Spectral Clustering

Brief Overview

Dmitry Adamskiy

Outline

1. Motivation
2. Similarity Graphs
3. Graph Laplacian(s)
4. The algorithm(s)
5. NCut and RatioCut approximations
6. Random walk viewpoint
7. Practical considerations

Motivation

Motivation

K-means has a number of drawbacks:

- Hard assignment

Motivation

K-means has a number of drawbacks:

- Hard assignment
- Works well only with balanced clusters...

Motivation

K-means has a number of drawbacks:

- Hard assignment
- Works well only with balanced clusters...
- ...and convex ones
- The solution depends on the initialisation and it is common to re-run K-means several times and keep the best one.

Some of these are addressed by mixture models and some by spectral clustering.

Overall plan

The idea in spectral clustering is this:

- Given a set of points $x_1, \ldots, x_n$ and some kind of similarity measure $s_{ij} \geq 0$ or distances d_{ij} between all pairs of data points, build an undirected similarity graph G.
- Define clustering objective as a certain partition problem for this graph.
- Relax this problem to make it solvable.

Similarity Graphs

ϵ-neighborhood graph

There are several ways of building similarity graphs. Here is one of them:

- Connect all the points with pairwise distances less than ϵ.
- As all connected points are roughly of the same scale, the graph is usually not weighted.

k-nearest neighbour graphs

- Connect v_i and v_j if x_j is among k nearest neighbours of x_i.

k-nearest neighbour graphs

- Connect v_i and v_j if x_j is among k nearest neighbours of x_i.
- But the nearest neighbour relationship is not symmetric. Two popular choices:
  - Ignore the direction of the edges: connect v_i and v_j if either x_i is in the k nearest neighbours of x_j or x_j is among k nearest neighbours of x_i
  - Mutual k-nearest neighbour graph: replace "or" with "and" in the definition above.
- In both cases we weight the edges by the similarity of the data points $w_{ij} = s_{ij}$

Fully connected graph

- Connect all the points with positive similarity and weight all the edges by s_{ij}.
- We would like the graph to model local neighborhood relationships, so need a similarity function which encodes this.
- Gaussian kernel is widely used: $s_{ij} = \exp(-\frac{||x_i - x_j||}{2\sigma^2})$. The parameter σ controls the width of the neighbourhood.

Graph Laplacian(s)

Basic Graph Definitions

- Before we get to Graph Laplacians, some basic graph theory. Let $G = (V, E)$ be an undirected weighted graph with non-negative edge weights. Let w_{ij} be the weight of an edge connecting v_i and v_j. The degree of the vertex i is

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

- The weighted adjacency matrix is the matrix $W = (w_{ij})$. If $w_{ij} = 0$ then v_i and v_j are not connected.
- The degree matrix D is a diagonal matrix with $d_1, \ldots, d_n$ on the diagonal.
- For a subset $A \subset V$ denote the complement $\bar{A} = V \setminus A$. Two ways of measuring the "size" of a subset:

$$|A| := \text{The number of vertices in } A$$

$$\text{vol}(A) := \sum_{i \in A} d_i$$

Graph Laplacian

The main tool in spectral clustering is Graph Laplacian. There are several versions of it.

- Unnormalised graph Laplacian:

$$L = D - W$$

- Properties of the Laplacian:
  1. For every vector $f \in \mathbb{R}^n$

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2.$$

  2. L is symmetric and positive semi-definite.
  3. The smallest eigenvalue of L is 0 with the corresponding eigenvector being a constant.
  4. L has n non-negative eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$

  Proof: on the board.

Laplacian and connected components

Proposition

The multiplicity k of eigenvalue 0 of L equals the number of connected components $A_1, \ldots, A_k$ in the graph. The eigenspace of eigenvalue 0 is spanned by the indicator vectors of those components

Proof: on the board.

Normalised Graph Laplacian

The other matrix which is used in Spectral Clustering is normalized Laplacian:

$$L_{rw} = D^{-1}L = I - D^{-1}W$$

rw stands for random walk and we'll see this connection soon.

