THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

AMA 2233

Data Analytics and Visualization

Yancheng Yuan

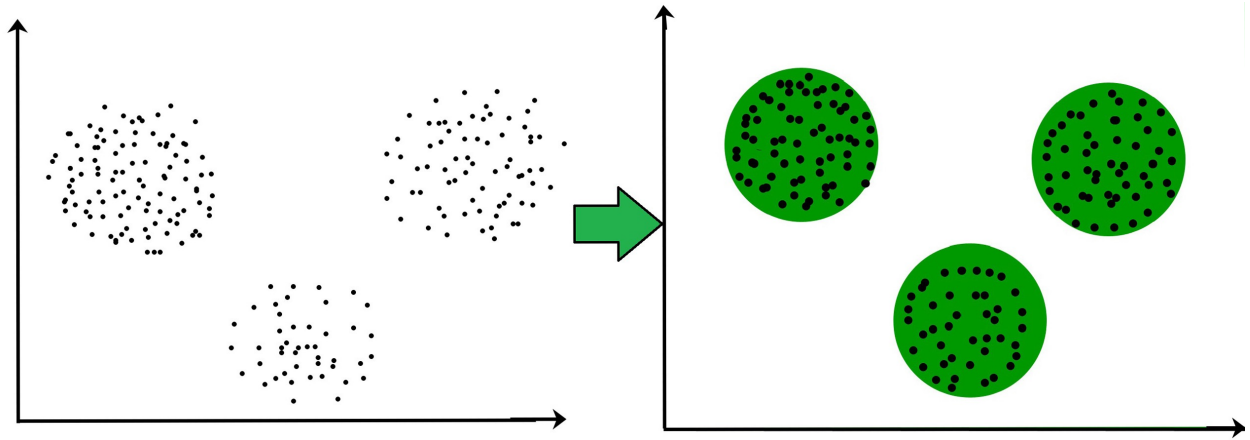
Department of Data Science and Artificial Intelligence

Opening Minds • Shaping the Future
啟迪思維 • 成就未來

Recap of last lecture

Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more **similar** (in some sense) to each other than to those in **other groups** (clusters).



Clustering in Euclidean Space

Clustering in Euclidean space: Giving n data points $\{x_1, x_2, \dots, x_n\} \subseteq R^d$, our target is to partite the data points into K clusters based on reasonable similarity measure.

K-means model: Giving n data points $\{x_1, x_2, \dots, x_n\} \subseteq R^d$, K-means model solves the following

$$\min_{S_1 \cup \cdots \cup S_K} \sum_{l=1}^{K} \sum_{i \in S_l} ||x_i - \frac{1}{|S_l|} \sum_{j \in S_l} x_j||^2$$

How to find the solution? Lloyd's algorithm

Lioyd's algorithm

Input: n data points $\mathrm{X} = \{x_1, x_2, \dots, x_n\} \subseteq R^d$, number of clusters K.

Initialization: Randomly initialize K centroid $\mu_1^0, \dots, \mu_K^0$.

for $\boldsymbol{t = 0, 1, 2, \cdots}$

Step 1: Update Cluster Assignment:

$$S_l^t = \{i \in [n] \mid ||x_i - \mu_l^t||^2 \leq ||x_i - \mu_j^t||^2, \forall\, j, 1 \leq j \leq K\}.$$

Step 2: Update Centroids:

$$\mu_l^{t+1} = \frac{1}{|S_l^t|} \sum_{i \in S_l^t} x_i$$

K-Means++: Run K-means with a better initialization

K-means++ algorithm

Input: n data points $\mathrm{X} = \{x_1, x_2, \dots, x_n\} \subseteq R^d$, number of clusters K.

Initialization of centroids:

Step 1: Choose an initial centroid $\mathbf{C} = \{\boldsymbol{\mu}_\mathbf{1}^\mathbf{0}\}$ uniformly at random from **X**.

