Introduction to Spectral Clustering

Ou Wei (Huynh's lab)

Japan Advanced Institute of Sciences and Technology

15, May, 2017

Overview

Plan

Which clustering techniques to use ?

K-means can be used when data:

- graphs are not involved
- neighbouring relation are not considered

Spectral clustering when data:

- graphs are involved
- neighbouring relation are considered

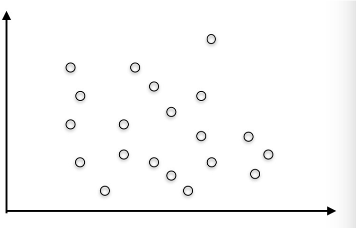


Figure 1: K-Means

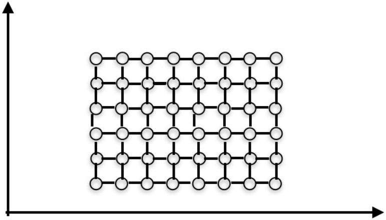


Figure 2: Spectral Clustering

Plan

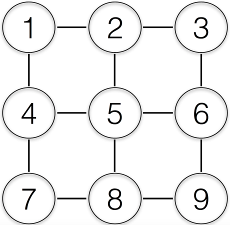
Procedure of Spectral Clustering

Given an input dataset

- predefine the number of clusters K
- Construct the similarity matrix W
- Compute the Laplacian L of W
- Compute the eigenvectors of L: $V = [v_1, v_2, ..., v_n]$
- Keep only the first K eigenvector $V' = [v_1, v_2, ..., v_k]$
- Use each row of V' as a datapoint and run K-means
- The K-means label of the i^{th} row of V' is the cluster label of the i^{th} instance of the input data

Constructing similarity matrix W

A number of options available, here we use k-nearest neighbour



	1	2	3	4	5	6	7	8	9
1	1.0	d_{12}	0	d_{14}	0	0	0	0	0
2	d_{21}	1.0	d_{23}	0	d_{25}	0	0	0	0
3	0	d_{32}	1.0	0	0	d_{36}	0	0	0
4	d_{41}	0	0	1.0	d_{45}	0	d_{47}	0	0
......									

Compute the Laplacians L of W

- First compute degree matrix D whose row $d_i = \sum_{j=1}^{n} w_{ij}$
- Then compute the unnormalised Laplacian $L = D - W$
- There exist other forms of Laplacian
- The Laplacian matrix must satisfy a certain set of conditions

Comparison between K-means and Spectral clustering in image segmentation

K=6

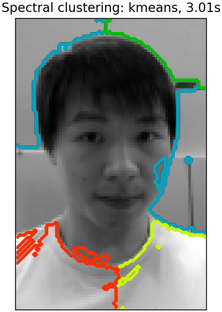


Figure 3: Spectral Clustering

Figure 4: K-Means

Plan

Generalised similarity matrix

- similarity matrix is usually task-wise
- difference choices of similarity matrix lead to very different performances
- any way to find a generalised similarity matrix ?
- Self-expression property comes to rescue !

Self-expression property

For an input dataset X, the value of a point x_i can be

- expressed as the weighted sum of all the points $x_i \approx \sum_j x_j z_{ij}$
- Vectorise the equality, we get $X \approx XZ$
- Z can be a generalised alternative to the traditional similarity matrix
- however, Z is unknown, how to find ?

Constructing the following cost function to fit z

$$||X - XZ||_F^2 + \alpha ||Z||_2^F \tag{1}$$

Then use the same spectral procedure on Z

Plan

1. Motivation
2. Spectral clustering
3. Spectral clustering with generalised similarity matrix
4. Q&A

- Can the generalised similarity matrix be used on categorical data ?