COMP 4432 Machine Learning

Lesson 9: Unsupervised Learning

Agenda

- Gaussian Mixtures
- DBScan
- Agglomerative

Brief Review

- Partition *n* instances of data into *k* groups
  - *k* is less than or equal to *n*
- The correct *answer* isn’t know beforehand
- Multiple algorithms
  - K-Means
  - Gaussian Mixtures
  - DBSCAN
  - Agglomerative

Brief Review

- Partition *n* instances of data into *k* groups
  - *k* is less than or equal to *n*
- The correct *answer* isn't know beforehand
- Multiple algorithms
  - K-Means
  - Gaussian Mixtures
  - DBSCAN
  - Agglomerative

Gaussian Mixtures

- Assumes the data originated from a mixture of several Gaussian (Normal) distributions.

Gaussian Mixtures

- Assumes the data originated from a mixture of several Gaussian (Normal) distributions.
  - Modeling a multimodal distribution could be simplified by assuming that it is a collection of multiple simple, unimodal distributions.

Gaussian Mixtures

- Assumes the data originated from a mixture of several Gaussian (Normal) distributions.
  - Modeling a multimodal distribution could be simplified by assuming that it is a collection of multiple simple, unimodal distributions.
  - The Gaussian distribution is frequently employed to model real-world unimodal data.

Gaussian Mixtures

- Assumes the data originated from a mixture of several Gaussian (Normal) distributions.
  - Modeling a multimodal distribution could be simplified by assuming that it is a collection of multiple simple, unimodal distributions.
  - The Gaussian distribution is frequently employed to model real-world unimodal data.
- The distribution parameters (mean and variance) are unknown.

Gaussian Mixtures

- Assumes the data originated from a mixture of several Gaussian (Normal) distributions.
  - Modeling a multimodal distribution could be simplified by assuming that it is a collection of multiple simple, unimodal distributions.
  - The Gaussian distribution is frequently employed to model real-world unimodal data.
- The distribution parameters (mean and variance) are unknown.
- Model learns which distribution a data point belongs.

Gaussian Mixtures

- Very High Level…
  - [SciKit Documentation]
  - Start with *k* means (“centroids”), and *k* equal covariances and class weights
  - Update each by employing Expectation Maximization

Gaussian Mixtures

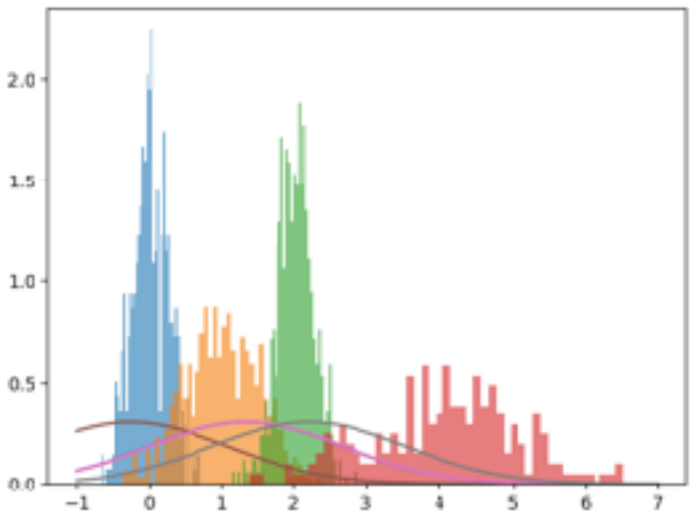
$$\gamma_{i,k} = \frac{\phi_k \mathcal{N}(x_i|\mu_k \Sigma_k)}{\sum_k \phi_k \mathcal{N}(x_i|\mu_k \Sigma_k)} = P(Class = k|x_i)$$

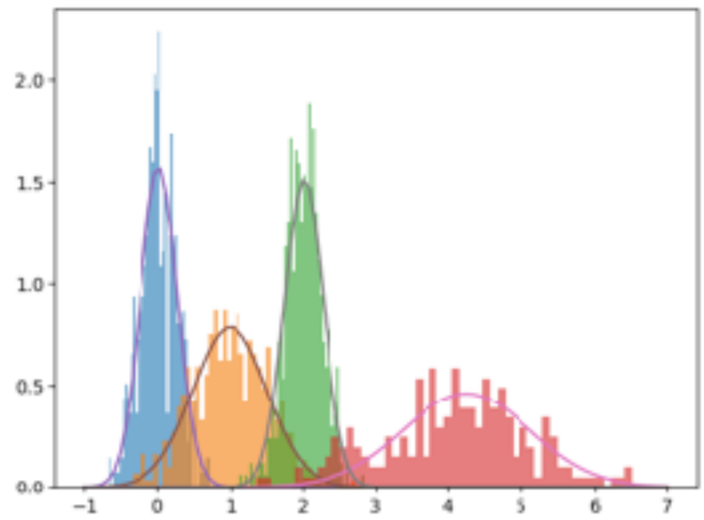
$$\phi_k = \sum_i^N \frac{\gamma_{i,k}}{N}$$

$$\mu_k = \frac{\sum_i^N \gamma_{i,k} x_i}{\sum_i^N \gamma_{i,k}}$$

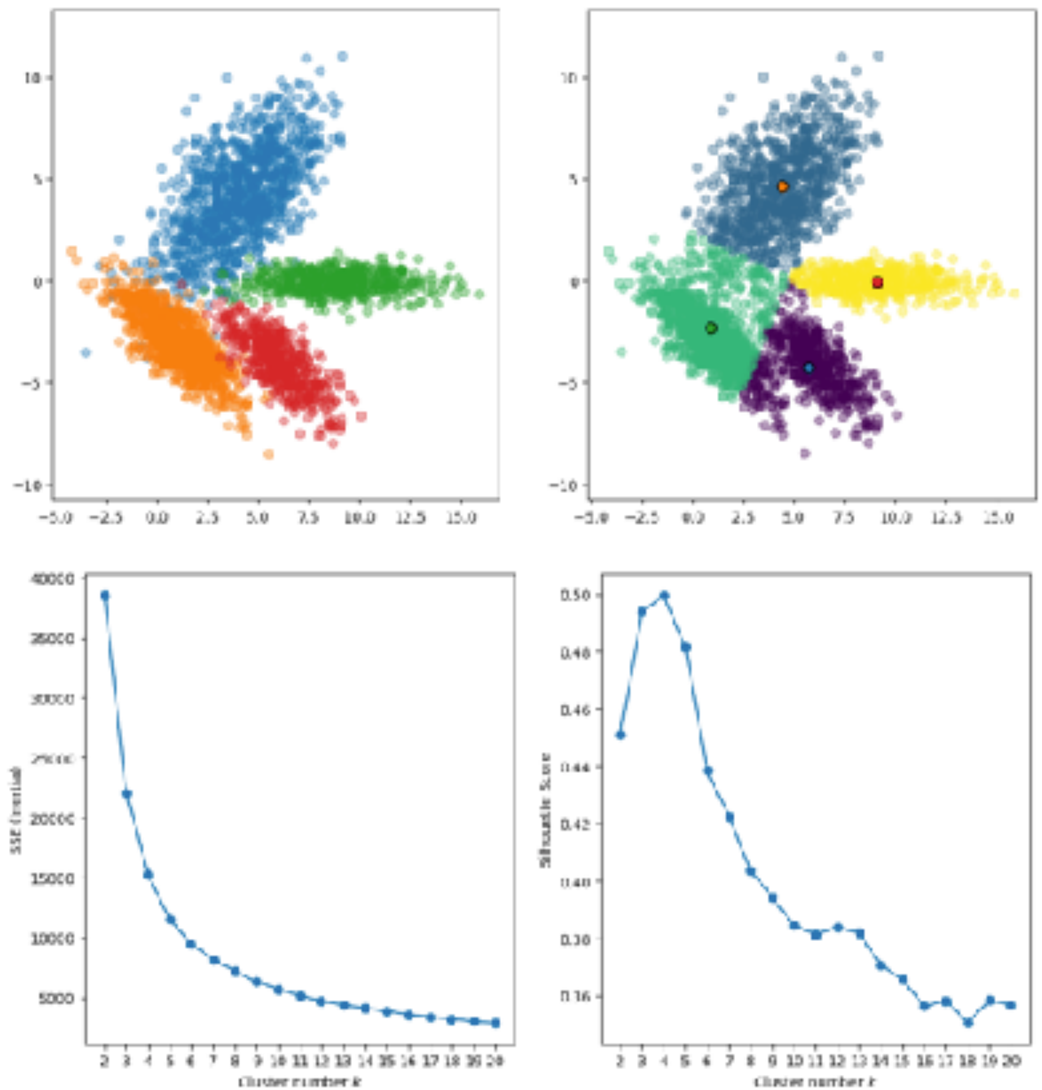
$$\Sigma_k = \frac{\sum_i^N \gamma_{i,k}(x_i - \mu_k)(x_i - \mu_k)}{\sum_i^N \gamma_{i,k}}$$

Gaussian Mixtures

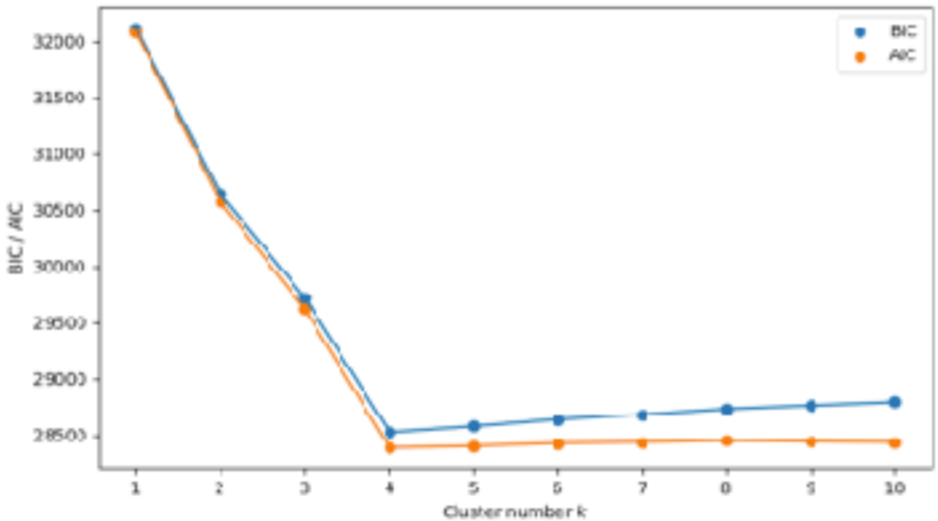


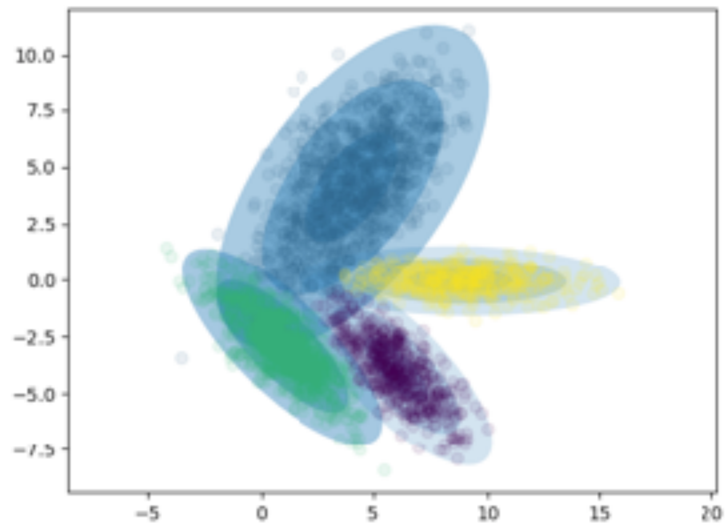


Value of *k* ?


