Spectral / K-means clustering

T-61.6020

Special Course in Computer and Information Science

Ville Lämsä
ville.lamsa@tkk.fi

April 2, 2008

Outline

- Introduction
- Spectral clustering
  - Algorithm
  - Example
- K-means clustering
  - Algorithm
  - Example
- Summary

Introduction

- *"Clustering is a process of organizing objects into groups whose members are similar in some way."*
- Spectral clustering: data points as nodes of a connected graph and clusters are found by partitioning this graph, based on its spectral decomposition, into subgraphs.
- K-means clustering: divide the objects into *k* clusters such that some metric relative to the centroids of the clusters is minimized.

Introduction

Spectral clustering

- How to define a graph for spectral methods?
- How to partition a graph into subgraphs?

K-means clustering

- How to choose K?
- Which metric to use?

Spectral clustering / Algorithm

By Ng, Jordan and Weiss

- Given a data set $S = \{s_1,\dots s_n\}$ to be clustered

1. Calculate the affinity matrix $\mathbf{A_{ij} = \exp(-\|s_i - s_j\|^2/2\sigma^2)}$, if $i \neq j$ and $A_{ii} = 0$ where σ^2 is the scaling parameter
2. Define $\boldsymbol{D}$ to be the diagonal matrix whose (i,i)-element is the sum of A's i-th row, and construct the matrix $\boldsymbol{L = D^{-1/2} A D^{-1/2}}$
3. Find k largest eigenvectors of L and form the matrix $\boldsymbol{X = [x_1\, x_2 \dots x_k]}$
4. Form the matrix Y from X by normalizing each of X's rows to have unit length, $\boldsymbol{Y_{ij} = X_{ij} / (\Sigma_j X_{ij}^2)^{1/2}}$
5. Treating each row of Y as a point, cluster them into k clusters via K-means or any other algorithm
6. Assign the original point s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j

Spectral clustering

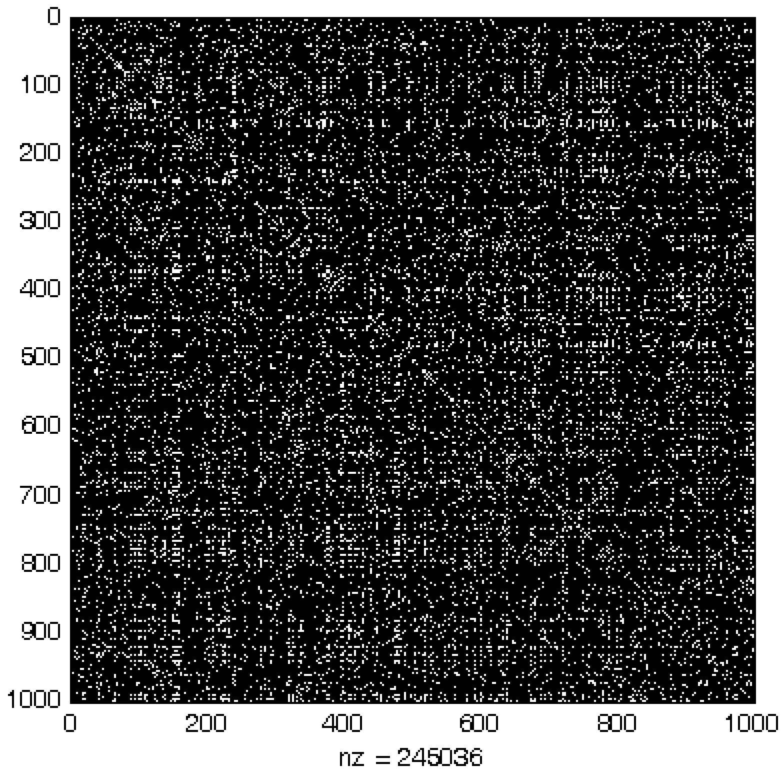
- Main difference between algorithms is the definition of L