Step 2: Choose the next centroid $\boldsymbol{\mu}_\mathbf{i}^\mathbf{0} \in \boldsymbol{X}$ with probability

$$P(x_i) = \frac{D_C(x_i)^2}{\sum_{x_j \in X} D_C\left(x_j\right)^2}$$

and set $\mathbf{C} = \mathbf{C} \cup \{\boldsymbol{\mu}_\mathbf{1}^\mathbf{0}\}$.

Step 3: Repeat **Step 2 until** $|\mathbf{C}| = \mathbf{K}$.

Run K-means with the initialized centroids $\mathbf{C}$.

Agenda

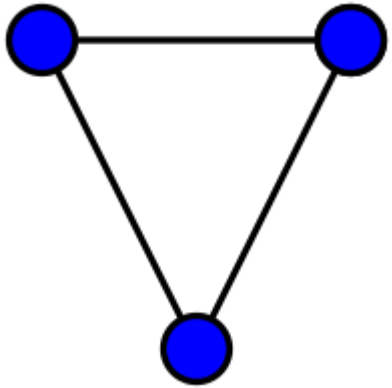


1. Graph
2. Page Rank Algorithm
3. Spectral Clustering
4. Applications

Graph

Graph

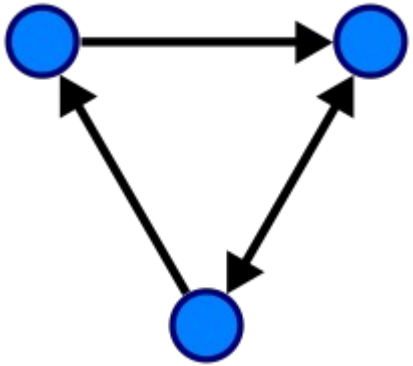
Graph: A graph (sometimes called an undirected graph to distinguish it from a directed graph) is a pair G = (V, E), where V is a set whose elements are called vertices, and E is a set of paired vertices, whose elements are called edges (sometimes links or lines).



Directed Graph

Directed graph: A directed graph or digraph is a graph in which edges have orientations. A directed graph is a pair G = (V, E) comprising:

- **V**, a set of vertices (also called nodes or points);
- **E**, a set of edges (also called directed edges, directed links, directed lines, arrows, or arcs), which are **ordered pair** of distinct vertices: $E = \{(x, y) | x, y \in V, x \neq y\}$.



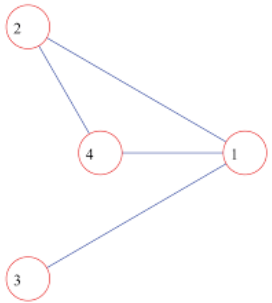
Adjacency matrix of an undirected graph

adjacency matrix is a square matrix used to represent a finite graph. The elements of the matrix indicate whether pairs of vertices are adjacent or not in the graph. For a graph G = (V, E) with n vertices, the adjacency matrix is

$$A_{ij} = \begin{cases} 1 \; if \; (i,j) \in E \\ 0, otherwise \end{cases}$$

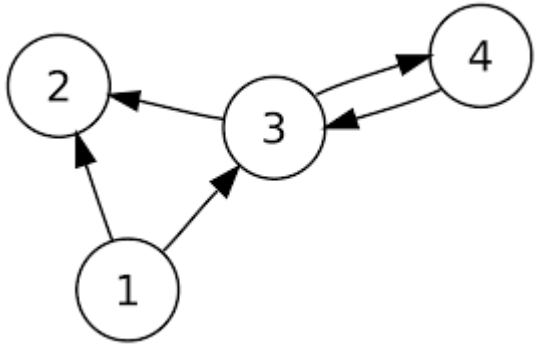
Note that for an undirected graph, A is a symmetric matrix.