SSE (Inertia)
Cluster number k
Silhouette Score
Cluster number k

Value of *k* ?





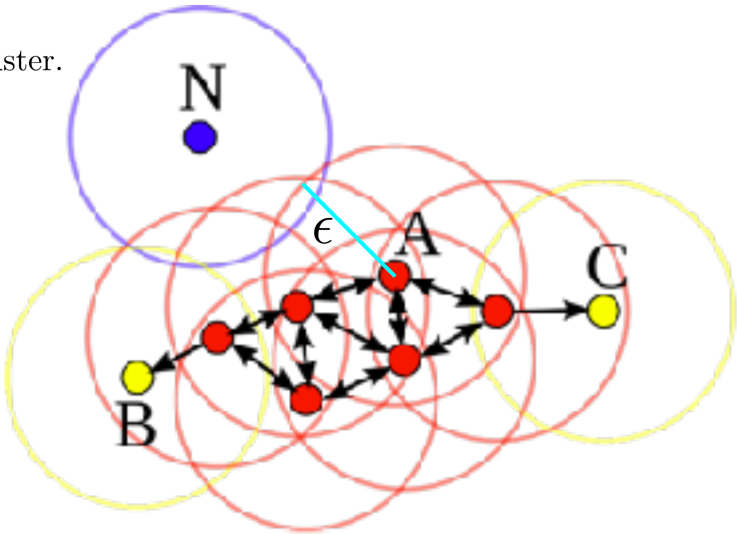
DBSCAN

- Density based algorithm
  - Groups points by considering the number and distances to nearest neighbors.

DBSCAN

- Density based algorithm
  - Groups points by considering the number and distances to nearest neighbors.

- Select an unvisited point at random.
- Identify neighbors within the distance ϵ.
- If a minimum number of data points (min_samples) are neighbors, a cluster is started.
- If a point is part of a cluster, its neighbors are also part of that cluster.
- Continue until the cluster is completed.



DBSCAN

- Density based algorithm
  - Groups points by considering the number and distances to nearest neighbors.
  - Resistant to noise: points in low density areas are outliers.

DBSCAN

- Density based algorithm
  - Groups points by considering the number and distances to nearest neighbors.
  - Resistant to noise: points in low density areas are outliers.
- Identifies clusters of different and non-symmetric shapes and sizes.

DBSCAN

- Density based algorithm
  - Groups points by considering the number and distances to nearest neighbors.
  - Resistant to noise: points in low density areas are outliers.
- Identifies clusters of different and non-symmetric shapes and sizes.
- Cannot cluster datasets with large differences in densities.
  - The distance and number of neighbors to start a cluster are not cluster specific, but apply to the entire dataset.

DBSCAN

- Consider the hyperparameters…
  - SciKit Documentation
  - eps
    - *The maximum **distance** between two samples for one to be considered as in the neighborhood of the other.*
    - *This is the most important DBSCAN parameter to choose appropriately for your data set and distance function.*
  - min_samples
    - *The number of samples in a neighborhood for a point to be considered as a core point. Includes the point itself.* (Minimum number of data points to be a cluster.)
    - *If set to a higher value, DBSCAN will find denser clusters*

DBSCAN

- Estimate the hyperparameters
  - min_samples
    - Set using SME and/or data familiarity
    - As the number of instances increases, so should min_samples
    - As the noise in the data increases, so should min_samples
    - Should be greater than or equal to the number of dimensions or features (f)
    - For f=2 data, min_samples = 4
    - For f > 2, min_samples = 2*f

DBSCAN

- Estimate the hyperparameters
  - eps

DBSCAN

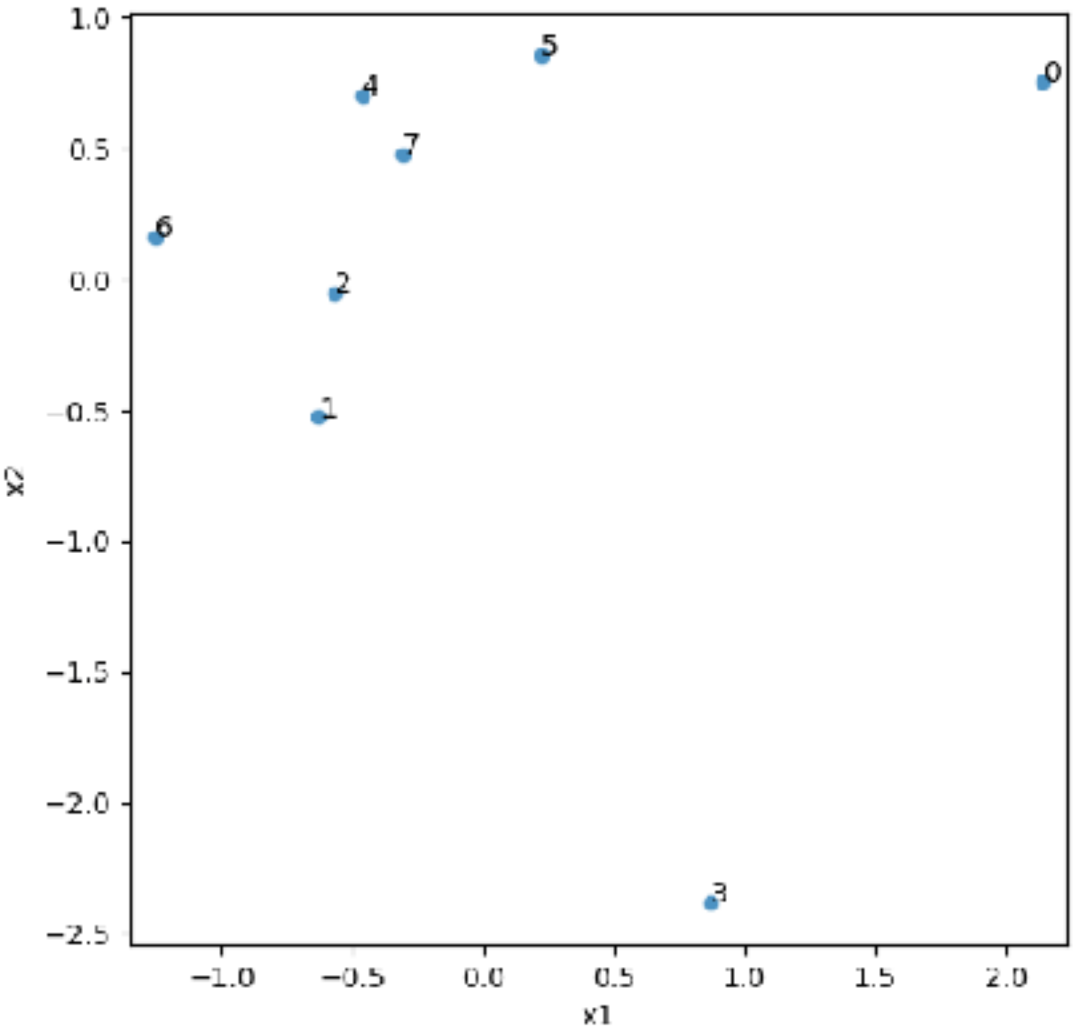
- Evaluation of clustering
  - Silhouette Analysis
    - Uses mean intra- and inter- cluster distances for each instance
  - Davies-Bouldin
    - Uses centroid locations to calculate distances

Agglomerative

- Assumes that all data points are initially their own cluster
- Successively merges pairs of clusters by considering their similarity or *linkage*:
  - *single*
    - Minimum distance between all instances of two clusters
  - *complete*
    - Maximum distance between all instances of two clusters
  - *average*
    - Average distance between all instances of two clusters
  - *ward*
    - Minimizes variances of the clusters being merged