  Meila & Shi: $L = D^{-1}A$, use largest eigenvectors, no normalization

  Shi & Malik: $L = D - A$, use smallest eigenvectors, no normalization

Spectral clustering / Example

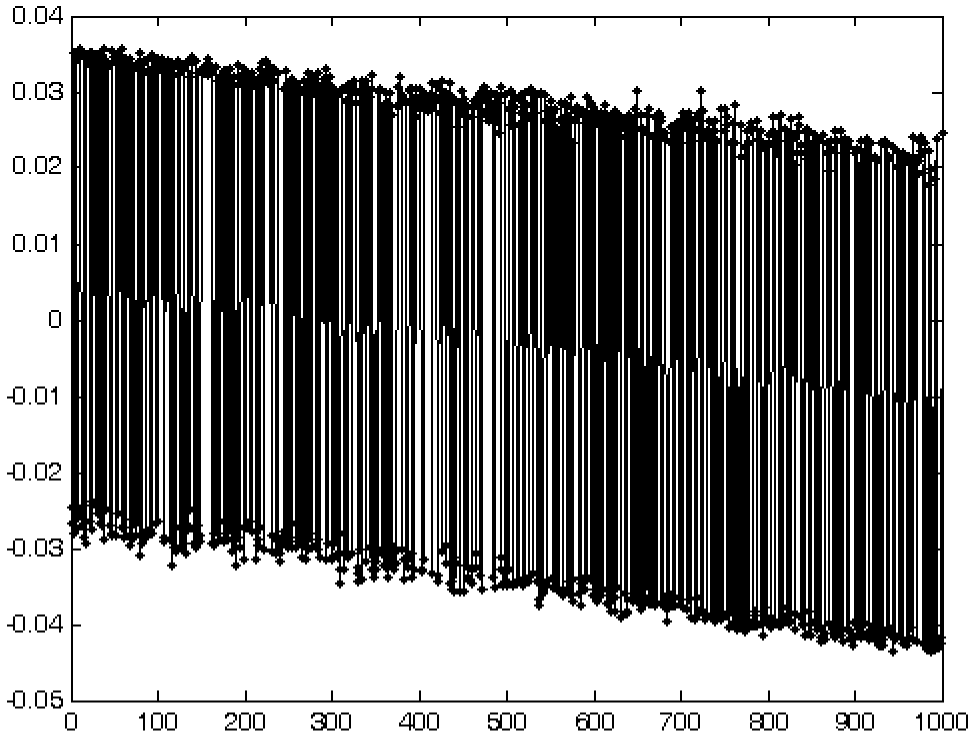
Graph of two groups...can you believe?



The adjacency matrix

Spectral clustering / Example

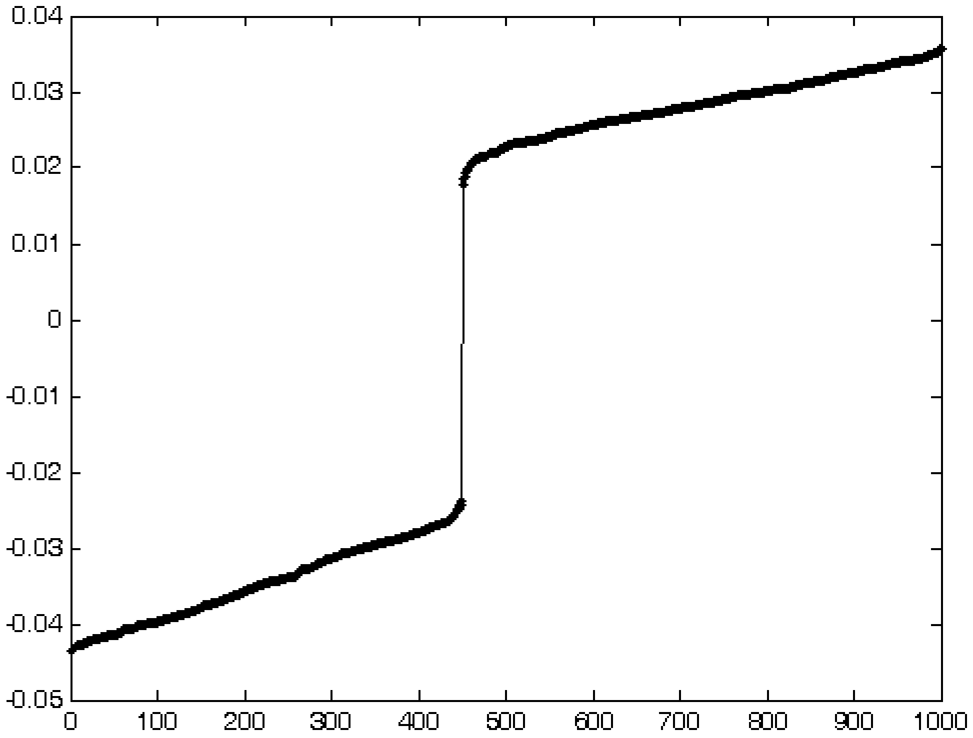
The second smallest eigenvector, unsorted



