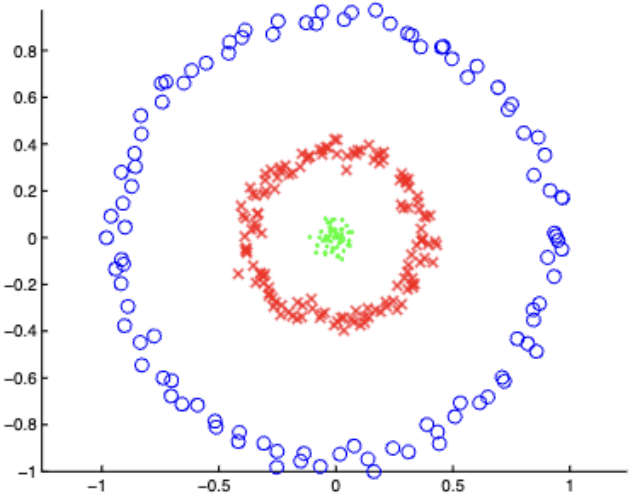
Lesson 12

Final Project, Python

guidelines

- Submission Date: 24/11/2024
- NO EXTENSION!!!
- Automatic tests, manual test

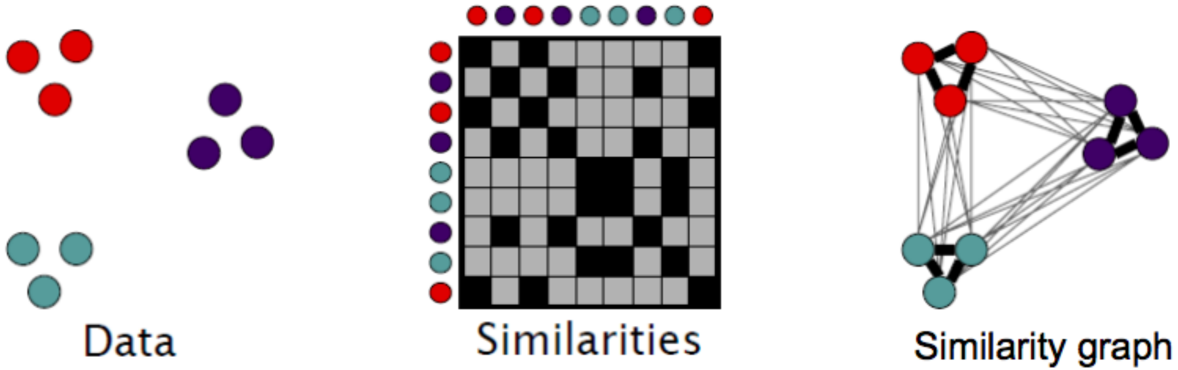
Motivation - Non Linear Data



[Kuang et al.(2012), Simon et al.(2005)]

From Data points to Similarity

- Given: d-dimensional points: $x_1, x_2, \ldots, x_n$.
- Transform them into a graph (Similarity Graph).
  - $G = (V, E; W)$ - undirected with no self loops.



Weighted Adjacency Matrix

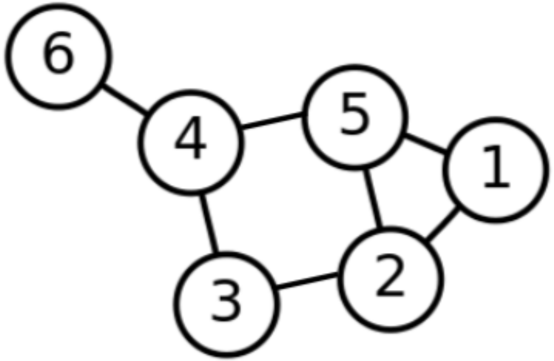
- Gaussian RBF

$$a_{ij} = \exp(\frac{-\|x_i - x_j\|^2}{2})$$

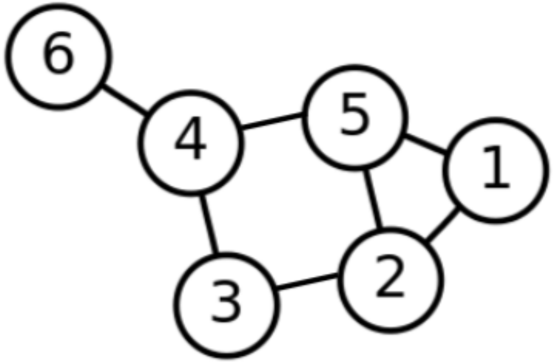
- A is symmetric, non-negative and no self loops $a_{ii} = 0$

Example

- For simplicity, in the next example we will show a non fully connected graph, with all weights set to 1
- We are given d-dimensional data points: x_1 … x_n
- Choose random points and connect them, and we get:



notations



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$W = \begin{bmatrix} 0 & \frac{\sqrt{6}}{6} & 0 & 0 & \frac{\sqrt{6}}{6} & 0 \\ \frac{\sqrt{6}}{6} & 0 & \frac{\sqrt{6}}{6} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{\sqrt{6}}{6} & 0 & \frac{\sqrt{6}}{6} & 0 & 0 \\ 0 & 0 & \frac{\sqrt{6}}{6} & 0 & \frac{1}{3} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{6}}{6} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{3}}{3} & 0 & 0 \end{bmatrix}$$

$$d_i = \sum_{j=1}^{n} a_{ij}$$

$$W = D^{-1/2} A D^{-1/2}$$

SymNMF Clustering

- Form the similarity matrix A from X
- Compute the Diagonal Degree Matrix
- Compute the normalized similarity W
- Factorize:

$$\text{Find } H_{n\times k} \text{ that solves: } \min_{H\geq 0} \|W - HH^T\|_F^2$$

Matrix factorization

Objective:

$$\min_{H \geq 0} \|W - HH^T\|_F^2$$

$H_{n \times k},\ k << n,$

Matrix factorization

- Initialize H randomly, from interval [0,2*sqrt(m/k)], where m is average of W.

- Update H

$$H_{ij}^{(t+1)} \leftarrow H_{ij}^{(t)} \left(1 - \beta + \beta \frac{(WH^{(t)})_{ij}}{(H^{(t)}(H^{(t)})^T H^{(t)})_{ij}} \right)$$

where $\beta = 0.5$

- Convergence

$$\|H^{(t+1)} - H^{(t)}\|_F^2 < \epsilon.$$

Assign clusters

Hard clustering, we choose for each element the cluster with the highest association score.

$$H = \begin{bmatrix} \boxed{0.06} & 0.01 \\ 0.01 & \boxed{0.05} \\ 0.01 & \boxed{0.04} \\ 0.02 & \boxed{0.04} \\ \boxed{0.05} & 0.02 \end{bmatrix}$$