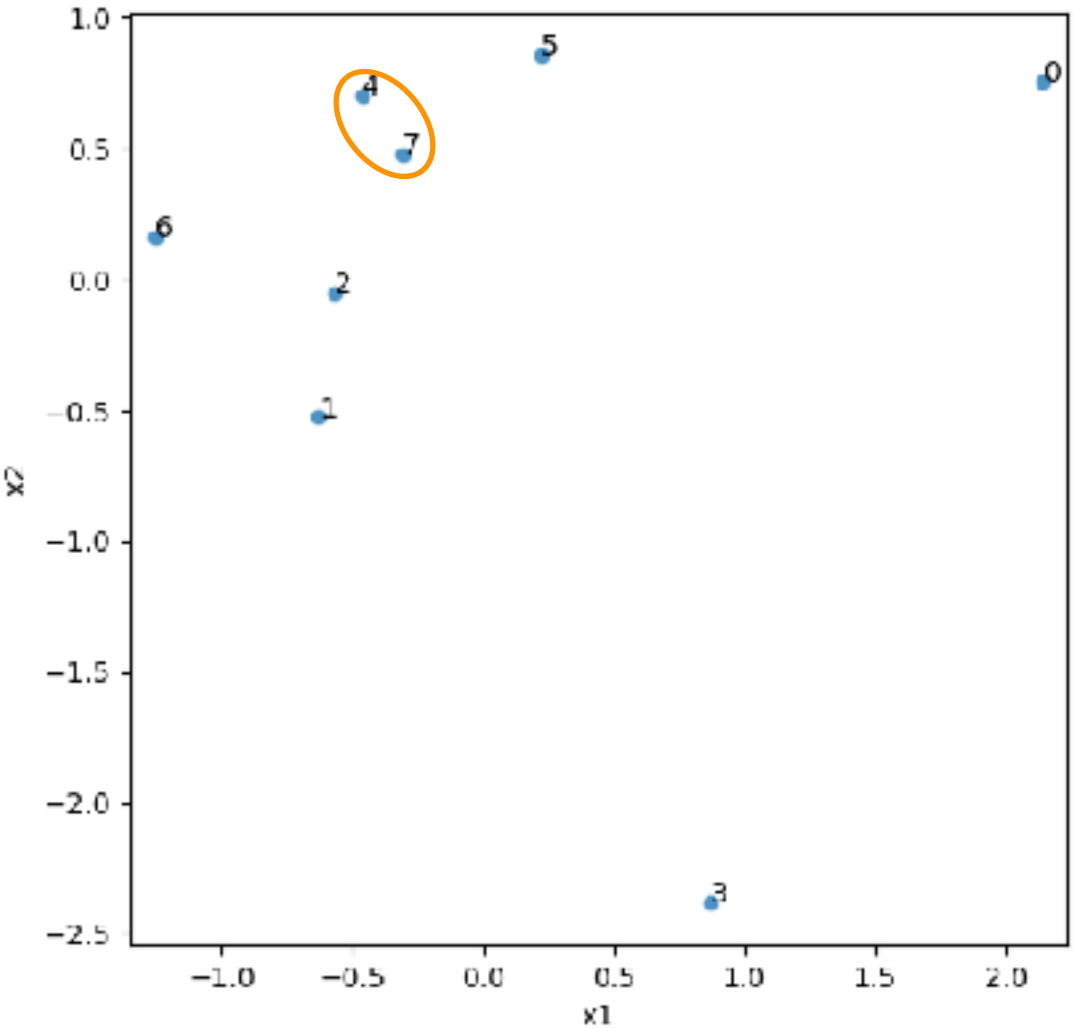
- SciKit Documentation

Agglomerative



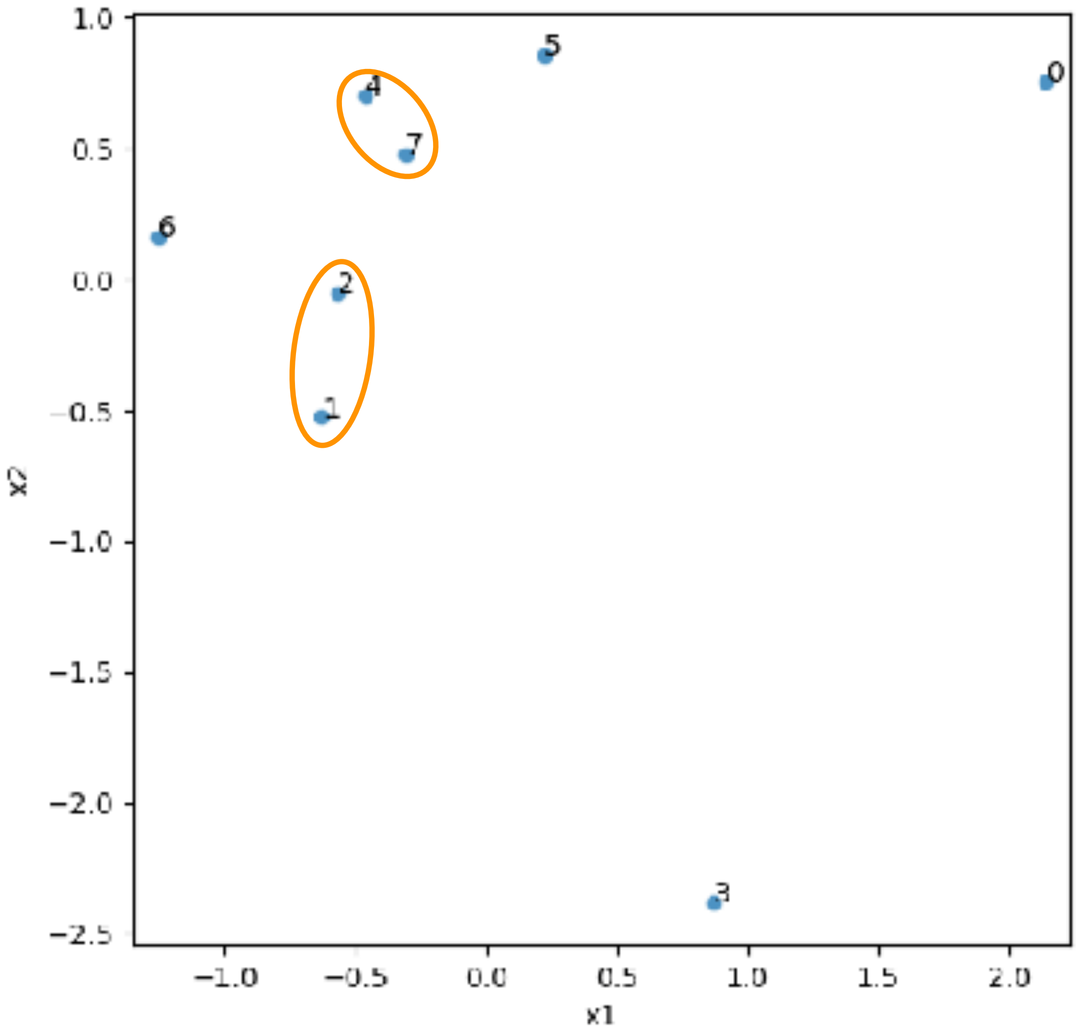
idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.591172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761675
1	6	0.930176
4	6	0.955851
6	7	0.998142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.627007
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833308
0	1	3.052951
3	7	3.090535
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

Agglomerative



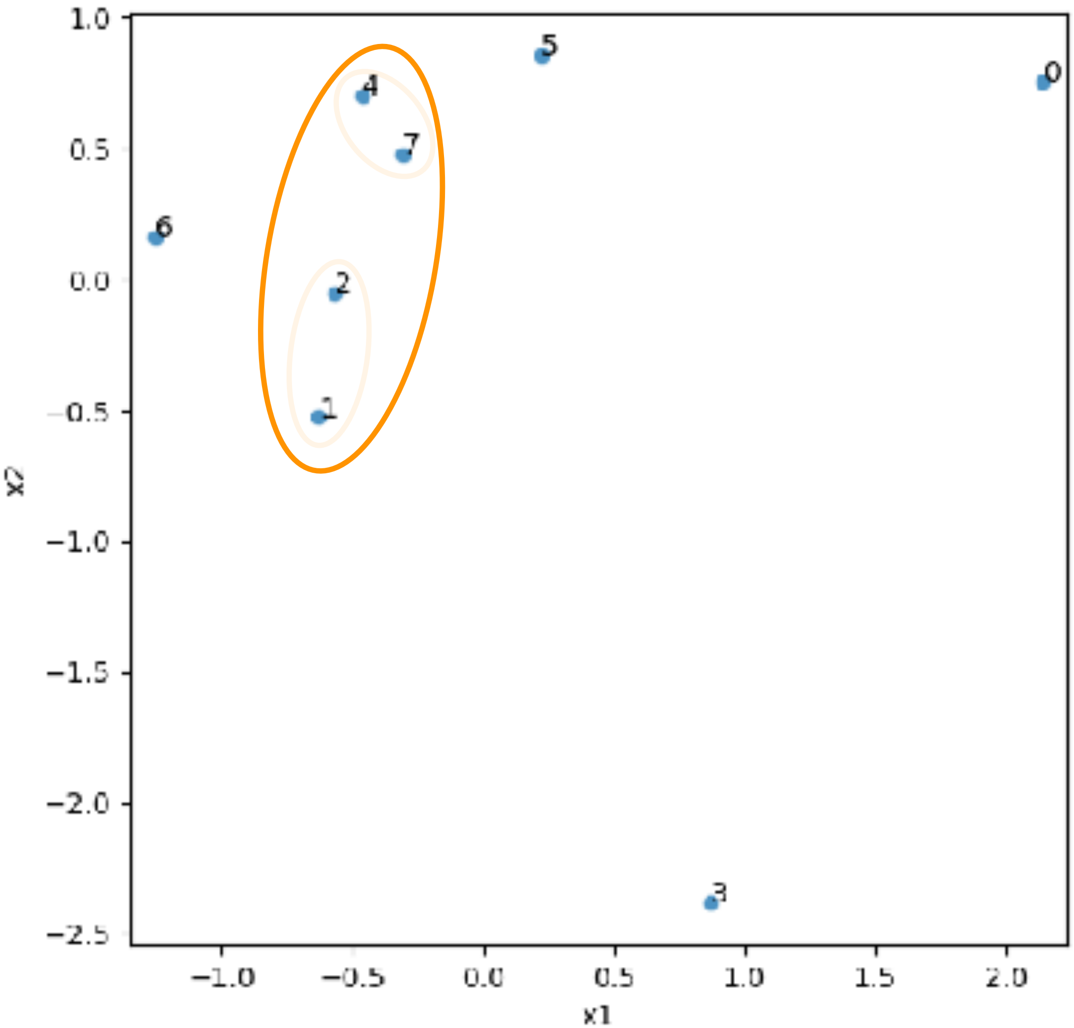
idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.592172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761675
1	6	0.930176
4	6	0.958851
6	7	0.998142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.627337
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833308
0	1	3.052951
3	7	3.090535
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

Agglomerative



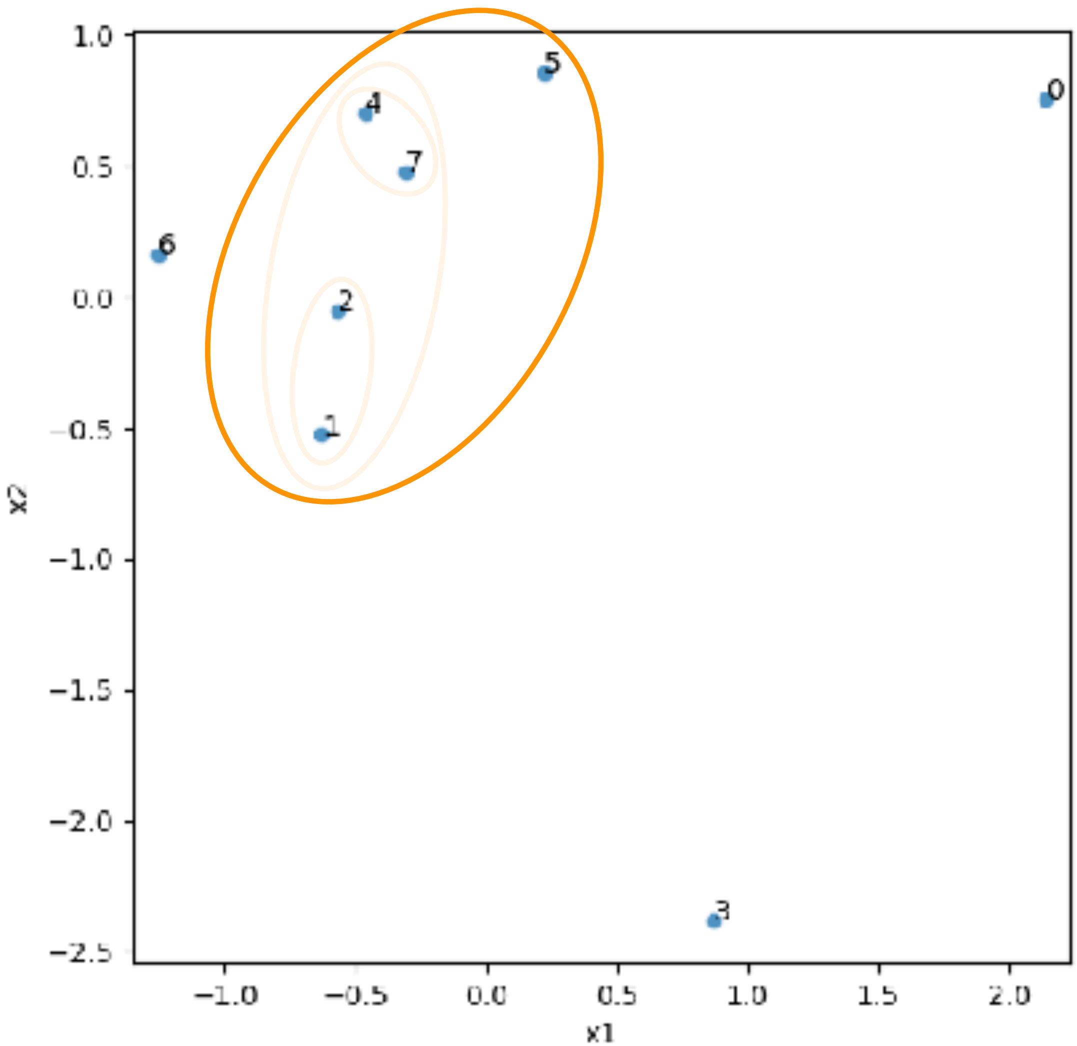
idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.592172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761675
1	6	0.930176
4	6	0.955851
6	7	0.996142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.627337
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833308
0	1	3.052951
3	7	3.090535
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