The second smallest eigenvector is the Fiedler vector, i.e. algebraic connectivity

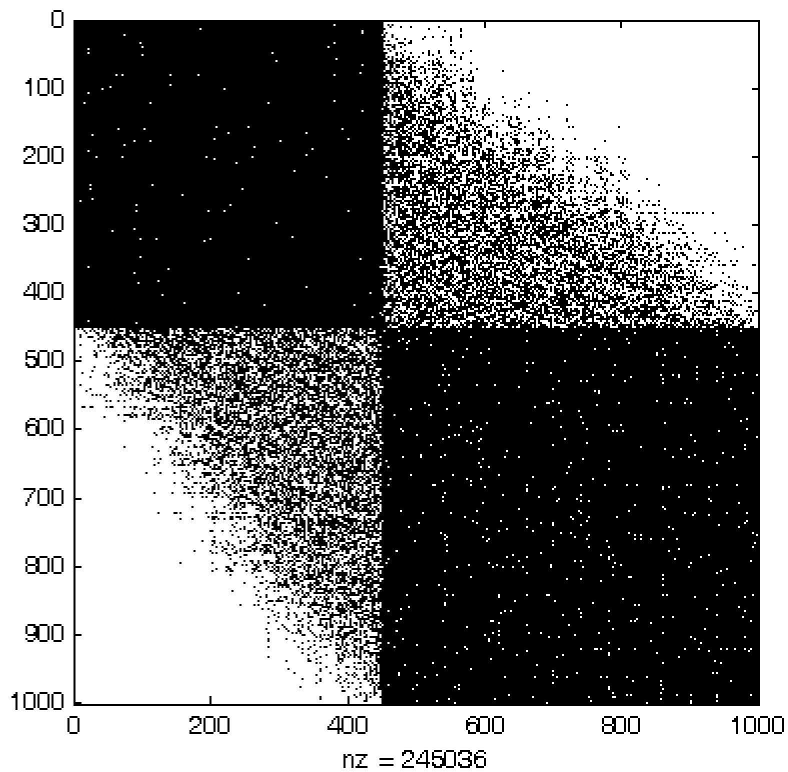
Spectral clustering / Example

The second smallest eigenvector, sorted



Spectral clustering / Example

Permutated graph...looks like there is two groups



~0.5 s

K-means clustering / Algorithm

<u>Standard version</u>

- Given an integer k and a set of n data points $X \subset R^d$

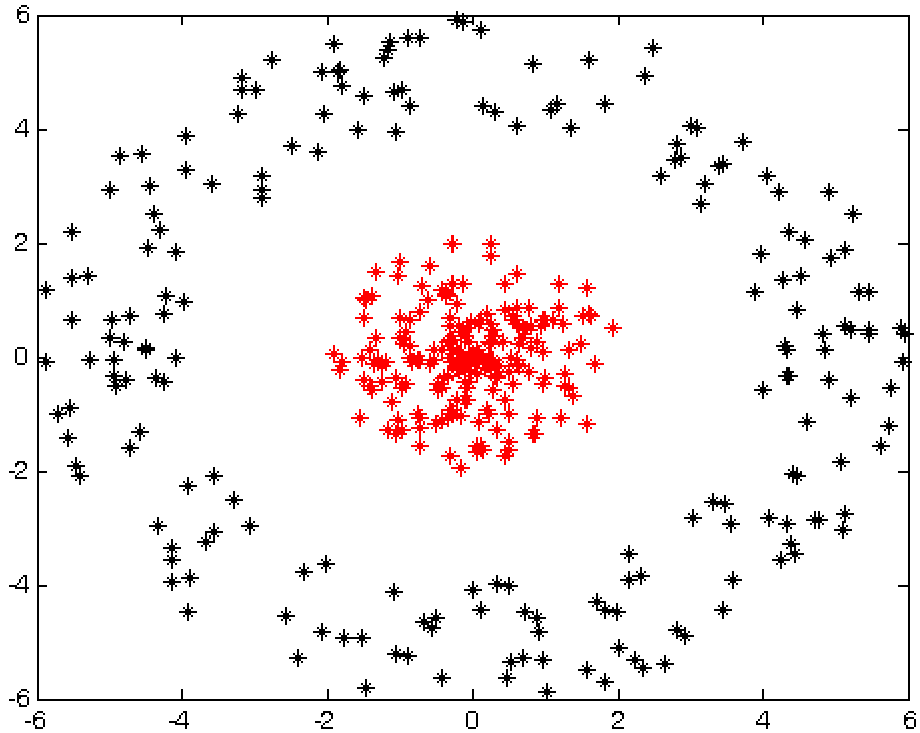
1. Arbitrarily choose initial k centers $C = \{c_1, c_2, ..., c_k\}$, $C \in R^d$
2. For each $i \in \{1,...,k\}$, set the cluster C_i to be the set of points X that are closer to c_i than they are to c_j for all $j \neq i$
3. For each $i \in \{1,...,k\}$, set c_i to be the center of mass of all points in C_i, i.e. $c_i = (1/|C_i|) \Sigma_{x \in C_i} x$
4. Repeat Steps 2 and 3 until C no longer changes