Example: The adjacency matrix of the following graph is



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

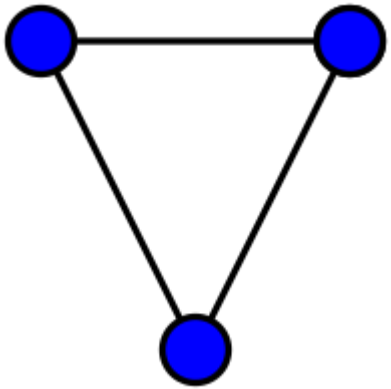
Example for Directed Graph

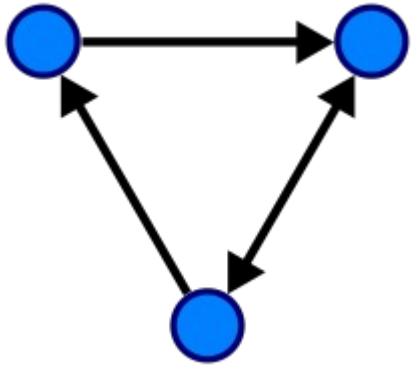


$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Connected Graph

Connected Graph: A graph is connected if, for all pairs of vertices, there is a path between these vertices on the graph.





Page Rank

Page Rank

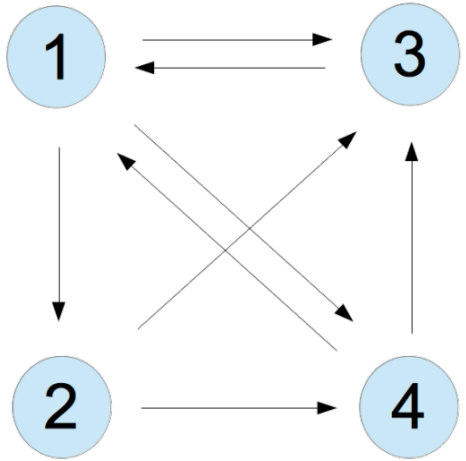
Page Rank: The goal of PageRank is to quantitatively rate the importance of each page on the web, allowing the search algorithm to rank the pages and thereby present to the user the more important pages first.

Main Steps in a Search Engine:

1. Crawl the web and locate all, or as many as possible, accessible webpages.
2. Index the data of the webpages from step 1.
3. Rate the importance of each page in the database, so that when a user does a search, more important pages can be presented first.

Key idea of the Page Rank Algorithm

Page Rank: A key idea in assigning a score to any given webpage is that the page's score is derived from the links made to that page from other webpages — "A person is important not if it knows a lot of people, but if a lot of people know that person".



The Page Rank Algorithm

Notation:
Denote $\boldsymbol{x_i}$ as the number of backlinks for page i.
Denote $\boldsymbol{n_j}$ as the number of outgoing links from page j.
Denote $\boldsymbol{L_k}$ as the set of pages with a link to page k.

Score calculation: In the page rank algorithm, if page j contains $\boldsymbol{n_j}$ outgoing links, one of which links to page k, then we will boost page k's score by $\frac{x_j}{n_j}$. Therefore,

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}$$

Page Rank Algorithm

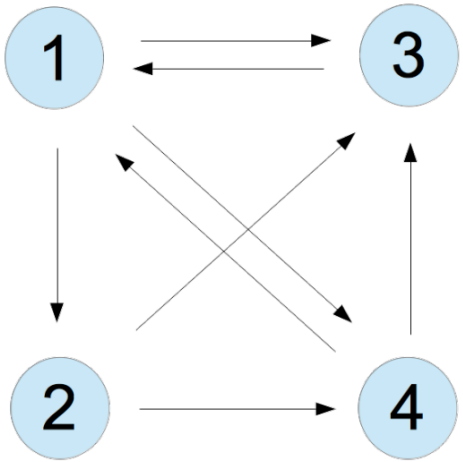
Key equation: Assume there are n pages in total. Denote $x = [x_1, \cdots, x_n]^T$. Then

$$x = Ax,$$

where

$$A_{ij} = \begin{cases} \frac{1}{n_j}, & \text{if } j \in L_i, \\ 0, & \text{otherwise.} \end{cases}$$

Example



$$A = \begin{pmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{pmatrix}$$

Page Rank Algorithm: based on the key equation for the page rank algorithm, x is the eigenvector corresponding eigenvalue 1 of the matrix A.