Agglomerative



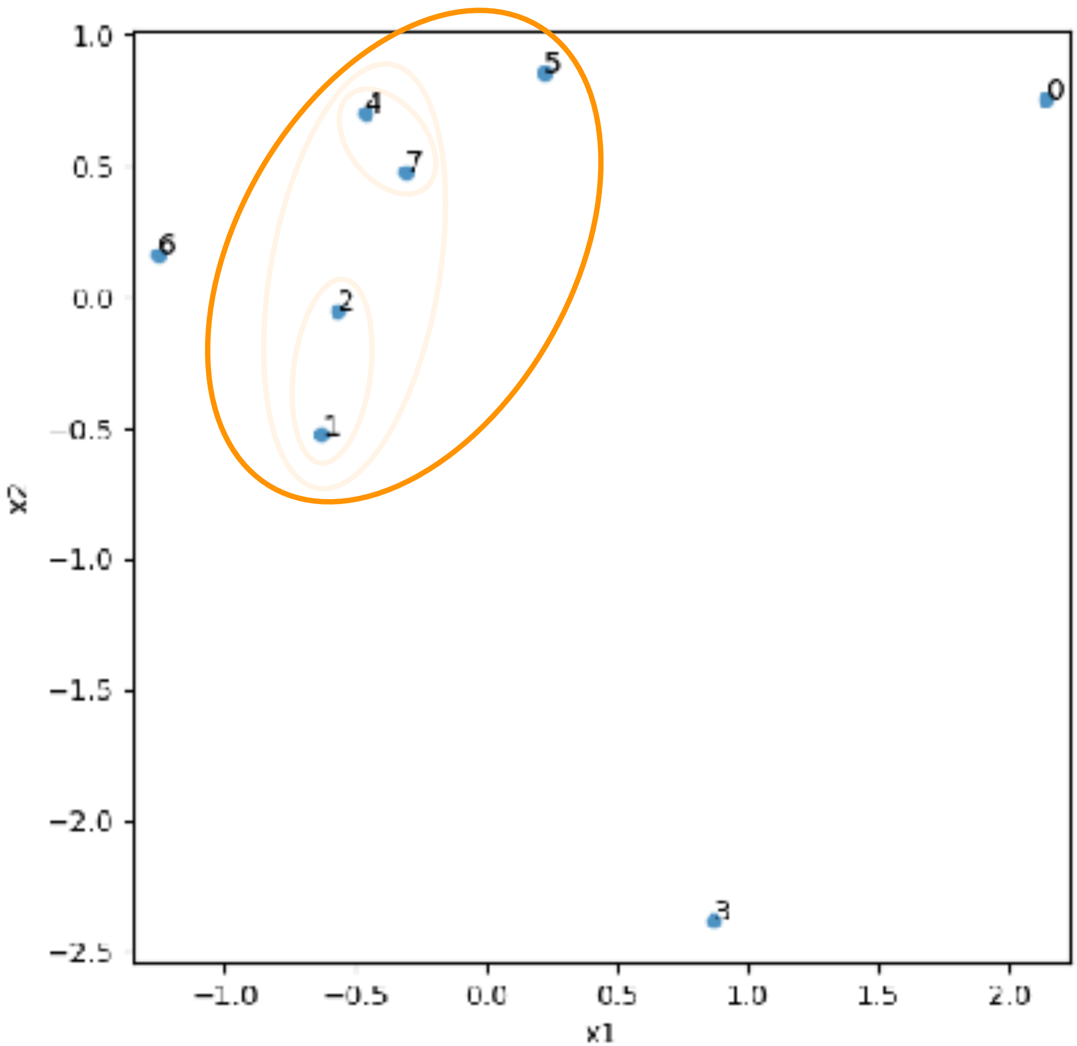
idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.591172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761675
1	6	0.930176
4	6	0.955851
6	7	0.998142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.627337
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833308
0	1	3.052951
3	7	3.090635
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

Agglomerative



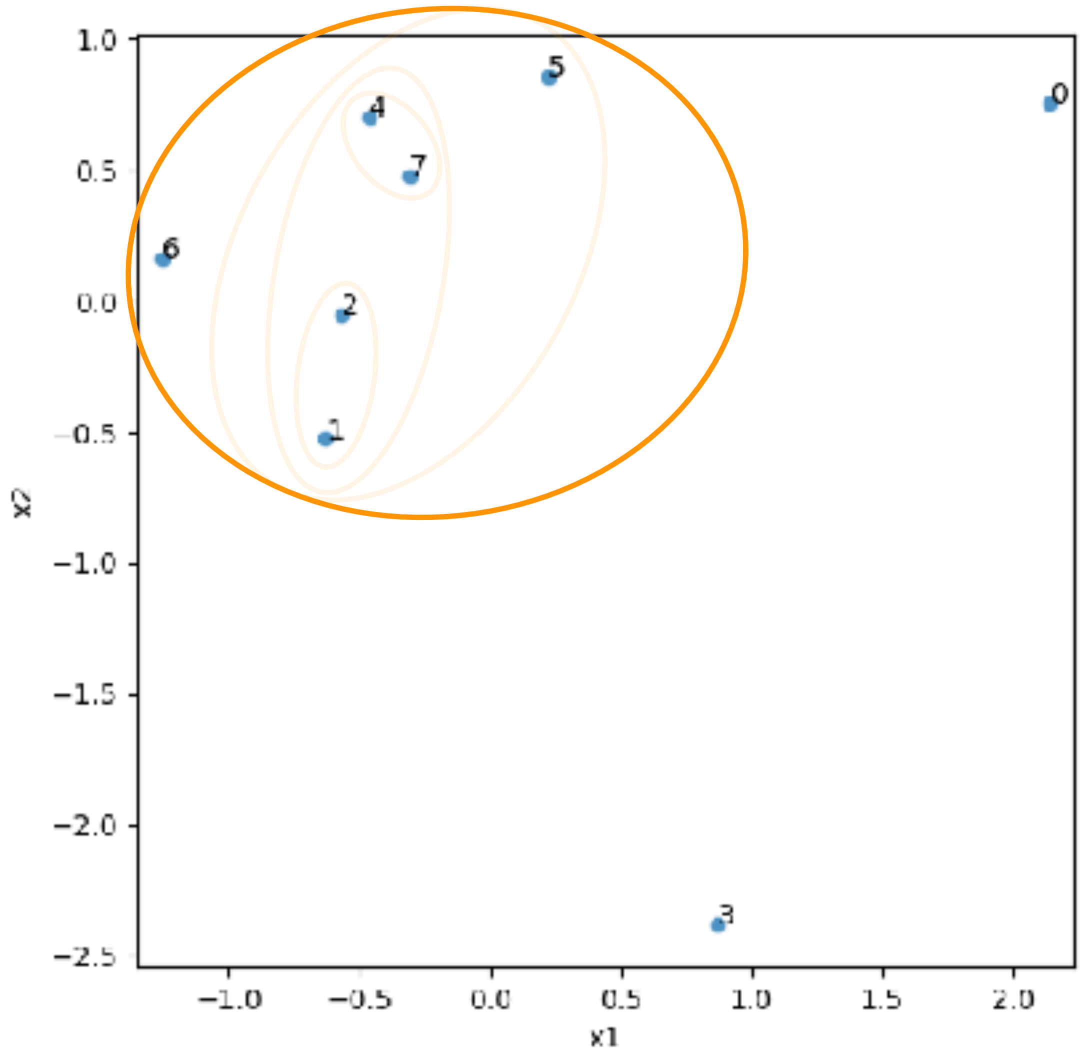
idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.592172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761676
1	6	0.930176
4	6	0.955851
6	7	0.996142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.627337
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833308
0	1	3.052951
3	7	3.090535
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

Agglomerative



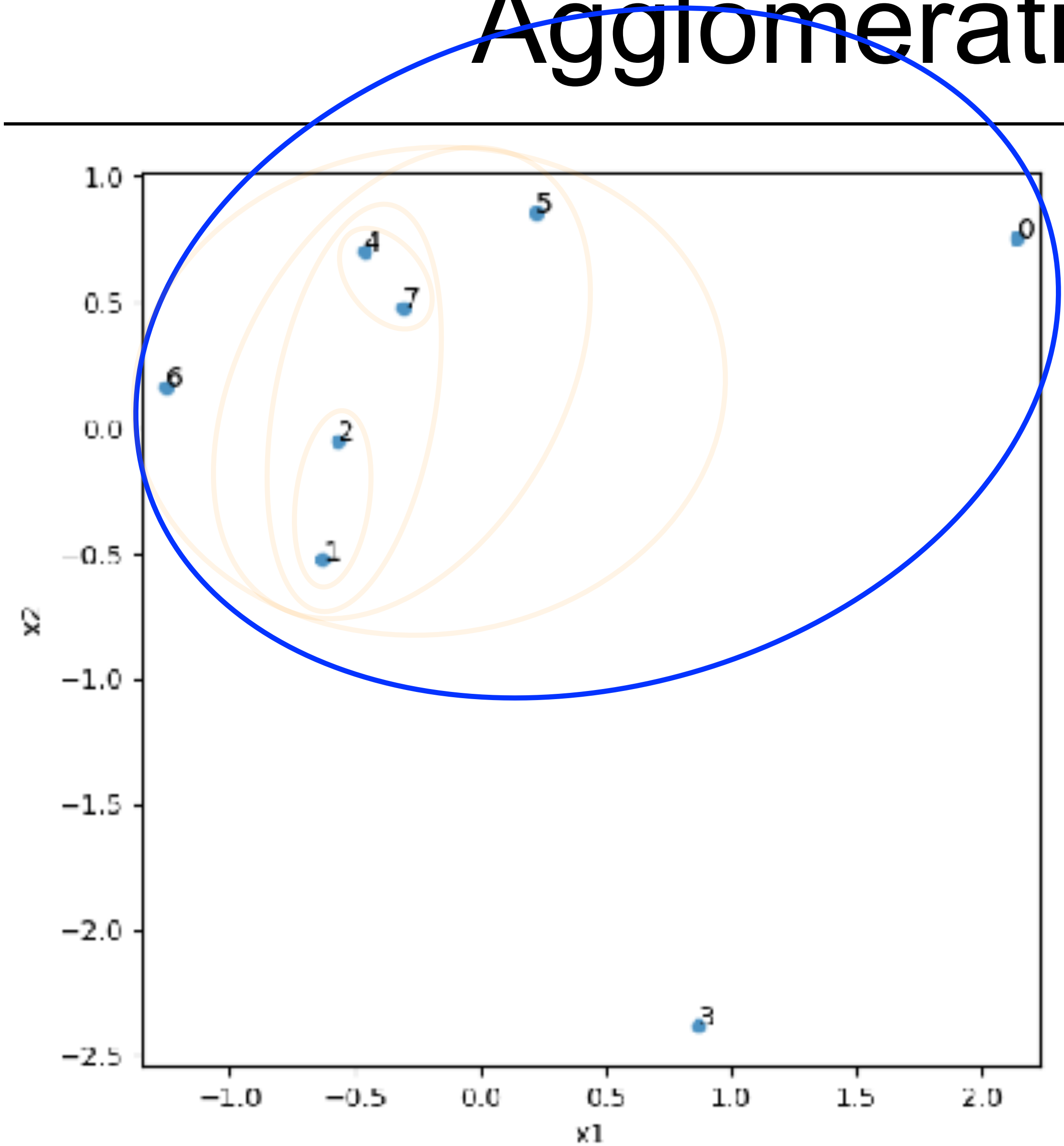
idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.592172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761675
1	6	0.930176
4	6	0.958851
6	7	0.998142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.627007
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833308
0	1	3.052951
3	7	3.090535
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

