2 (cont.)

Step 1: Merge the two closest clusters (largest similarity)

	P1	P2	P3	P4	P5
P1	1.00	0.90	0.10	0.65	0.20
P2	0.90	1.00	0.70	0.60	0.50
P3	0.10	0.70	1.00	0.40	0.30
P4	0.65	0.60	0.40	1.00	0.80
P5	0.20	0.50	0.30	0.80	1.00

Similarity matrix

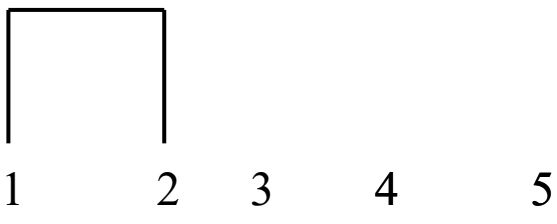


Question 2 (cont.)

- Step 2: Update proximity matrix based on MAX: proximity of two clusters is based on the two farthest points in different clusters (smallest similarity)

	P1∪P2	P3	P4	P5
P1∪P2	1.00	0.10	0.60	0.20
P3	0.10	1.00	0.40	0.30
P4	0.60	0.40	1.00	0.80
P5	0.20	0.30	0.80	1.00

Similarity matrix

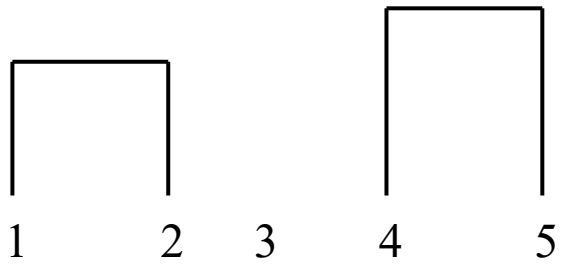


Question 2 (cont.)

Step 1: Merge the two closest clusters (largest similarity)

	P1∪P2	P3	P4	P5
P1∪P2	1.00	0.10	0.60	0.20
P3	0.10	1.00	0.40	0.30
P4	0.60	0.40	1.00	0.80
P5	0.20	0.30	0.80	1.00

Similarity matrix

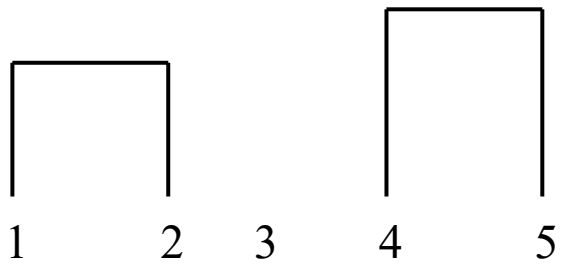


Question 2 (cont.)

- Step 2: Update proximity matrix based on MAX: proximity of two clusters is based on the two farthest points in different clusters (smallest similarity)

	P1∪P2	P3	P4∪P5	
P1∪P2	1.00	0.10	0.20	
P3	0.10	1.00	0.30	
P4∪P5	0.20	0.30	1.00	
	0.20	0.30	0.80	1.00

Similarity matrix

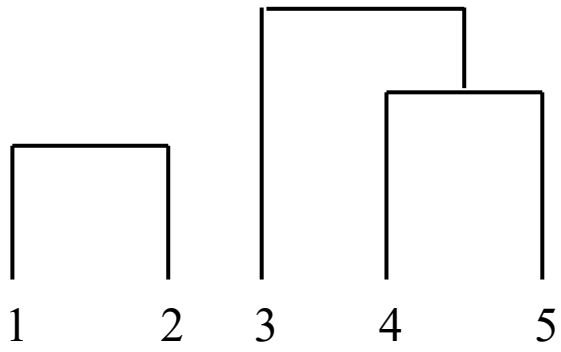


Question 2 (cont.)

Step 1: Merge the two closest clusters (largest similarity)

	P1∪P2	P3	P4∪P5
P1∪P2	1.00	0.10	0.20
P3	0.10	1.00	0.30
P4∪P5	0.20	0.30	1.00

Similarity matrix

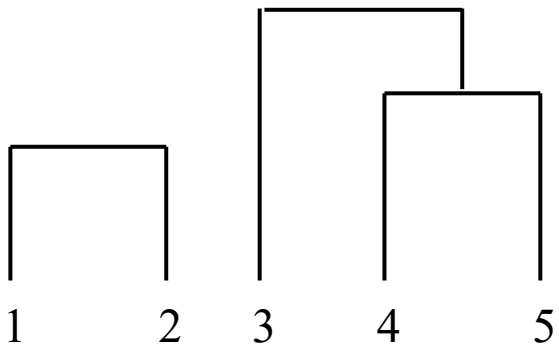


Question 2 (cont.)

- Step 2: Update proximity matrix based on MAX: proximity of two clusters is based on the two farthest points in different clusters (smallest similarity)

	P1∪P2	P3 ∪P4∪P5	
P1∪P2	1.00	0.10	
P3∪P4∪P5	0.10	1.00	
	0.20	0.30	1.00

Similarity matrix



Question 2 (cont.)

Step 1: Merge the two closest clusters (largest similarity)

	P1∪P2	P3∪P4∪P5
P1∪P2	1.00	0.10
P3∪P4∪P5	0.10	1.00

Similarity matrix