Theorem: $\boldsymbol{\lambda = 1}$ is the leading eigenvalue of the matrix A in the page rank algorithm.

Damping Factor

Damping factor: The PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue following links is a damping factor d. The probability that they instead jump to any random page is 1 - d.

$$x_k = \frac{1-d}{n} + d \sum_{j \in L_k} \frac{x_j}{n_j}$$

In other words, x is the solution to the following linear system

$$x = \frac{1-d}{n} * 1_{\mathrm{n}} + \mathrm{d} * \mathrm{Ax}.$$

Spectral Clustering (Part I)

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

A motivation question

Question: How to cluster the data points with only the similarity matrix (also called affinity matrix) A?

A commonly used similarity matrix for a given collection of data $\{x_1, \ldots, x_n\}$ is defined by

$$A_{ij} = \begin{cases} exp(-\|x_i - x_j\|^2/2\sigma^2), & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases}$$

The Laplacian Matrix

Laplacian Matrix: The Laplacian matrix with respect to the similarity matrix A is defined by

$$L = D - A,$$

where D is a diagonal matrix defined by

$$D_{ii} = \sum_j A_{ij}.$$

An example: Consider three data points $x_1 = (0.54, 1.83), x_2 = (-2.26, 0.86), x_3 = (0.32, -1.31)$ and $\sigma = 1$. We can form

$$A = \begin{pmatrix} 0 & 0.013 & 0.007 \\ 0.013 & 0 & 0.003 \\ 0.007 & 0.003 & 0 \end{pmatrix} \qquad L = \begin{pmatrix} 0.02 & -0.013 & -0.007 \\ -0.013 & 0.016 & -0.003 \\ -0.007 & -0.003 & 0.01 \end{pmatrix}$$

Normalized Laplacian Matrix

Normalized Laplacian Matrix: The normalized Laplacian matrix with respect to the similarity matrix A is defined by

$$L^{norm} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}.$$

An example: Consider three data points $x_1 = (0.54, 1.83), x_2 = (-2.26, 0.86), x_3 = (0.32, -1.31)$ and $\sigma = 1$. We can form

$$A = \begin{pmatrix} 0 & 0.013 & 0.007 \\ 0.013 & 0 & 0.003 \\ 0.007 & 0.003 & 0 \end{pmatrix} \qquad L^{\text{norm}} = \begin{pmatrix} 1 & -0.727 & -0.495 \\ -0.727 & 1 & -0.237 \\ -0.495 & -0.237 & 1 \end{pmatrix}$$

Spectral Clustering algorithm

Input: Similarity matrix A, number of clusters K.

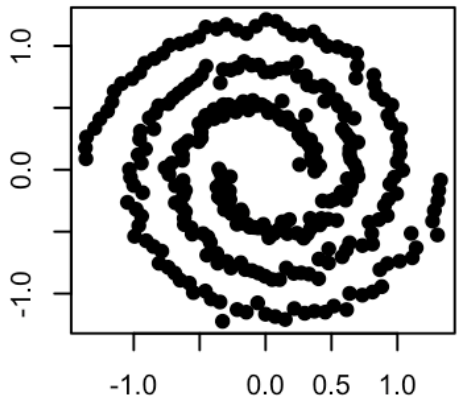
Step 1: Compute the noralized Laplacian matrix L^{norm}.