Agglomerative



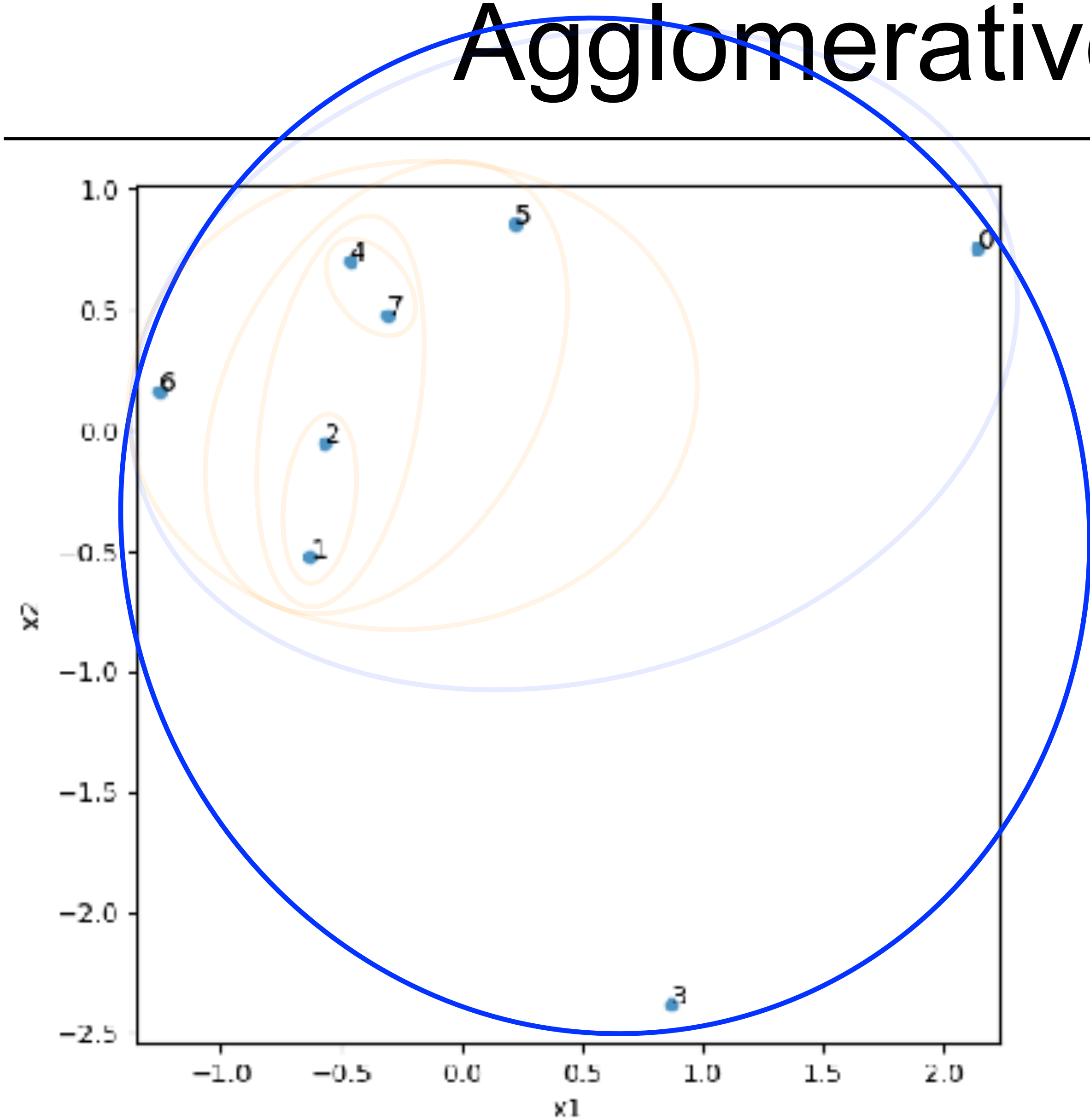
idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.592172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761675
1	6	0.930176
4	6	0.955851
6	7	0.998142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.427037
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833308
0	1	3.052951
3	7	3.090635
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

Agglomerative



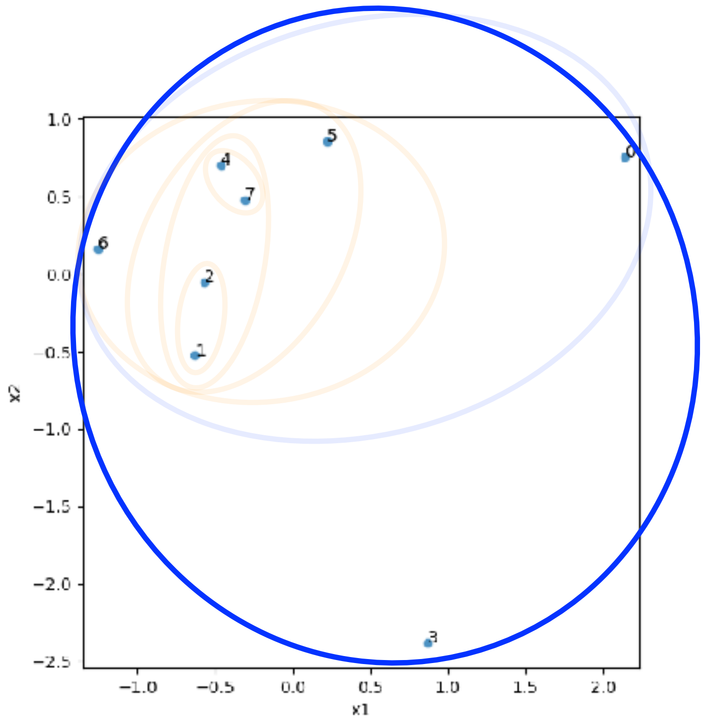
idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.591172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761675
1	6	0.930176
4	6	0.955851
6	7	0.998142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.627337
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833308
0	1	3.052951
3	7	3.090535
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

Agglomerative

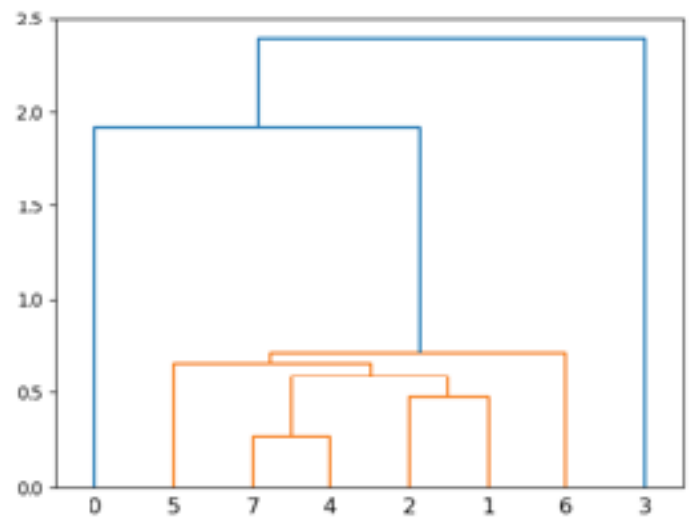


idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.592172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761675
1	6	0.930176
4	6	0.955851
6	7	0.998142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.027337
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833308
0	1	3.052951
3	7	3.090535
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

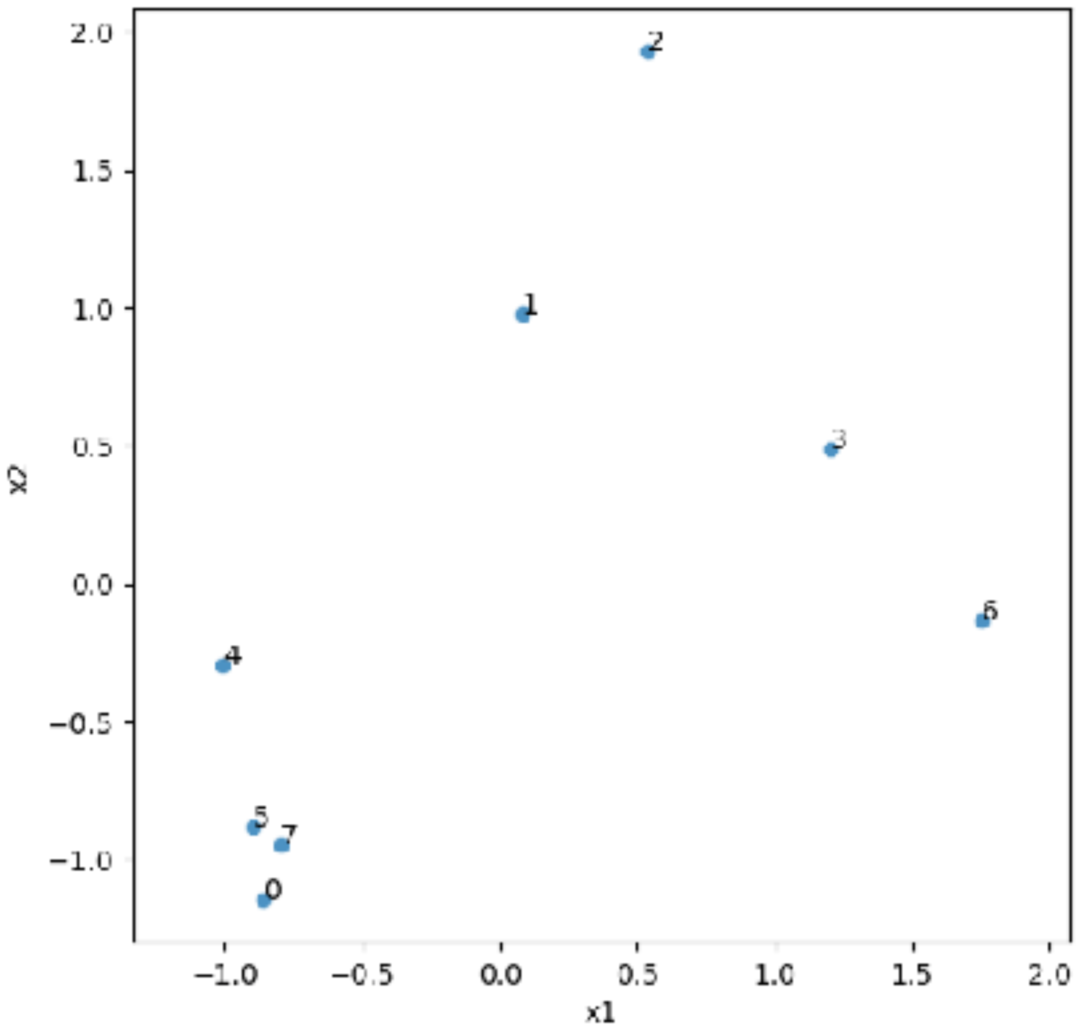
Agglomerative



idx1	idx2	dist
4	7	0.273128
1	2	0.477100
2	7	0.591172
5	7	0.657239
4	5	0.708838
2	6	0.712901
2	4	0.761675
1	6	0.930176
4	6	0.955851
6	7	0.998142
1	7	1.053129
2	5	1.212528
1	4	1.238645
1	5	1.627307
5	6	1.635960
0	5	1.916347
1	3	2.383826
0	7	2.465498
0	4	2.605734
2	3	2.739238
0	2	2.833306
0	1	3.052951
3	7	3.090635
3	5	3.301018
3	6	3.312018
3	4	3.357863
0	3	3.385676
0	6	3.446489