Properties of the normalised Laplacian

Properties of the Normalised Laplacian:

1. L_{rw} is symmetric and positive semi-definite.
2. The smallest eigenvalue of L_{rw} is 0 with the corresponding eigenvector being a constant.
3. L_{rw} has n non-negative eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$
4. λ is an eigenvalue of L_{rw} with eigenvector v iff λ and v solve the generalised eigenproblem $Lv = \lambda Dv$

The algorithm(s)

Spectral clustering algorithm:

1. Input: Similarity matrix S, number of clusters k
2. Construct a similarity graph by one of the methods we described.
3. Compute the Graph laplacian L.
4. Compute the first k eigenvectors $v_1, \ldots, v_k$ of L.
5. Compute the vectors $y_i \in \mathbb{R}$, such that j-th component of i-th vector is the i-th coordinate of v_j.
6. Cluster the points y_i using k-means algorithm.

Based on which graph Laplacian is used the algorithm is called normalised or unnormalised spectral clustering.

NCut and RatioCut approximations

Ratio Cut and NCut

Recall the definition of the graph cut. For two disjoint subsets $A, B \in V$ we define

$$\text{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij}$$

- One way to partition a graph into k disjoint subsets is to solve a mincut problem, defining $\text{cut}(A_1, \ldots, A_k) = \sum_{i=1}^{k} \text{cut}(A_i, \bar{A}_i)$
- However, this leads to an unbalanced partitions, for example in case of $k = 2$ the minimum is often achieved by splitting off one vertex.
- Thus, we want to enforce the clusters A_i to be "reasonably large". Two natural ways of doing it: by the number of vertices and by the volume. This leads to the two definitions:

$$\text{RatioCut}(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{|A_i|}$$

$$\text{NCut}(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{\text{cut}(A_i, \bar{A}_i)}{\text{vol}(A_i)}$$

RatioCut relaxation for $k = 2$

Here is a sketch of how RatioCut for $k = 2$ is related to unnormalised spectral clustering.

- Suppose we have some partition $(A, \bar{A})$. Define a vector $f_i = \sqrt{|\bar{A}|/|A|}$ if $v_i \in A$ and $f_i = -\sqrt{|A|/|\bar{A}|}$ if $v_i \in \bar{A}$.
- Now one can show (see the board) that

$$f^T L f = 2|V|\text{RatioCut}(A, \bar{A})$$

- Also, $\sum_i f_i = 0$ and $||f||^2 = n$.
- So we can rewrite RatioCut problem as minimising across all the partitions the function $f^T L f$ subject to $||f|| = \sqrt{n}, ||f|| \perp \mathbb{1}$ and f as defined above.
- The relaxation is to drop the "as defined above" requirement and allow f to take any values:

$$\min_f f^T L f, \text{subject to } f \perp \mathbb{1}, ||f|| = \sqrt{n}$$

RatioCut relaxation (contd.)

- The relaxation is to drop the "as defined above" requirement and allow f to take any values:

$$\min_f f^T L f, \text{subject to } f \perp \mathbb{1}, ||f|| = \sqrt{n}$$

- The solution to this problem is the second eigenvector of L.
- Now we need to transform back the solution to the indicator vector.
- This is done using k-means clustering of the components.

Relaxation (contd.)

- One can show similarly that unnormalised spectral clustering corresponds to the relaxation of RatioCut...
- ...whereas Normalised spectral clustering corresponds to the relaxation of NCut.
- Example where the relaxed problem is a really bad approximation: "cockroach graph".

Random walk viewpoint

Spectral clustering could be viewed as finding the partition such that the random walk with transition probabilities proportional to edge weights stays long within the same cluster.

- Transition matrix $P = (p_{ij})$, $p_{ij} = w_{ij}/d_i$.

$$P = D^{-1}W$$

- We can see that $L_{rw} = I - P$.
- If the graph is connected and non-bipartite there is a unique stationary distribution $\pi = (\pi_1, \ldots, \pi_n)$ given by $\pi_i = d_i/\text{vol}(G)$
- Random walk and NCut equivalence. Let G be a connected non-bipartite graph. Suppose we run the random walk starting from X_0 in the stationary distribution. Then

$$\text{NCut}(A, \bar{A}) = P(\bar{A}, A) + P(A, \bar{A})$$

Practical considerations

Practical considerations

- The choice of similarity measure is important.
- So is the choice of similarity graph.
- Computations are easier for k-nn or ϵ-ball graphs (the matrices are sparse)
- Normalised clustering incorporates within-cluster similarity in the objective function:

$$\sum_{i,j \in A} w_{ij} = \text{vol}(A) - \text{cut}(A, \bar{A})$$

- For more details, see [1]

References I


U. von Luxburg.
A tutorial on spectral clustering.
CoRR, abs/0711.0189, 2007.