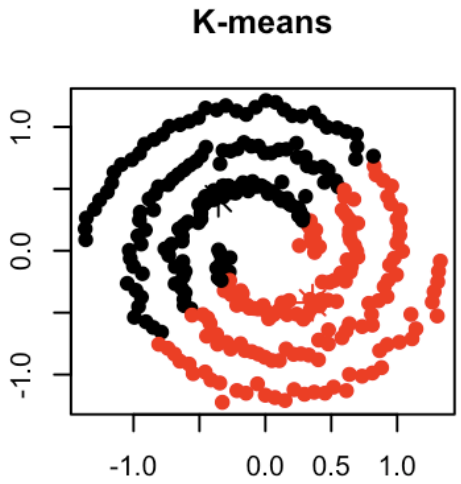
Step 2: Calculate the eigenvectors corresponding to the smallest K eigenvalues of L^{norm}. Form it as $V = [v_1, v_2, \dots, v_K]^T$.

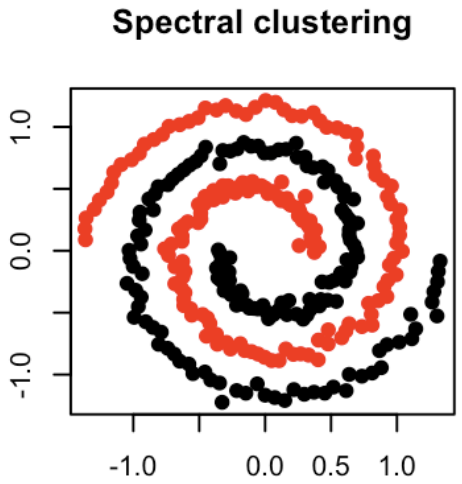
Step 3: Run K-means algorithm on the columns of V.