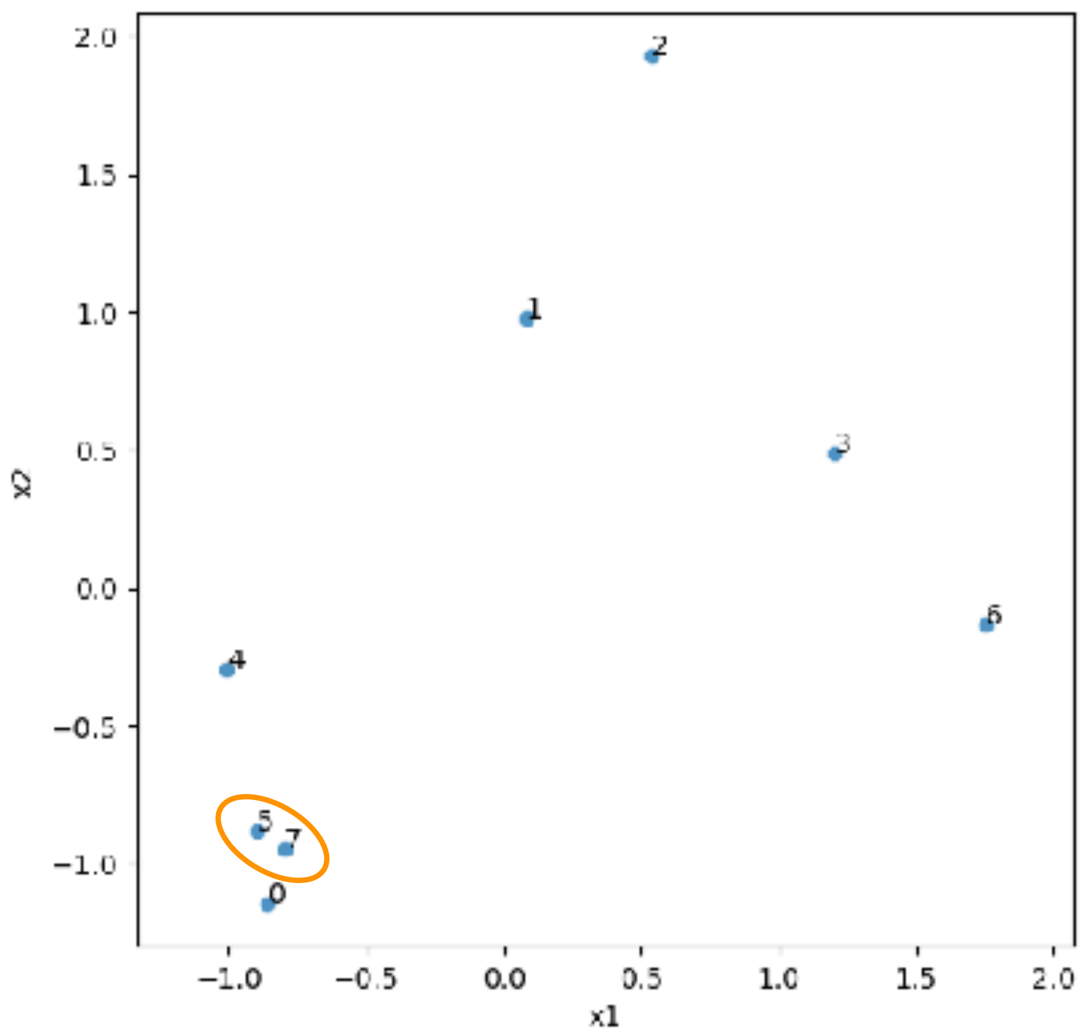


Agglomerative



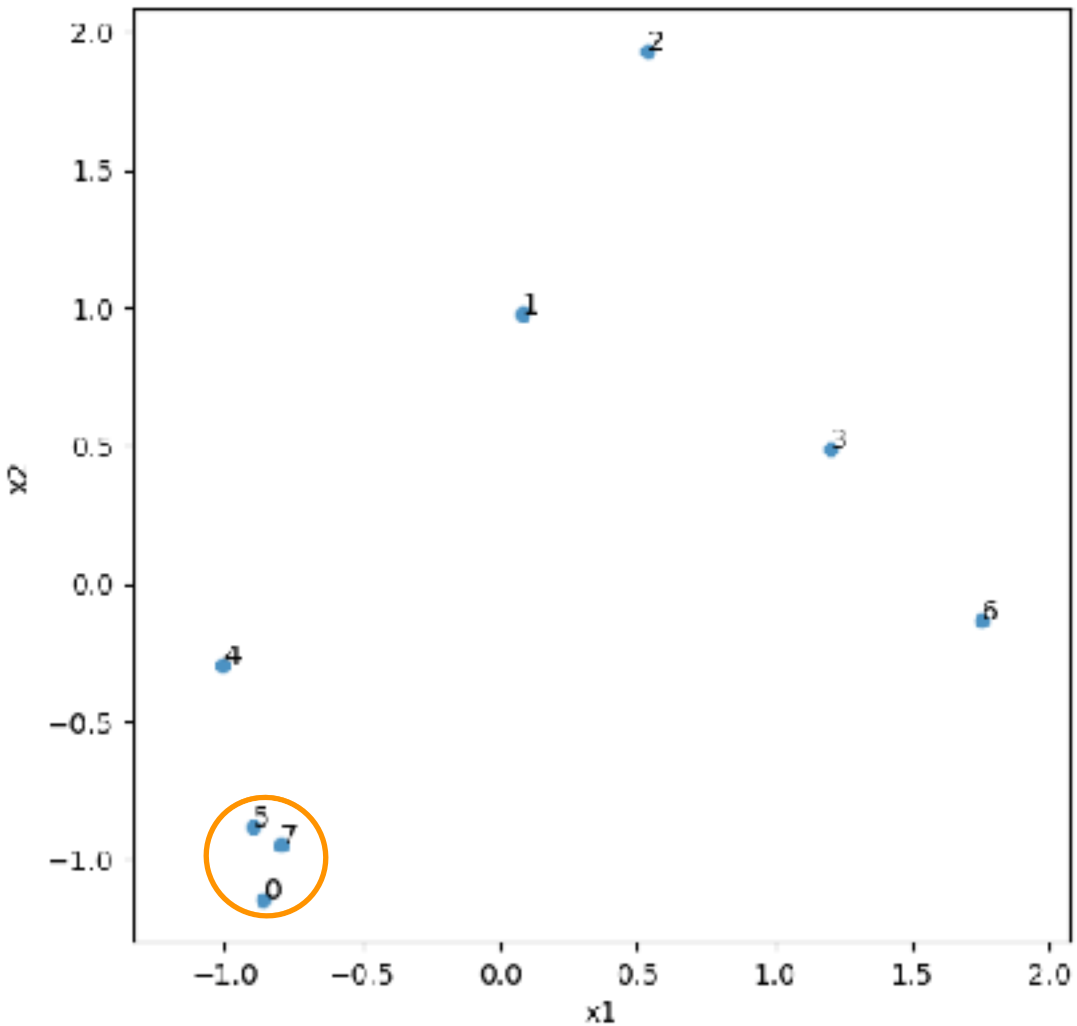
idx1	idx2	dist
5	7	0.120626
0	7	0.204960
0	5	0.266416
4	5	0.593721
4	7	0.685002
3	6	0.830519
0	4	0.859904
1	2	1.057046
1	3	1.226104
2	3	1.589177
1	4	1.673179
1	6	2.008112
1	5	2.055691
1	7	2.113592
0	1	2.317238
3	4	2.347471
2	6	2.357036
3	7	2.463425
3	5	2.506639
0	3	2.630727
6	7	2.634089
2	4	2.708019
5	6	2.749506
4	6	2.763791
0	6	2.797673
2	5	3.152207
2	7	3.170526
0	2	3.373003

Agglomerative



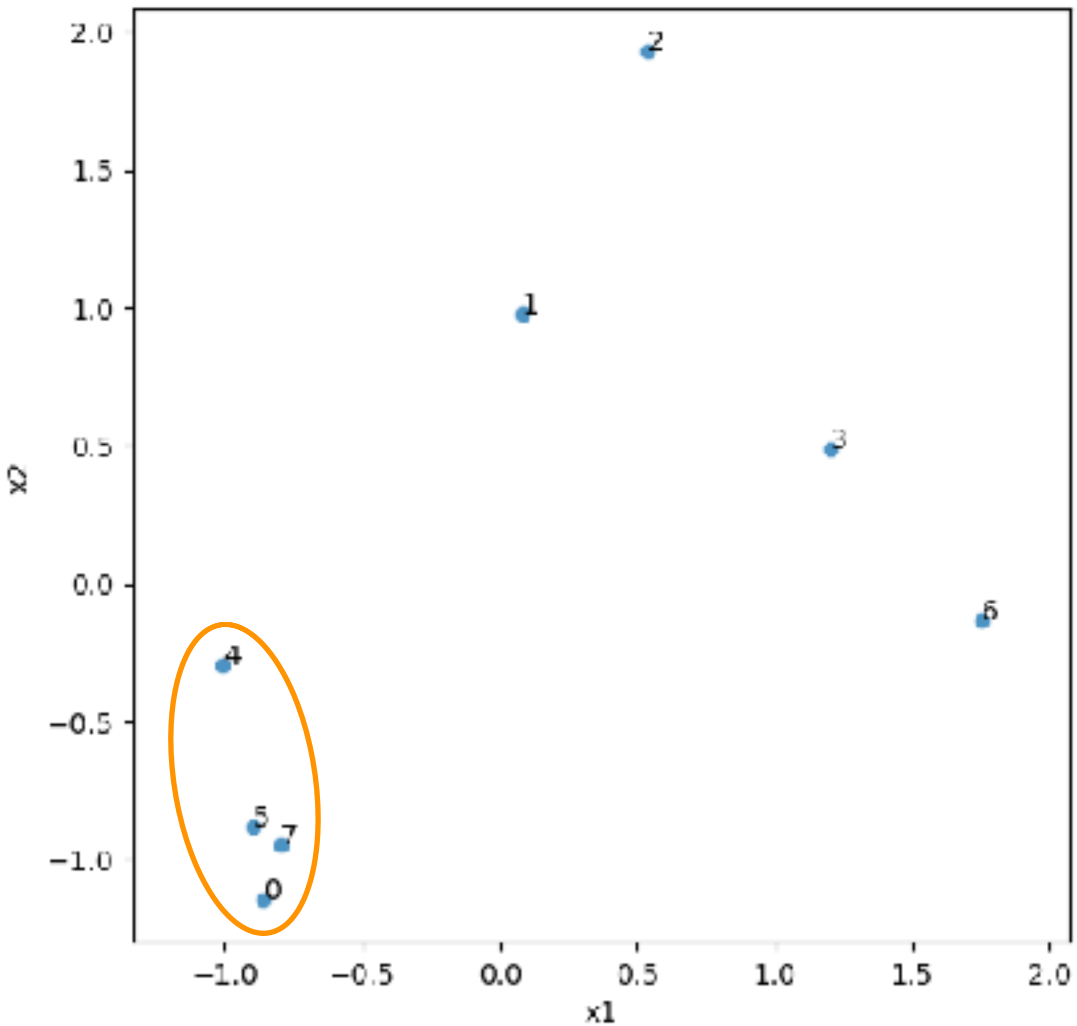
idx1	idx2	dist
5	7	0.120626
0	7	0.204960
0	5	0.266416
4	5	0.593721
4	7	0.685002
3	6	0.830519
0	4	0.859904
1	2	1.057046
1	3	1.226104
2	3	1.689177
1	4	1.673179
1	6	2.008112
1	5	2.055691
1	7	2.113592
0	1	2.317238
3	4	2.347471
2	6	2.357036
3	7	2.463425
3	5	2.506639
0	3	2.630727
6	7	2.634089
2	4	2.708019
5	6	2.749506
4	6	2.763791
0	6	2.797673
2	5	3.152207
2	7	3.170526
0	2	3.373003