K-means clustering

- Standard practice to choose the initial centers uniformly at random from X
- For Step 2, ties may be broken arbitrarily, as long as the method is consistent
- Steps 2 and 3 guarantee to decrease the intra-cluster variance, i.e. to minimize the potential function $\phi = \Sigma_{x \in X} \min_{c \in C} ||x-c||^2$, until it is no longer possible to do so
- Some extension:
    - Different ways to choose *k* initial centers, e.g. K-means++
    - Force the center point of each cluster to be one of the actual points, i.e. K-medoids
- All in all: fast, simple, no approximation guarantees

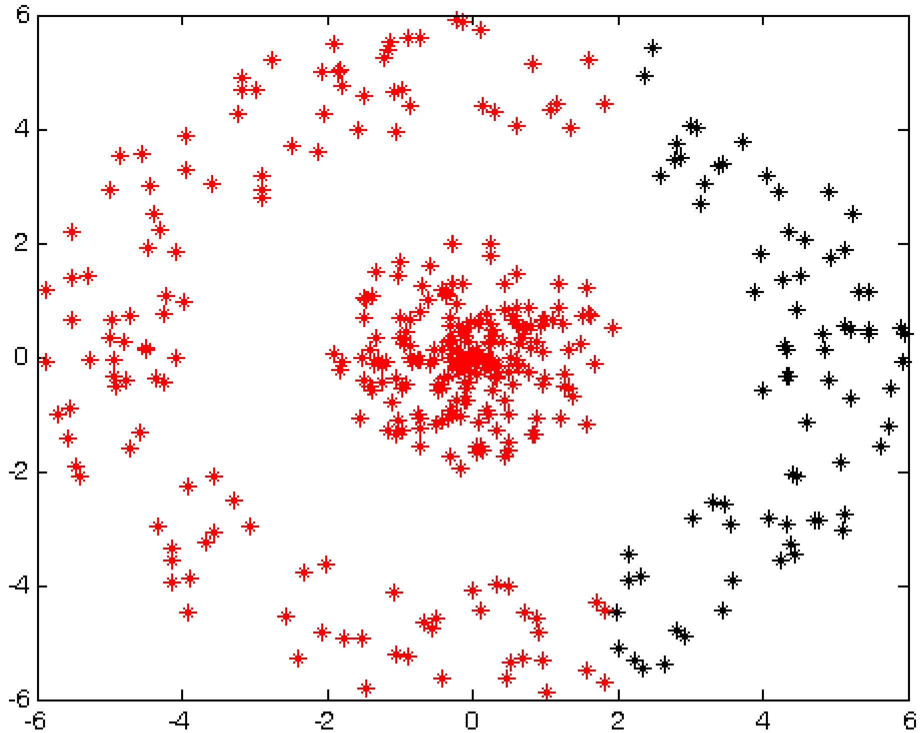
K-means clustering / Example

Two circles, originally



K-means clustering / Example

Two circles, K-means clustered (standard Matlab kmeans)



~0.01 s

Summary

- If successful, algorithm by Ng, Jordan & Weiss will dramatically improve the results of the standard K-means
- Tip: eigs(xxx,xxx,'opt'); , where opt has essential role
- Remember to normalize *Y*

References

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. *On spectral clustering: Analysis and an algorithm*. In: Advances in Neural Information Processing Systems 14, 2002.

D. Arthur, S. Vassilvitskii. k-means++*: The Advantages of Careful Seeding*. Symposium on Discrete Algorithms (SODA), 2007.