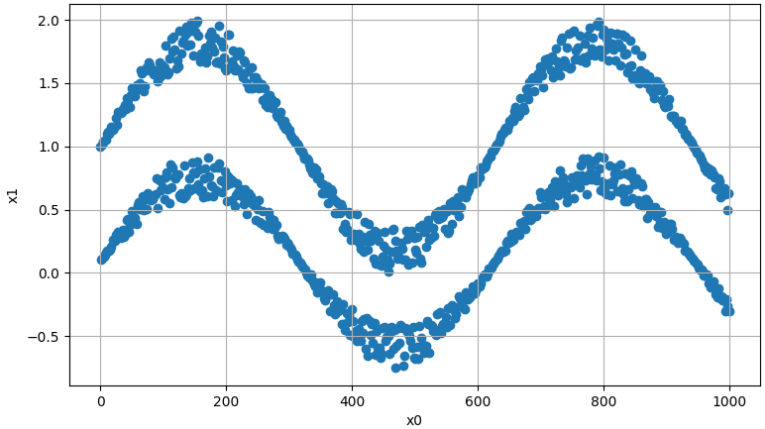
Applications

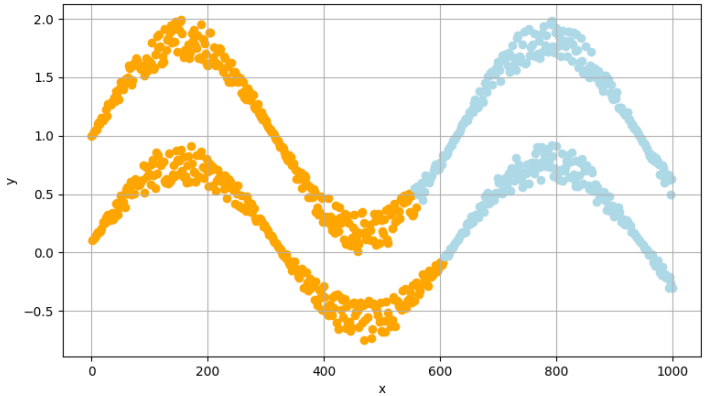
Spectral Clustering



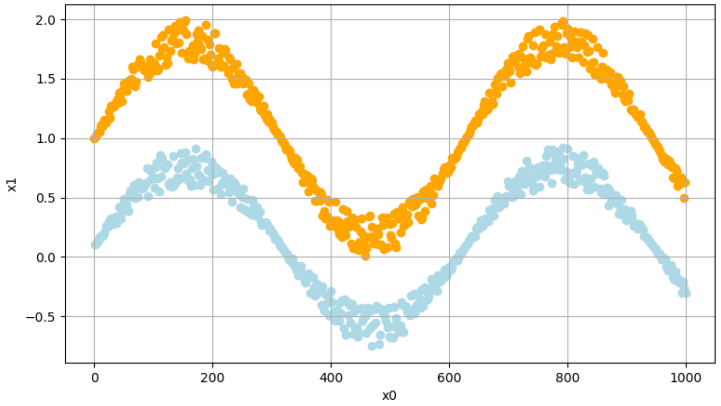




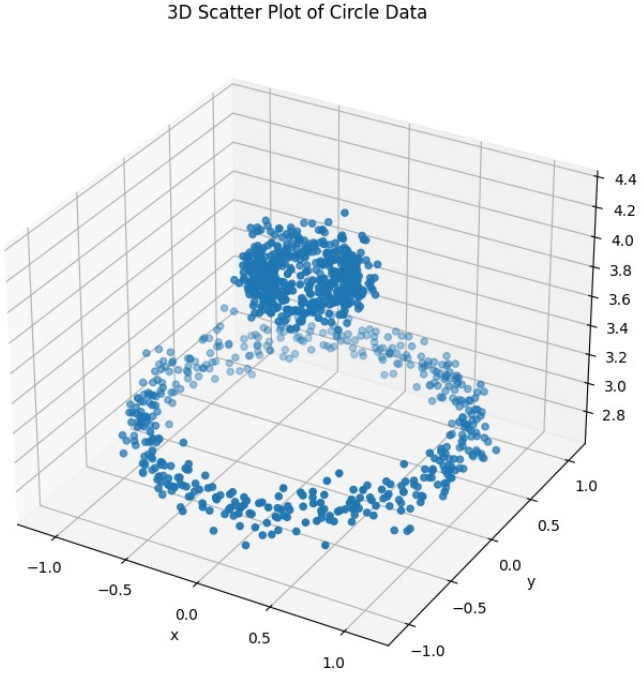


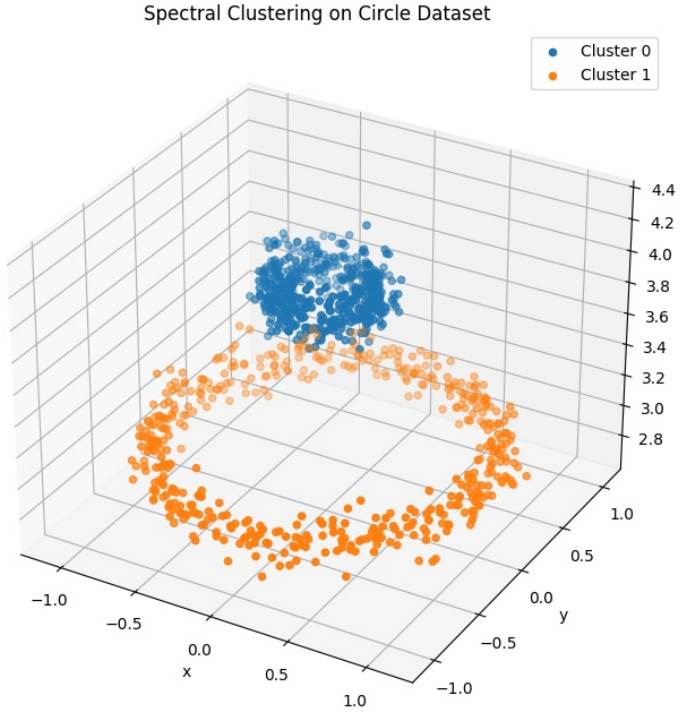


K-means



Spectral Clustering


3D Scatter Plot of Circle Data
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
2.8
−1.0
−0.5
0.0
0.5
1.0
x
y


Spectral Clustering on Circle Dataset
Cluster 0
Cluster 1
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
2.8
−1.0
−0.5
0.0
0.5
1.0
x
y


Thank You!