Agglomerative



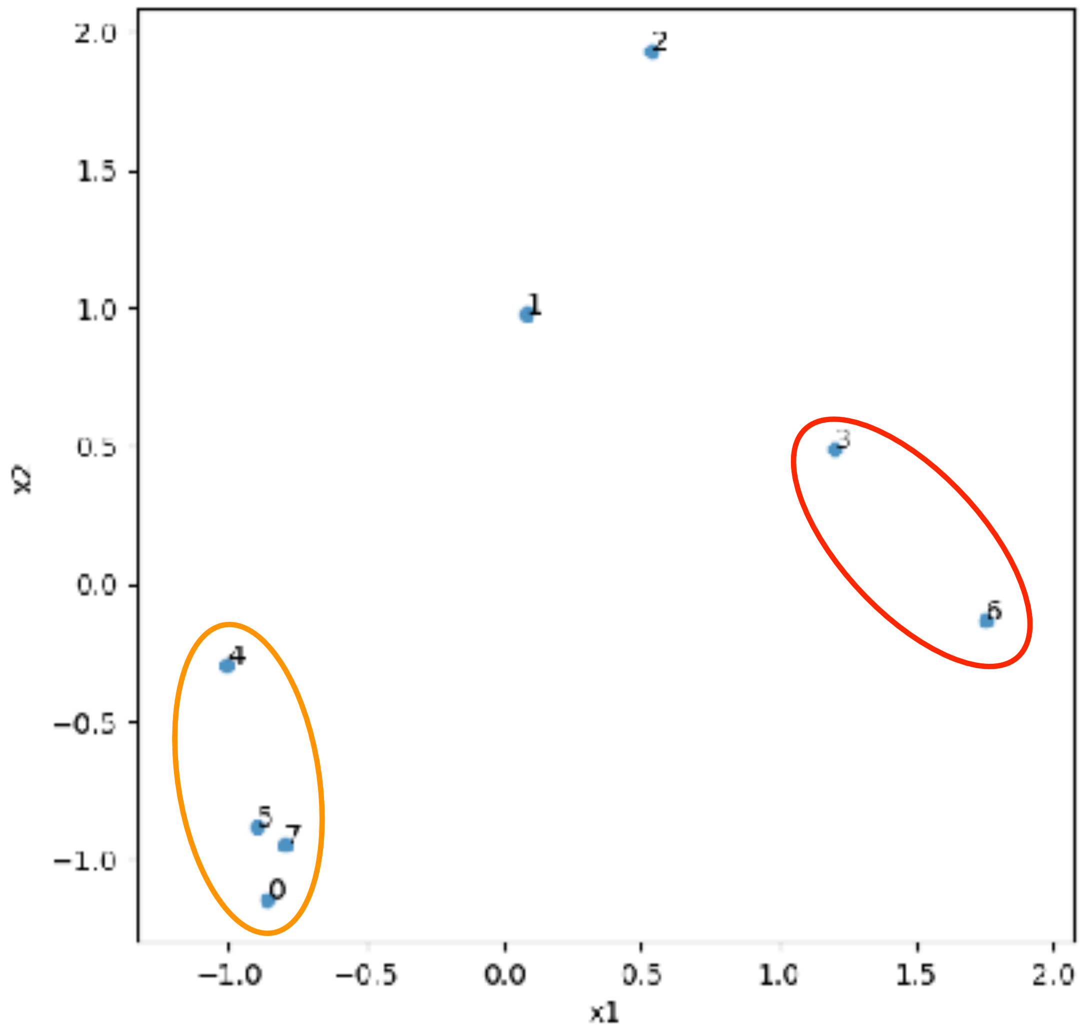
idx1	idx2	dist
5	7	0.120626
0	7	0.204960
0	5	0.266416
4	5	0.593721
4	7	0.685002
3	6	0.830519
0	4	0.859904
1	2	1.057046
1	3	1.226104
2	3	1.589177
1	4	1.673179
1	6	2.008112
1	5	2.055691
1	7	2.113592
0	1	2.317238
3	4	2.347471
2	6	2.357030
3	7	2.463425
3	5	2.506639
0	3	2.630727
6	7	2.634089
2	4	2.708019
5	6	2.749506
4	6	2.763791
0	6	2.797673
2	5	3.152207
2	7	3.170526
0	2	3.373003

Agglomerative



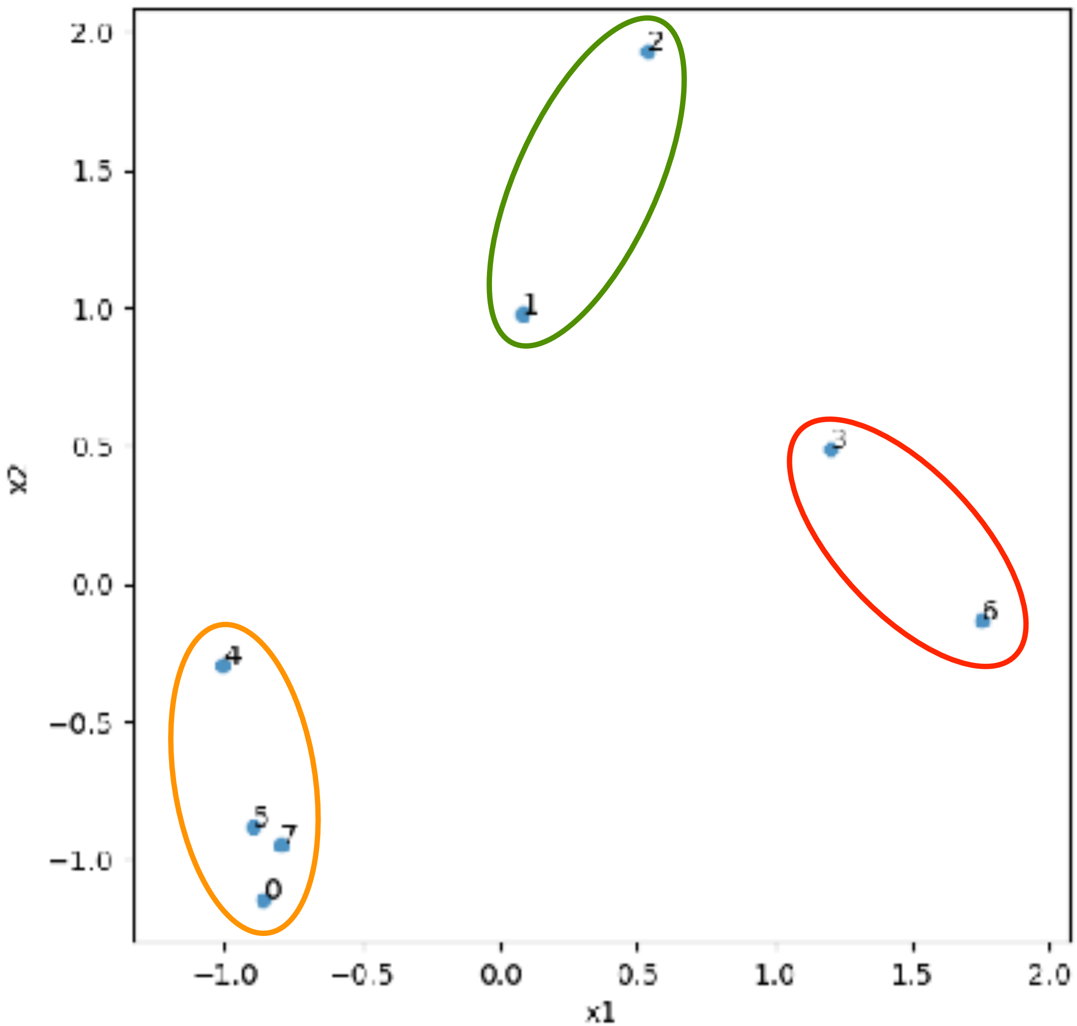
idx1	idx2	dist
5	7	0.120626
0	7	0.204960
0	5	0.266416
4	5	0.593721
4	7	0.685002
3	6	0.830519
0	4	0.859904
1	2	1.057046
1	3	1.226104
2	3	1.609177
1	4	1.673179
1	6	2.008112
1	5	2.055691
1	7	2.113592
0	1	2.317238
3	4	2.347471
2	6	2.357036
3	7	2.463425
3	5	2.506639
0	3	2.630727
6	7	2.634089
2	4	2.708019
5	6	2.749506
4	6	2.763791
0	6	2.797673
2	5	3.152207
2	7	3.170526
0	2	3.373003

Agglomerative



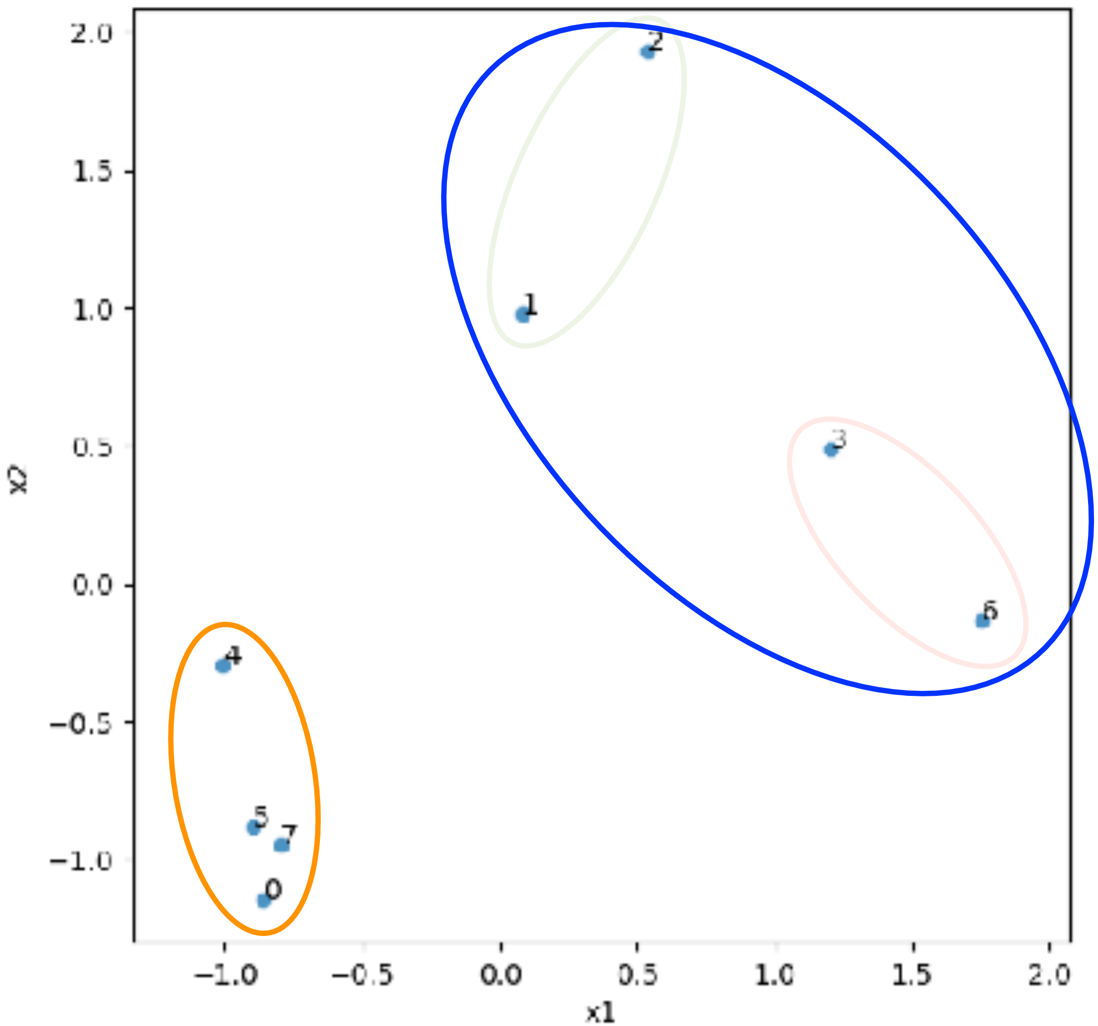
idx1	idx2	dist
5	7	0.120626
0	7	0.204960
0	5	0.266416
4	5	0.593721
4	7	0.685002
3	6	0.830519
0	4	0.859904
1	2	1.057046
1	3	1.226104
2	3	1.589177
1	4	1.673179
1	6	2.008112
1	5	2.055691
1	7	2.113592
0	1	2.317238
3	4	2.347471
2	6	2.357030
3	7	2.463425
3	5	2.506639
0	3	2.630727
6	7	2.634089
2	4	2.708019
5	6	2.749506
4	6	2.763791
0	6	2.797673
2	5	3.152207
2	7	3.170526
0	2	3.373003

Agglomerative



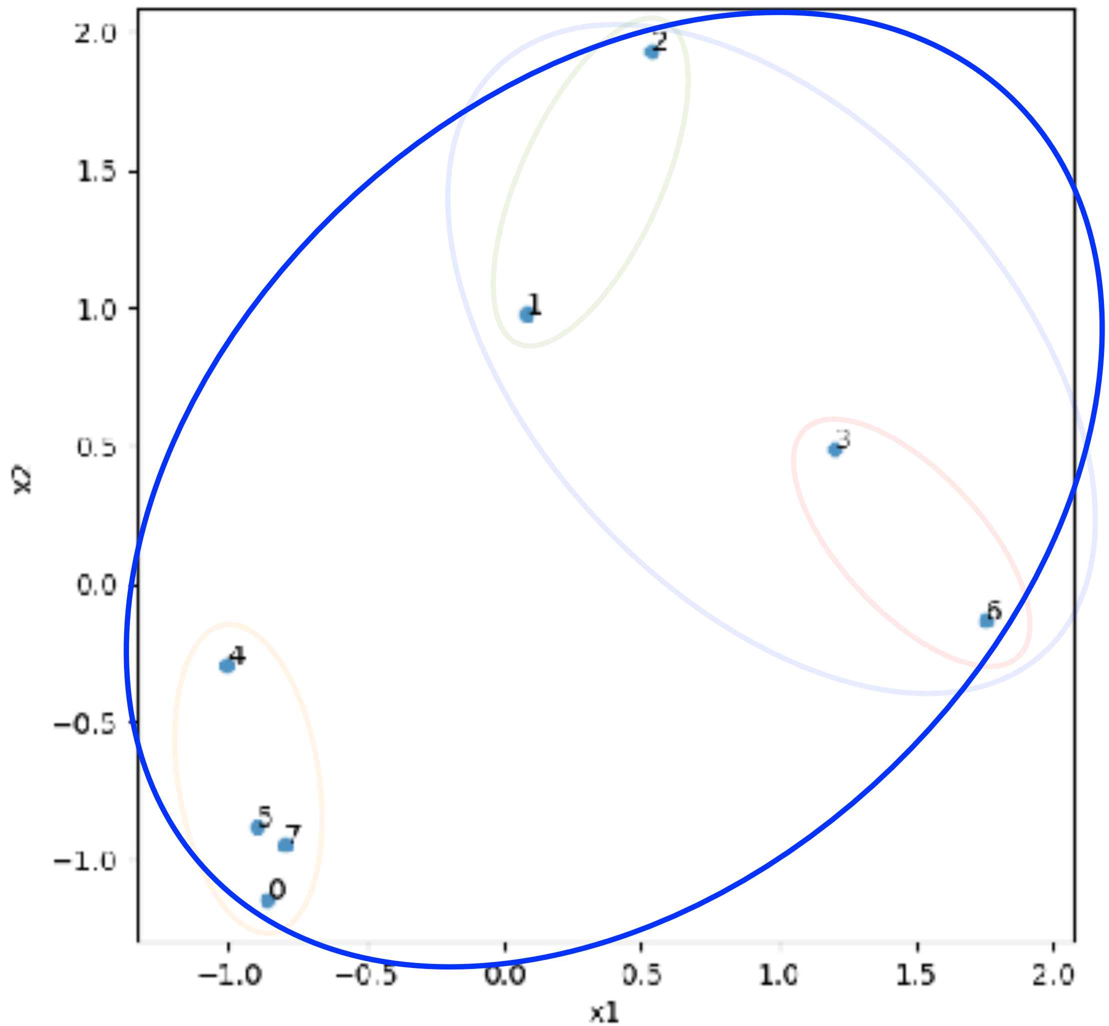
idx1	idx2	dist
5	7	0.120626
0	7	0.204960
0	5	0.266416
4	5	0.593721
4	7	0.685002
3	6	0.830519
0	4	0.859904
1	2	1.057046
1	3	1.226104
2	3	1.589177
1	4	1.673179
1	6	2.008112
1	5	2.055691
1	7	2.113592
0	1	2.317238
3	4	2.347471
2	6	2.357036
3	7	2.463425
3	5	2.506639
0	3	2.630727
6	7	2.634089
2	4	2.708019
5	6	2.749506
4	6	2.763791
0	6	2.797673
2	5	3.152207
2	7	3.170526
0	2	3.373003

Agglomerative



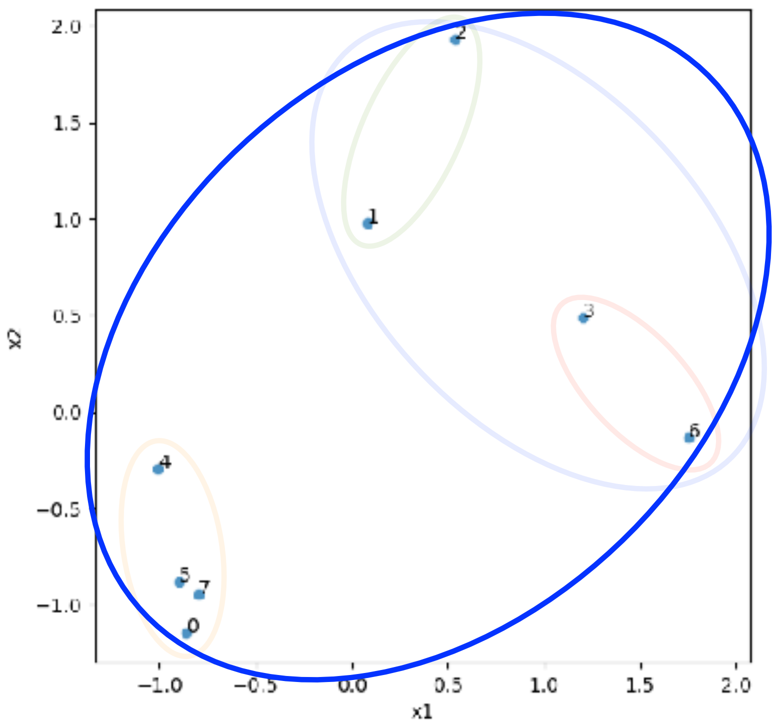
idx1	idx2	dist
5	7	0.120626
0	7	0.204960
0	5	0.266416
4	5	0.593721
4	7	0.685002
3	6	0.830619
0	4	0.859904
1	2	1.057046
1	3	1.226104
2	3	1.589177
1	4	1.673179
1	6	2.008112
1	5	2.055691
1	7	2.113592
0	1	2.317238
3	4	2.347471
2	6	2.357036
3	7	2.463425
3	5	2.506639
0	3	2.630727
6	7	2.634089
2	4	2.708019
5	6	2.749506
4	6	2.763791
0	6	2.797673
2	5	3.152207
2	7	3.170526
0	2	3.373003

Agglomerative



idx1	idx2	dist
5	7	0.120626
0	7	0.204960
0	5	0.266416
4	5	0.593721
4	7	0.685002
3	6	0.830619
0	4	0.859904
1	2	1.057046
1	3	1.226104
2	3	1.589177
1	4	1.673179
1	6	2.008112
1	5	2.055691
1	7	2.113592
0	1	2.317238
3	4	2.347471
2	6	2.357036
3	7	2.463425
3	5	2.506639
0	3	2.630727
6	7	2.634089
2	4	2.708019
5	6	2.749506
4	6	2.763791
0	6	2.797673
2	5	3.152207
2	7	3.170526
0	2	3.373003

Agglomerative



idx1	idx2	dist
5	7	0.120626
0	7	0.204960
0	5	0.266416
4	5	0.593721
4	7	0.685002
3	6	0.830519
0	4	0.859904
1	2	1.057046
1	3	1.226104
2	3	1.689177
1	4	1.673179
1	6	2.068112
1	5	2.095691
1	7	2.113592
0	1	2.317238
3	4	2.347471
2	6	2.367036
3	7	2.463425
3	5	2.506639
0	3	2.630727
6	7	2.634089
2	4	2.708019
5	6	2.749506
4	6	2.763791
0	6	2.797673
2	5	3.152207
2	7	3.170526
0	2	3.373003